New Phytologist

Research

Methods

Automated hyperspectral vegetation index derivation using a hyperparameter optimisation framework for high-throughput plant phenotyping

Joshua C.O. Koh[1]*, Bikram P. Banerjee[1]*, German Spangenberg[2,3] and Surya Kant[1,2,3]

[1]Agriculture Victoria, Grains Innovation Park, 110 Natimuk Rd, Horsham, Vic. 3400, Australia; [2]Agriculture Victoria, AgriBio, Centre for AgriBioscience, 5 Ring Road, Bundoora, Vic. 3083, Australia; [3]School of Applied Systems Biology, La Trobe University, Bundoora, Vic. 3083, Australia

Author for correspondence:
Surya Kant
Email: *surya.kant@agriculture.vic.gov.au*

Received: *18 October 2021*
Accepted: *16 December 2021*

New Phytologist (2022) **233:** 2659–2670
doi: 10.1111/nph.17947

Key words: automated vegetation index development, chlorophyll estimation, high-throughput plant phenotyping, hyperparameter optimisation, hyperspectral vegetation indices, sugar estimation, wheat.

Summary

- Hyperspectral vegetation indices (VIs) are widely deployed in agriculture remote sensing and plant phenotyping to estimate plant biophysical and biochemical traits. However, existing VIs consist mainly of simple two-band indices that limit the net performance and often do not generalise well for traits other than those for which they were originally designed.
- We present an automated hyperspectral vegetation index (AutoVI) system for the rapid generation of novel two- to six-band trait-specific indices in a streamlined process covering model selection, optimisation and evaluation, driven by the tree parzen estimator algorithm. Its performance was tested in generating novel indices to estimate chlorophyll and sugar contents in wheat.
- Results showed that AutoVI can rapidly generate complex novel VIs (at least a four-band index) that correlated strongly ($R^2 > 0.8$) with measured chlorophyll and sugar contents in wheat. Automated hyperspectral vegetation index-derived indices were used as features in simple and stepwise multiple linear regressions for chlorophyll and sugar content estimation, and outperformed the results achieved with the existing 47 VIs and those provided using partial least squares regression.
- The AutoVI system can deliver novel trait-specific VIs readily adoptable to high-throughput plant phenotyping platforms and should appeal to plant scientists and breeders. A graphical user interface for the AutoVI is provided here.

Introduction

High-throughput plant phenotyping (HTP) is integral in meeting the demand for large-scale evaluation of genotypes in breeding programmes and crop management systems (Tardieu *et al.*, 2017; Mir *et al.*, 2019). In recent years, controlled-environment and field-based HTP platforms have been developed to monitor plants at the canopy or plot level for a large number of crop genotypes (Tardieu *et al.*, 2017; Mir *et al.*, 2019; Lu *et al.*, 2020). Central to the success of these HTP platforms is the use of various imaging sensors to acquire morphological, physiological and biochemical parameters in a noninvasive manner. Hyperspectral imaging has been a promising HTP technology for measuring biochemical and morphophysiological traits, in a fast and nondestructive way, by detecting signatures in the reflectance spectrum of vegetation in narrow (e.g. 1–2 nm in spectral resolution) and contiguous/broad (e.g. ≥ 20 nm in spectral resolution) spectral bands (Lu *et al.*, 2020). The recent availability of lightweight hyperspectral sensors has stimulated a rapid adoption of these sensors for use in unmanned aerial vehicle systems for HTP and precision agriculture (Adão *et al.*, 2017; Lu *et al.*, 2020). Hyperspectral data have been applied to estimate biophysical (e.g. leaf area index, green canopy cover) and biochemical (e.g. chlorophyll, nitrogen, sugar) traits for the leaf and canopy, as well as traits relating to the whole plant, particularly those linked with growth (e.g. plant biomass, plant height) and productivity (e.g. grain yield) (Adão *et al.*, 2017; Lu *et al.*, 2020). Apart from this, close-range hyperspectral imaging (ground based or glasshouse) characterised using high spatial resolution and signal-to-noise ratio is also increasingly common in HTP facilities. These systems allow the fine-scale investigation of vegetative features at the leaf or canopy level with applications in plant water content and biochemical compounds estimation, and the detection of abiotic and biotic stresses in plants (Mishra *et al.*, 2017).

*These authors contributed equally to this work.

However, the large amount of data collected using hyperspectral sensors poses challenges in analytical implementations. Redundancy problems are linked to the multicollinearity of bands and the curse of dimensionality imposes high computational costs on analytical pipelines (Bajwa & Kulkarni, 2011; Burger & Gowen, 2011). Multicollinearity occurs when independent variables (e.g. wavebands) are highly correlated, leading to inaccurate estimation of coefficients (effect of independent variables on response/target trait) in a regression analysis and less reliable statistical inferences (Bajwa & Kulkarni, 2011; Burger & Gowen, 2011). Conversely, the curse of dimensionality refers to the problem of optimising a given function or model due to the exponential increase in possible solutions or sets of parameters associated with the increase in the number of variables (i.e. the dimensionality), which often necessitates exhaustive enumeration of the solution/search space to achieve satisfactory optimisation (Donoho, 2000). Dedicated efforts are often required to develop efficient hyperspectral data processing workflows for a specific plant phenotyping task (Aasen *et al.*, 2014, 2018). To this end, vegetation indices (VIs) offer an alternate simple and fast approach to hyperspectral data analysis. The VIs are formulated as ratios or algebraic combinations of vegetative reflectance at different wavebands selected from the visible (VIS; 400–700 nm), near infrared (NIR; 700–1000 nm) and shortwave infrared (SWIR; 1000–2500 nm) regions (Silleos *et al.*, 2006; Xue & Su, 2017). The application of VIs for plant phenotyping is straightforward, requiring the user to compute index values directly from the relevant waveband reflectance and use them either as proxy measures to the target trait or as features (or variables) in regression modelling to predict the target trait values. Since the introduction of the first VI by Pearson & Miller (1972), more than 500 hyperspectral VIs have been developed, demonstrating a strong and continued interest in the development and adoption of novel VIs for specific remote sensing applications (Henrich *et al.*, 2017). However, existing VIs consist predominantly of two-band indices and, to some extent, three-band indices; this limits the amount of information represented and therefore the net performance produced by these VIs (Henrich *et al.*, 2017). In addition, existing VIs are designed for specific plant traits and often do not generalise well for other traits. As such, particularly in agriculture, there is a continued demand for novel VIs that can target specific traits associated with crop growth, biochemical parameters, yield and quality.

The development of VIs is technically challenging and time consuming, and often requires a comprehensive understanding of the dynamic changes of the plant optical properties in relation to the intrinsic biochemical or biophysical trait(s) of interest. To this end, wide varieties of experiments have been established to acquire a comprehensive spectral library (Rao *et al.*, 2007; Chauhan & Mohan, 2013). Ideally, knowledge on wavebands associated with plant traits may be enriched or expanded with successive developments of new VIs when different regions of the reflectance spectrum corresponding to vegetative features are identified. Biochemical and biophysical traits can be described comprehensively with more wavebands in which each waveband adds supplementary information. However, this is seldom the case, as the VIs rarely constitute a complex or cohesive assemblage of wavebands, but rather a limited selection of a few wavebands (usually up to four bands). In addition, multicollinearity of bands and the curse of high dimensionality inherent in hyperspectral data complicate the identification of wavebands linked to the underlying trait of interest. Several attempts have been made to accelerate the development of novel VIs, these include the use of correlation matrices between VIs and the target traits of interest to retrieve new waveband or index combinations (Thenkabail *et al.*, 2004; Aasen *et al.*, 2014; Xu *et al.*, 2019). Careful selection of hyperspectral features (wavebands) has been effective in overcoming the curse of dimensionality and selected bands could be combined to develop VIs (Aasen *et al.*, 2014, 2018). Recently, a brute force indices mining approach was applied in our laboratory to identify a new normalised difference chlorophyll index ($NDCI_w$) for the estimation of chlorophyll content in wheat (Banerjee *et al.*, 2020). However, these approaches are suited for the evaluation of a limited number of wavebands and/or index model combinations for the development of new VIs; and may be computationally less efficient when dealing with a greater number of wavebands and index models. An efficient VIs evaluation and waveband selection strategy is crucial to the development of trait-specific hyperspectral VIs for HTP and agriculture remote sensing.

This study aimed to report the development and evaluation of an automated hyperspectral vegetation index (AutoVI) derivation system for the rapid generation of trait-specific novel two-band to six-band VIs based on a hyperparameter optimisation framework. The term 'hyperparameter optimisation' (HPO) is often associated with the machine learning discipline, in which specific algorithms are deployed to select optimum values in a defined search space for model parameters (values learned from data) and hyperparameters (values associated with the model function or architecture) to maximise the model performance (Bergstra *et al.*, 2011; Yu & Zhu, 2020). Building upon an HPO framework, the AutoVI is designed to deliver an end-to-end VI development pipeline that covers index model evaluation and optimum waveband selection with minimal user input. In this study, novel VIs for chlorophyll and sugar estimation in wheat were generated using the AutoVI and compared against existing VIs as features in simple and multiple regression modelling, with the results also compared against those computed using partial least squares regression (PLSR), a well established method for plant trait prediction using hyperspectral data (Ely *et al.*, 2019; Wu *et al.*, 2019; Burnett *et al.*, 2021). The potential application of the AutoVI for HTP and agriculture remote sensing is discussed.

Materials and Methods

Experiment in a high-throughput phenotyping facility

Data used in this study were collected in a high-throughput controlled-environment phenotyping facility in the Plant Phenomics Victoria, Horsham (PPVH), Agriculture Victoria as previously described (Banerjee *et al.*, 2020). In brief, the PPVH facility is equipped with a conveyor belt system, automated

weighing and watering stations, and an automated phenotyping Scanalyzer 3D system (LemnatecGmBH, Aachen, Germany), which includes a hyperspectral imaging sensor. For the experiment conducted at the PPVH, wheat plants were grown under 2, 5, 10, or 20 mM nitrogen (N) levels. One plant per pot was grown in a nutrient-free growth medium consisting of perlite covered with a layer of vermiculite. Individual pots were weighed and equalised to a fixed pot weight and watered uniformly. The pots were loaded onto the system 10 d after the emergence of seedlings. Nutrient solution (4 mM $MgSO_4$, 4 mM KCl, 5 mM $CaCl_2$, 3 mM KH_2PO_4/K_2HPO_4 pH 6.0, 0.1 mM Fe-EDTA, 10 μM $MnCl_2$, 10 μM $ZnSO_4$, 2 μM $CuSO_4$, 50 μM H_2BO_3 and 0.2 μM Na_2MoO_4) with the indicated N concentrations was supplied as 100 ml per pot every week. The growing conditions were 16 h : 8 h, 24 C : 15°C, day : night. The experiment was conducted as biological repeats with 20 replicate plants per N treatment. A subset of five plants per N treatment was destructively harvested at 14, 21, 28 and 35 d after sowing (DAS) and, in total, 80 samples were collected for biochemical assays.

Hyperspectral image acquisition and processing

Plants were scanned in an imaging station with a push broom hyperspectral sensor (Micro-Hyperspec, VNIR-E Series; Headwall Photonics, Fitchburg, MA, USA) over a spectral range (475–1710 nm) and three viewing angles (0°, 120°, and 240°). Raw data acquired in 12-bit digital numbers (DNs) were transformed to radiance following spectral and radiometric calibration using a white reference target (Spectralon panel; Labsphere Inc., North Sutton, NH, USA) and dark reference (spectrum collected with halogen lamps turned off) in a data-acquisition software (Hyperspec III; Headwall Photonics, Inc., Bolton, MA, USA). Further processing was applied to remove interchannel variation and correct illumination variations (Banerjee *et al.*, 2020). Non-plant pixels (cage, pot, soil, and background) in the hyperspectral image were first classified using the spectral information divergence method (Chein, 1999) and a binary mask was applied to segment out the remaining pixels (i.e. the plant pixels). Detected plant pixels were averaged to generate a representative reflectance spectrum with 256 spectral bands and resampled to a spectral width of 1 nm using a linear resampling approach (Lewitt, 1990). In total, 47 published VIs within the spectral range 475–1710 nm were then computed (Supporting Information Table S1).

Biochemical assays

Whole plant samples were finely ground in liquid nitrogen using a pestle and mortar, then subsampled separately for chlorophyll and sugar analysis and stored at −80°C until biochemical analysis. For chlorophyll analysis, chlorophyll was extracted from 100 mg of sample with 100% methanol and centrifuged for 10 min at 10 016 ***g***; this process was repeated twice. Extracts were analysed by recording the absorbance at 750, 665, 652 and 470 nm using an ultraviolet–VIS (UV–VIS) light spectrophotometer (Shimadzu UV-1800; Shimadzu Inc., Kyoto, Japan). Total chlorophyll was calculated using the formula described in Lichtenthaler (1987) and ranged between 396.2 and 821.9 μg g^{-1} in this study. For sugar analysis, soluble sugars were extracted from 100 mg samples with 80% ethanol and centrifuged for 10 min at 10 016 ***g***; this process was repeated twice. Total soluble sugars were assayed according to the colorimetric method described in DuBois *et al.* (1956) and ranged between 2600 and 28 300 μg g^{-1} in this study.

Automated hyperspectral vegetation index derivation

An automated system for hyperspectral vegetation index derivation named AutoVI, was developed to deliver effective VIs in a streamlined process covering model selection, model parameter generation, model parameter tuning and model evaluation, and driven by a hyperparameter optimisation framework, that is the optimiser, for plant phenotyping (Fig. 1). The optimiser is the core of the AutoVI method that seeks to generate a VI model for the desired trait based on a model evaluation metric, that is the objective function score (R^2, described in the following paragraphs) using an optimisation algorithm given time and computing resource constraints. However, unlike simple optimisation challenges that typically search for optimal solutions for a single model or function (static search space), multiple index models

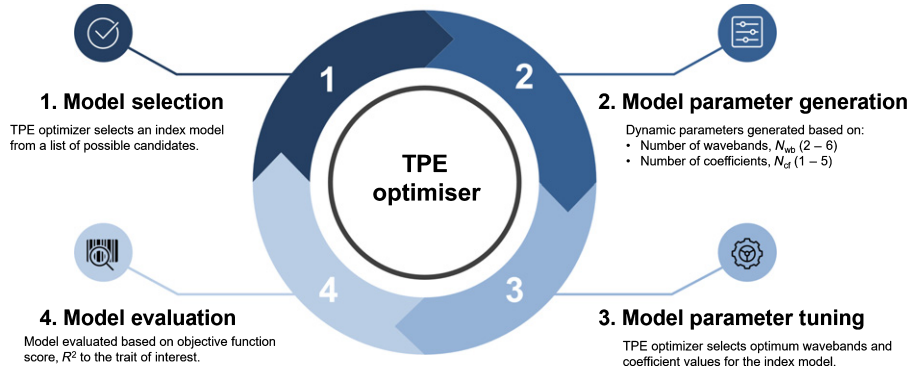

Fig. 1 Overview of the automated hyperspectral vegetation index (AutoVI) framework. The tree parzen estimator (TPE) algorithm is implemented as the optimiser for AutoVI. A single iteration consists of model selection, model parameter generation, model parameter tuning and model evaluation. The AutoVI is programmed to repeat the optimisation process until a predetermined number of iterations is reached. The repeated iterations seek to maximise the objective function score, R^2, by selecting the best candidate model and an optimum set of hyperparameters.

(dynamic search spaces) were optimised and evaluated in the AutoVI (Fig. 1). This is made possible using dynamic parameter programming or 'define-by-run' coding (Akiba *et al.*, 2019), which generates the search space or set of model parameters during code execution, depending on the index model or equation under evaluation.

The mathematical expression (or model) defining a VI is primarily needed to combine two or more wavebands to decipher certain biochemical or biophysical traits of interest. The development of a new VI requires the selection of both a suitable model and a set of wavebands. A library of 33 index models was created (Table S2) after an exhaustive review of more than 500 previously developed VIs (Henrich *et al.*, 2017). The listed model equations are of varying complexities, differing in the number of distinct wavebands (N_{wb} = 2–6, i.e. B1, B2, . . ., B6) and the number of coefficients (N_{Cf} = 1–5, i.e. α, β, . . ., ρ). The AutoVI system begins with model selection (step 1, Fig. 1) in which the optimiser selects a mathematical model at random from the library of 33 index models and generates the model parameters corresponding to the N_{wb} and N_{Cf} (step 2, Fig. 1). This is followed by model parameter tuning (step 3, Fig. 1) in which optimum wavebands and coefficient values (between 0 and 1) are selected to maximise the objective function score (step 4, Fig. 1), which in this study is the coefficient of determination, R^2, derived from a simple linear regression fitted to calculated index values and ground truth values for the measured trait of interest (in this case chlorophyll or sugar content). One run of optimisation (steps 1–4) is referred to as a single iteration; the AutoVI system is programmed to repeat the optimisation process until a predetermined number of iterations is reached. At each iteration, an index model is selected, and the computed objective function score (R^2) is compared with the previous iteration. Additionally, unique sets of model parameters are computed for the respective index models. The repeated iterations seek to maximise the objective function score, R^2, by selecting the best candidate model and an optimum set of model parameters.

In this study, the tree parzen estimator (TPE) (Bergstra *et al.*, 2011; Yu & Zhu, 2020) was implemented as the optimisation algorithm in the AutoVI, as it is widely used for hyperparameter optimisation in machine learning problems and showed better accuracy and efficiency compared with other algorithms when dealing with dynamic search spaces (Bergstra *et al.*, 2015; Akiba *et al.*, 2019; Yu & Zhu, 2020). The TPE algorithm is a variant of Bayesian optimisation approaches that tries to construct a probabilistic model, also known as a 'surrogate' model for mapping hyperparameters based on the probability of an objective function score given the set of hyperparameters. The AutoVI system, including the TPE algorithm customised to handle hyperspectral data, was implemented in PYTHON using the open source hyperparameter optimisation library, OPTUNA v.2.0 (Akiba *et al.*, 2019) with default settings. A graphical user interface (GUI) for the AutoVI and source codes for TPE implementation are hosted on the public repository GitHub (see the Data availability section). The AutoVI system was tested on an AMD Threadripper 3970X (32-cores) system with 256 GB RAM at the SmartSense iHub, Agriculture Victoria.

Automated hyperspectral vegetation index for chlorophyll content estimation

We evaluated the ability of AutoVI to derive high-quality novel hyperspectral VIs for plant phenotyping using wheat total chlorophyll content as a biochemical trait. Chlorophyll content, either measured or estimated, can be a direct indicator of a plant's primary production and has been used to determine the N status and stress response of crop plants (Richardson *et al.*, 2002; Murchie & Lawson, 2013). The existing dataset (n = 80) was split randomly in a ratio of 65 : 35 into training and test datasets, with both datasets having the same sample distribution (i.e. stratified sampling) according to time points. The resulting train-test split was used for all subsequent AutoVI computations and regression modelling in this study for chlorophyll content prediction. The AutoVI system was trained on the training dataset to derive novel indices for chlorophyll content estimation, with the performance of these indices validated using the test dataset.

One possible issue with any optimisation system is a selection bias towards index models with lower complexity, for example models with a N_{wb} value between 2 and 3 compared with those with higher complexity, for example models with $N_{wb} \geq 4$. This is because the size of the solution search space increases exponentially with the increase in the number of input features, that is N_{wb} (Winston, 1992; Yao & Liu, 1997). Consequently, more computational time or resource is required to optimise the complex models ($N_{wb} \geq 4$) compared with the simpler models ($N_{wb} \leq 3$), but their performances tend to scale better over time and do not plateau as fast as do simpler models. When all model computations are grouped together, simpler models tend to outperform complex models in the early stages of optimisation, causing these to be favoured by the optimiser and leading towards a 'locally maximal solution', which is the tendency of the computation to become stuck at a suboptimal solution (Hinneburg & Keim, 1999). To address this issue in the AutoVI, computations were performed on model groups consisting of models with the same N_{wb} (Fig. 2). Five parallel instances of the AutoVI corresponding to the five model groups (M2, M3, . . ., M6, equivalent to N_{wb} = 2, 3, . . ., 6) were executed with 20 000 iterations each, and with coefficients fixed at 1 (Fig. 2). These were repeated five times on the same train-test split as described earlier, with the best performing index model from each group logged at each repetition (Fig. 2). As TPE is initialised randomly and common to HPO algorithms, it can exhibit varied performance on the same dataset (Bergstra *et al.*, 2011; Yu & Zhu, 2020); the repeated AutoVI computations provided a measurement of the stability of the TPE algorithm. In addition, the best index model for each model group and the overall best index model could be reliably identified. The effect of longer optimisations and the inclusion of a coefficient on model performance was determined using the overall best index model, with five repeated AutoVI computations at 20 000 and 40 000 iterations with or without coefficient tuning.

To determine the quality of AutoVI-derived indices for chlorophyll estimation, they were used as features in simple linear regression (SLR) modelling to predict chlorophyll content. Simple linear regression with each of the derived indices was first

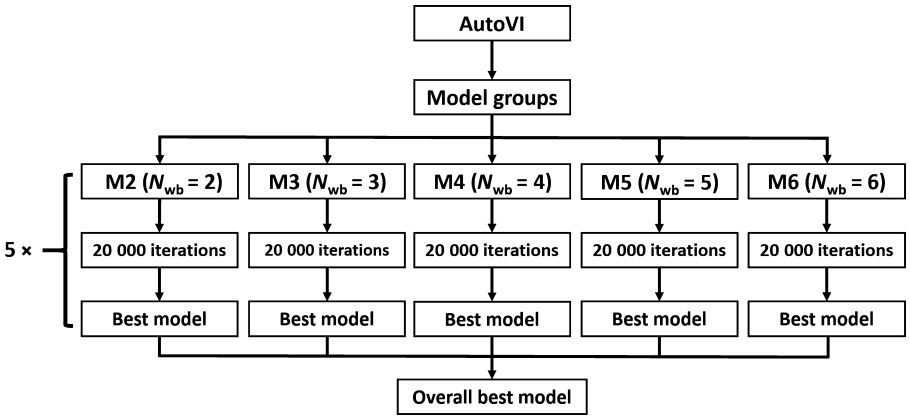

Fig. 2 Flowchart for grouped model evaluations in the automated hyperspectral vegetation index (AutoVI). Index models were grouped according to the number of wavebands (N_{wb}) and parallel computations with 20 000 iterations each were performed on the model groups (M2, M3, M4, M5 and M6). The best index model for each model group and the overall best index model were identified from five repeated computations.

trained on the training dataset and then used to predict chlorophyll values for the test dataset. Model performance was evaluated using the R^2 score calculated between predicted and actual chlorophyll values for the test dataset. In addition, performance for a stepwise multiple linear regression (SMLR) model (described in the following paragraphs) with VIs selected from 25 AutoVI-derived indices (Fig. 2) is also included for comparison. Results achieved using the AutoVI indices were compared with those produced using SLR and SMLR with 47 published VIs, in addition to results provided using PLSR modelling using the full spectrum of reflectance values, that is reflectance values from all 1235 wavebands (described in the following paragraphs). For comparison across different regression models, additional performance metrics, such as root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), were also provided.

Automated hyperspectral vegetation index for sugar content estimation

Sugar plays an important role in the osmotic adjustment of plants in response to drought stress. Some studies have shown that genotypes that have higher accumulation of sugar content in leaves or stems are more drought tolerant (Adams *et al.*, 2013; Piaskowski *et al.*, 2016). Using the grouped model evaluation approach, AutoVI computations were conducted across five repetitions with 20 000 iterations each, with the coefficients fixed at 1. The existing dataset was split at a ratio of 65 : 35 into training and test datasets as described previously, with the training of the AutoVI conducted on the training dataset and validation of derived indices performed on the test dataset. The resulting train-test split was used for all subsequent AutoVI computations and regression modelling in this study for estimation of sugar content. The effect of longer optimisations and inclusion of a coefficient on model performance was determined as described previously. Results for SLR and SMLR using AutoVI-derived indices were compared with those produced using 47 published VIs and PLSR modelling, as described in the previous section.

Stepwise multiple linear regression

Stepwise multiple linear regression is a feature selection method that iteratively adds (forward selection) or removes (backward selection) features to a multiple linear regression model to improve model performance, as indicated using an evaluation metric or score. In this study, we implemented a stepwise forward selection strategy based on a five-fold cross-validated R^2 score of a multiple linear regression model using the PYTHON package, SCIKIT-LEARN v.0.24. The maximum number of features to select was set to between 1 and 20 and selection was performed on 25 AutoVI-derived indices and 47 published VIs for both chlorophyll and sugar estimation on the training dataset. A multiple linear regression model was then fitted to the training dataset using the optimum selected features and used to predict target values (chlorophyll or sugar content) for the test dataset.

Partial least squares regression

Partial least squares regression modelling is one of the most effective methods for plant trait prediction using hyperspectral data (Ely *et al.*, 2019; Wu *et al.*, 2019; Burnett *et al.*, 2021). Partial least squares regression was designed to address both the collinearity between predictors, that is the different wavebands of a reflectance spectrum and the large number of predictor variables when compared with trait observations. This study implemented PLSR modelling using the PYTHON package SCIKIT-LEARN v.0.24 for chlorophyll and sugar content estimations. The optimal number of PLSR components was first determined based on five-fold cross-validated R^2 scores of PLSR models fitted on the training dataset with the number of components set to between 1 and 20 (Fig. S1). A PLSR model was then fitted to the training dataset using the optimum number of components ($n = 6$ for chlorophyll and $n = 7$ for sugar; Fig. S1) and used to predict target values (chlorophyll or sugar content) for the test dataset.

Results

Automated hyperspectral vegetation index for chlorophyll content estimation

Automated hyperspectral vegetation index performance was measured using R^2 scores generated using SLR models on the test dataset with the respective AutoVI-derived indices as features. Automated hyperspectral vegetation index performance across five repetitions was relatively stable, with the grouped evaluation

strategy allowing for comparison across different model groups and identification of the best performing index model within each model group (Fig. 3). Between model groups, the M4 group (N_{wb} = 4) had the best mean R^2 of 0.7818, followed by M5 (N_{wb} = 5) with R^2 of 0.7747, M6 (N_{wb} = 6) with R^2 of 0.7689, M2 (N_{wb} = 2) with R^2 of 0.7637 and finally M3 (N_{wb} = 3) with R^2 of 0.7555. Overall, Index25 (N_{wb} = 4, N_{cf} = 1) produced the best VI (R^2 = 0.8007) and generated the best results across all five repetitions within the M4 group (Fig. 3). The performance of the best VIs according to model group is summarised in Table 1. Wavebands selected by the AutoVI for the best chlorophyll indices were derived predominantly from the red (600–720 nm) and red-edge (RE; 720–780 nm) regions, with a few wavebands from the blue (470–490 nm), NIR (1000–1300 nm) and SWIR (1600–1700 nm) regions (Table 1). The R^2 score produced by the best VI, termed from this point forwards as the AutoVI chlorophyll index (AutoVI-Chl), was achieved using wavebands at 610, 716, 1384 and 1607 nm, without coefficient tuning (i.e. value set to 1), as depicted in Eqn 1:

$$\text{AutoVI} - \text{Chl} = \frac{(R_{1607} - R_{1384}) - (R_{1607} - R_{716})(R_{1607}/R_{1384})}{2 \times ((R_{610} - R_{1384})/(R_{610} + R_{1384} + 1))} \quad \text{Eqn 1}$$

(R_{wb}, reflectance measured at a discrete waveband (wb)).

The performance of the AutoVI-Chl with or without a coefficient variable, alpha (α), as depicted in its original equation (Index25; Table S2) was determined across five computational repetitions of 20 000 and 40 000 iterations (Fig. 4). At 20 000 iterations, the inclusion of the coefficient had minimal impact on the AutoVI-Chl performance, as boxplots for R^2 scores obtained with the coefficient (min = 0.7657, median = 0.7771, max = 0.7959) or without the coefficient (min = 0.7673, median = 0.7875, max = 0.7922) were comparable (Fig. 4). However, AutoVI computational time when the coefficient was included (*c.* 2.1 h for five repetitions) was up to 1.6× higher than without the coefficient (*c.* 1.3 h for five repetitions), suggesting that the inclusion of coefficient(s) in AutoVI optimisations were likely to incur computational costs. At 40 000 iterations, AutoVI-Chl performance deteriorated significantly with or without the coefficient (Fig. 4), suggesting that overfitting, in which a model performs significantly better on the training (i.e. overfitted) but not the test dataset (i.e. unable to generalise to new data), may be a concern with longer AutoVI runs. The effect of coefficients and longer AutoVI runs will need to be determined for individual target traits.

The quality of AutoVI-derived indices for chlorophyll content estimation was evaluated further against 47 published VIs, as features in SLR modelling. First, the best SLR model resulting from the AutoVI indices and the best SLR model with existing VIs were compared (Table 2). The model with the AutoVI-Chl (R^2 = 0.8007, RMSE = 38.52, MAE = 30.51, MAPE = 4.69%) significantly outperformed the model with the normalised difference chlorophyll index (NDCI; R^2 = 0.6018, RMSE = 54.45, MAE = 46.05, MAPE = 7.09%) (Tables 2, S3). Next, SMLR models using the optimum subset of features selected from AutoVI indices and existing VIs were compared (Table 2). For the existing VIs, SMLR with seven VIs selected (Table S4) led to a significant improvement in model performance (R^2 = 0.7136, RMSE = 46.17, MAE = 38.12, MAPE = 5.88%) compared with SLR and NDCI, but was still inferior to SLR with the AutoVI-Chl; SMLR with four AutoVI indices selected (Table S5) did not perform better (R^2 = 0.7989, RMSE = 38.69, MAE = 30.98, MAPE = 4.88%) compared with SLR and AutoVI-Chl. Finally, PLSR modelling performance for chlorophyll estimation was included as an additional benchmark for comparison. The PLSR model (R^2 = 0.7379, RMSE = 44.17, MAE = 33.81, MAPE = 5.36%) did not perform as well as SLR with AutoVI-Chl or SMLR with the selected AutoVI indices (Table 2). Overall, the best modelling performance was provided using

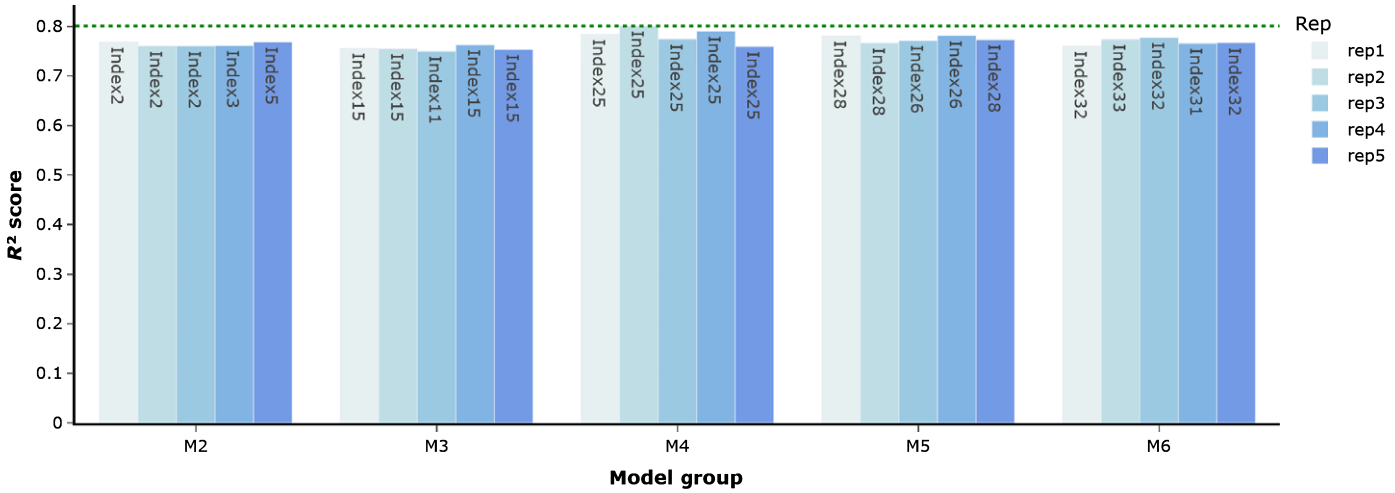

Fig. 3 Best performing index models from grouped model evaluations for total chlorophyll content estimation. The best index models from grouped (M2, M3, M4, M5 and M6) model evaluations were identified across five repeated automated hyperspectral vegetation index (AutoVI) computations. Index model names are indicated within the bar figure. The overall best index model was Index25 from the M4 group with objective function score R^2 of 0.8007 (indicated by the green dashed line).

Table 1 Performance of the best AutoVI-derived indices according to model group for chlorophyll estimation.

Group	Index model	Selected wavelengths (nm)	R^2	RMSE	MAE	MAPE
M2	2	637, 711	0.7683	41.52	33.78	5.14%
M3	15	607, 716, 1712	0.7629	42.01	34.27	5.23%
M4	25	610, 716, 1384, 1607	**0.8007**	**38.52**	**30.51**	**4.69%**
M5	26	497, 650, 707, 778, 1017	0.7813	40.35	32.74	5.13%
M6	32	477, 635, 713, 1033, 1035, 1668	0.7776	40.69	33.01	5.03%

Performance metrics calculated for simple linear regression using the indicated AutoVI-derived index for chlorophyll estimation on the test dataset. The best scores are indicated in bold. MAE, mean absolute error; MAPE, mean absolute percentage error; R^2, objective function score; RMSE, root mean square error.

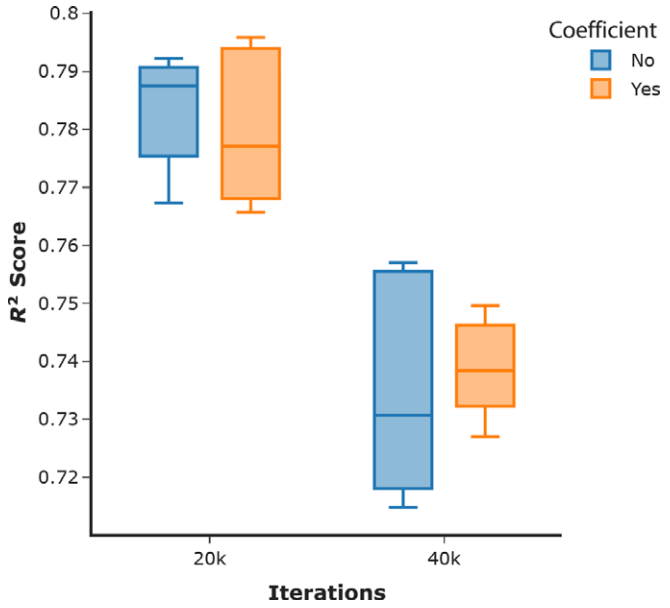

Fig. 4 Effect of coefficient tuning and longer iterations on the automated hyperspectral vegetation index-chlorophyll index (AutoVI-Chl) performance. The inclusion (Yes) or exclusion (No) of a single coefficient, α, on AutoVI-Chl performance was determined across five repeated computations at 20 000 (20k) and 40 000 (40k) iterations each. The distributions of objective function scores R^2 are shown in the boxplot. Error bars are the 95% confidence intervals, the lines inside the boxes indicate the median.

SLR with AutoVI-Chl. These results supported the AutoVI as an efficient system for novel trait-specific hyperspectral VI derivation.

Automated hyperspectral vegetation index for sugar content estimation

Automated hyperspectral vegetation index performance across five repetitions was relatively stable, with the M6 group having the best mean R^2 of 0.8201, followed by the M3 group with R^2 of 0.8127, the M4 group with R^2 of 0.7933, the M5 group with R^2 of 0.7877 and finally the M2 group with R^2 of 0.7591 (Fig. 5). The overall best VI was produced by Index33 (N_{wb} = 6, N_{cf} = 1), which also generated the best results for three repetitions within the M6 group (Fig. 5). The performance of the best VIs according to model group is summarised in Table 3. Wavebands selected by the AutoVI for the best sugar indices were derived predominantly from the SWIR (1400–1700 nm) and NIR (770–1370 nm) regions, with a few bands from the VIS (499–644 nm) region (Table 3). The R^2 score produced by the best VI, termed from this point forwards as the AutoVI sugar index (AutoVI-Sgr), was achieved using wavebands at 499, 773, 1179, 1291, 1425 and 1661 nm, without coefficient tuning (i.e. value set to 1), as depicted in Eqn 2:

$$\text{AutoVI} - \text{Sgr} = \frac{(R_{773} - R_{1425}) - (R_{773} - R_{1179})(R_{773}/R_{1425})}{2 \times ((R_{1661} - R_{1425})/(R_{1661} + R_{1425} + R_{499} + R_{1291} + 1))} \quad \text{Eqn 2}$$

(R_{wb}, reflectance measured at a discrete waveband (wb)).

The inclusion of a coefficient, alpha (α), in the AutoVI-Sgr as depicted in its original equation (Index33; Table S2) and longer optimisations at 40 000 iterations did not significantly improve its performance (Fig. 6). For the longer runs (40k iterations) with or without the coefficient, performance appeared to converge closer and plateau at an R^2 value of *c.* 0.83, suggesting that this may be the maximum performance produced by the underlying model equation (Fig. 6). Future studies using novel index equations may yield further performance enhancement. Based on these results, the recommended starting point for AutoVI training was up to 20 000 iterations and without coefficient tuning.

The quality of AutoVI-derived indices for sugar content estimation was evaluated further against 47 published VIs, as features in SLR modelling. The best SLR performance with the AutoVI-derived indices was achieved using the AutoVI-Sgr (R^2 = 0.8339, RMSE = 2612.41, MAE = 2148.15, MAPE = 20.17%), which significantly outperformed the best SLR model achieved with

Table 2 Comparison between different regression models for chlorophyll estimation.

Model	Feature(s)	R^2	RMSE	MAE	MAPE
SLR	AutoVI-Chl	**0.8007**	**38.52**	**30.51**	**4.69%**
SLR	NDCI	0.6018	54.45	46.05	7.09%
SMLR	Seven existing VIs	0.7136	46.17	38.12	5.88%
SMLR	Four AutoVI indices	0.7989	38.69	30.98	4.88%
PLSR	Reflectance values	0.7379	44.17	33.81	5.36%

Performance metrics calculated for simple linear regression (SLR) and stepwise multiple regression (SMLR) using AutoVI-derived indices or existing 47 vegetation indices (VI), in addition to partial least squares regression (PLSR) using the full spectrum of reflectance values for chlorophyll estimation on the test dataset. The best scores are indicated in bold. AutoVI-Chl, AutoVI chlorophyll index; MAE, mean absolute error; MAPE, mean absolute percentage error; NDCI, normalised difference chlorophyll index; R^2, objective function score; RMSE, root mean square error.

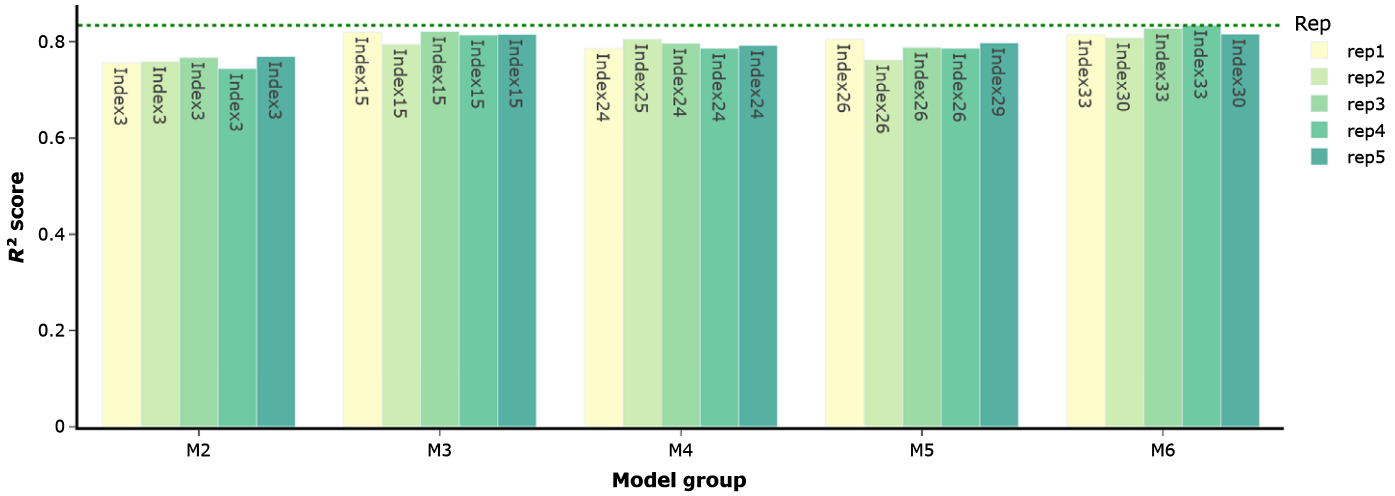

Fig. 5 Best performing index models from grouped model evaluations for total sugar content estimation. The best index models from grouped (M2, M3, M4, M5 and M6) model evaluations were identified across five repeated automated hyperspectral vegetation index (AutoVI) computations. Index model names are indicated within the bar figure. The overall best index model was Index33 from the M6 group with an objective function score R^2 of 0.8339 (indicated by green dashed line).

Table 3 Performance of the best AutoVI-derived indices according to model group for sugar estimation.

Group	Index model	Selected wavelengths (nm)	R^2	RMSE	MAE	MAPE
M2	3	1420, 1588	0.7693	3078.89	2493.21	22.94%
M3	15	1271, 1417, 1650	0.8209	2712.72	2237.33	20.31%
M4	24	1371, 1422, 1650, 1712	0.8055	2827.07	2466.19	23.57%
M5	29	644, 990, 1406, 1418, 1649	0.7860	2964.88	2387.41	21.66%
M6	33	499, 773, 1179, 1291, 1425, 1661	**0.8339**	**2612.41**	**2148.15**	**20.17%**

Performance metrics calculated for simple linear regression using the indicated AutoVI-derived index for chlorophyll estimation on the test dataset. The best scores are indicated in bold. MAE, mean absolute error; MAPE, mean absolute percentage error; R^2, objective function score; RMSE, root mean square error.

the published VI, Gitelson & Merzlyak Index 2 (GMI2) (R^2 = 0.4695, RMSE = 4668.35, MAE = 3939.59, MAPE = 38.96%). In general, SLR modelling performance with existing VIs was very poor (Table S6). Stepwise multiple linear regression with five existing VIs selected (Table S7) produced dramatically better results (R^2 = 0.7387, RMSE = 3276.50, MAE = 2401.14, MAPE = 22.30%) compared with SLR and GMI2 but was still inferior to the model with the AutoVI-Sgr (Table 4). Conversely, SMLR with four AutoVI indices selected (Table S8) performed better (R^2 = 0.8587, RMSE = 2409.19, MAE = 2071.47, MAPE = 19.16%) compared with SLR and AutoVI-Sgr (Table 4). Partial least squares regression modelling performance for sugar estimation was also included as a benchmark for comparison. The PLSR model produced a similar performance (R^2 = 0.8322, RMSE = 2625.99, MAE = 2212.89, MAPE = 21.19%) as SLR with the AutoVI-Sgr but was outperformed by SMLR with the AutoVI indices (Table 4). These results further supported the AutoVI as an efficient system for novel VI derivation, with the AutoVI indices as high-quality features for trait prediction.

Discussion

In this study, we describe the design and implementation of an automated system for hyperspectral vegetation index derivation (AutoVI), for plant phenotyping using as examples chlorophyll and sugar content estimations in wheat. In both cases, indices generated by the AutoVI significantly outperformed existing VIs in simple and multiple linear regression modelling, with performance exceeding that of a more complex model such as PLSR. The success of the AutoVI can be attributed to the use of a highly performing hyperparameter optimisation algorithm, TPE (Bergstra *et al.*, 2011; Akiba *et al.*, 2019).

Results in this study showed that the AutoVI was able to efficiently generate high-quality two-band to six-band indices specific for chlorophyll and sugar contents. A review of the existing literature highlighted that most VIs currently deployed consisted predominantly of two-band indices, and to a lesser extent three-band indices (Henrich *et al.*, 2017). Indices with four bands or more are rare, presumably due to the difficulty in optimising such indices (e.g. band selection), which has been attributed to the curse of high dimensionality and issues associated with the multicollinearity of bands. Although a myriad of hyperspectral band selection strategy exists (Sun & Du, 2019), most of these were explicitly developed for image classification or regression models. Previous studies that focused on generating novel hyperspectral VIs used single or multiple correlation matrices between VI pairs and the trait of interest to uncover new band or index combinations (Thenkabail *et al.*, 2004; Aasen *et al.*, 2014; Xu *et al.*,

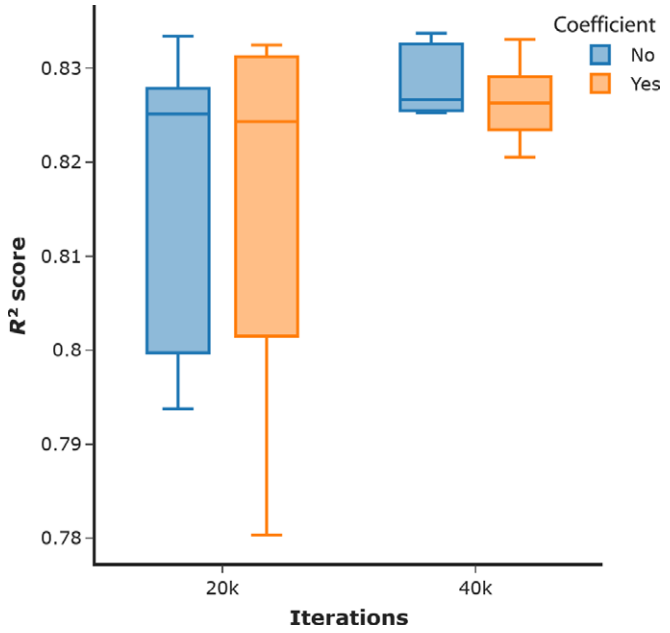

Fig. 6 Effect of coefficient tuning and longer iterations on automated hyperspectral vegetation index-sugar index (AutoVI-Sgr) performance. The inclusion (Yes) or exclusion (No) of a single coefficient, α, on AutoVI-Sgr performance was determined across five repeated computations at 20 000 (20k) and 40 000 (40k) iterations each. The distributions of objective function scores R^2 are shown in the boxplot. Error bars are the 95% confidence interval, the lines inside the boxes indicate the median.

Table 4 Comparison between different regression models for sugar estimation.

Model	Feature(s)	R^2	RMSE	MAE	MAPE
SLR	AutoVI-Sgr	0.8339	2612.41	2148.15	20.17%
SLR	GMI2	0.4695	4668.95	3939.59	38.96%
SMLR	Five existing VIs	0.7387	3276.50	2401.14	22.30%
SMLR	Four AutoVI indices	**0.8587**	**2409.19**	**2071.47**	**19.16%**
PLSR	Reflectance values	0.8322	2625.99	2212.89	21.19%

Performance metrics calculated for simple linear regression (SLR) and stepwise multiple regression (SMLR) using AutoVI-derived indices or existing 47 vegetation indices (VI), in addition to partial least squares regression (PLSR) using the full spectrum of reflectance values for sugar estimation on the test dataset. The best scores are indicated in bold. AutoVI-Sgr, AutoVI sugar index; MAE, mean absolute error; MAPE, mean absolute percentage error; R^2, objective function score; RMSE, root mean square error.

2019). However, these approaches are computationally expensive as every possible combination of available bands (filtered or unfiltered) needs to be computed. Therefore, previous efforts have focused only on limited combinations between a few bands (two-band and three-band indices), therefore also limiting the net performance of the derived VIs. Consequently, the limited availability of specific VIs as biomarkers for plant traits has forced researchers to generalise upon the applicability of existing VIs, for example the NDVI (two-band index), across almost all aspects of research and analysis, leading to suboptimal results. The AutoVI system is well positioned to address this issue through the rapid generation of trait-specific novel two-band to six-band VIs without requiring band filtering or dimension reduction techniques to limit the number of input hyperspectral bands before processing.

Biochemical constituents in plants absorb electromagnetic energy in specific wavelength regions. Vegetation indices profiled around these characteristic spectral absorption regions can detect or estimate the biochemical trait of interest. The AutoVI system was constructed to automate the identification of these critical spectral regions using the underlying TPE algorithm. For chlorophyll, the selected wavebands centred around the red (600–700 nm) and RE (700–740 nm) regions, with a few bands from the blue (470–490 nm), NIR (1000–1300 nm) and SWIR (1600–1700 nm) regions. Chlorophylls (chlorophyll *a* and *b*) are the most important plant pigments that function as photoreceptors and catalysts for photosynthesis, the photochemical synthesis of carbohydrates (Blackburn, 2006). As such, chlorophyll content in leaves and canopies is a key indicator of physiological measures such as photosynthetic capacity, developmental stage, productivity and stress (Richardson *et al.*, 2002; Murchie & Lawson, 2013). Studies have shown that the reflectance of wavelengths in the red region (*c.* 530–630 nm and a narrower band *c.* 700 nm) is most sensitive to chlorophyll pigment concentrations across the normal range found in most leaves and canopies (Lichtenthaler *et al.*, 1996; Gitelson *et al.*, 2005). In addition, research has also shown that bands within the RE region (680–740 nm), which delineates the border between chlorophyll absorption in red wavelengths and leaf scattering in the NIR wavelengths, are strongly correlated with chlorophyll content (Curran *et al.*, 1991; Gitelson *et al.*, 1996). Existing chlorophyll VIs consist mainly of two-band indices derived from the ratios of narrow bands within regions of the spectrum that are sensitive to chlorophyll pigments (VIS–RE, 400–740 nm) and those areas not sensitive to the pigments and/or related to some other control on reflectance (typically NIR, 750–900 nm) (Blackburn, 2006; Wu *et al.*, 2008). Wavebands selected by the AutoVI agree with published studies, and additional wavebands selected from the SWIR region were likely to enhance the sensitivity of some of the AutoVI indices to chlorophyll by acting as a control on reflectance.

The sugar metabolic pathway is intrinsically linked with the regulation of plant growth and development, and response to stress (Julius *et al.*, 2017; Kaur *et al.*, 2021). Studies have shown that abiotic stress such as drought or heat triggers sugar accumulation, particularly soluble sugars such as sucrose in plants (Lemoine *et al.*, 2013; Zhou *et al.*, 2017). In this study, the AutoVI generated novel indices with a strong correlation to total sugar content, with specific wavebands selected mainly from the SWIR (1400–1700 nm) and NIR (770–1370 nm) regions. It is known that leaf reflectance properties in the SWIR region (1300–2500 nm) are governed by water content and biochemical compounds such as cellulose, sugars and starch (Elvidge, 1990; Kokaly *et al.*, 2009). Indeed, more recent studies have used NIR spectroscopy (750–2500 nm) coupled with machine learning and/or statistical models to estimate soluble carbohydrates, including total sugar in various plant tissues and organs such as leaf and stem (Adams *et al.*, 2013; Piaskowski *et al.*, 2016).

Research in rice has also identified wavebands in the NIR region (800–1100 nm) as being important for sugar content estimation in leaves (Das *et al.*, 2018). These studies provide support for bands selected by the AutoVI as being specific for total sugar content. Models for sugar estimation in this study were less robust overall (e.g. MAPE = 19.16–22.30%) compared with models for chlorophyll estimation (e.g. MAPE = 4.69–5.23%) as reflected in the higher error scores. This may be due to difficulties in obtaining high-quality spectral signatures from the SWIR region as water vapour is known to obscure spectral signatures for biochemical compounds in this region in plants (Elvidge, 1990; Kokaly *et al.*, 2009). Nevertheless, it is noteworthy that the AutoVI was able to select bands in the SWIR region, whilst avoiding the water absorption peak at 1450 nm (Elvidge, 1990; Kokaly *et al.*, 2009). Moving forwards, further biological or physiological association of specific wavebands detected by the AutoVI can be studied on individual traits under varied environments.

Novel AutoVI-derived indices for chlorophyll and sugar content estimations were first compared against 47 published VIs as features in SLR. This provided a good baseline to compare VIs, as model performance would depend solely on the quality of the VI. For chlorophyll estimation, the best SLR was achieved with the AutoVI-derived index, AutoVI-Chl (R^2 = 0.8007), which significantly outperformed the best SLR with the published VI, NDCI (R^2 = 0.6018). Similarly, for sugar estimation, the best performance provided using SLR with AutoVI-Sgr (R^2 = 0.8339) significantly outperformed the SLR with the published VI, GMI2 (R^2 = 0.4695). Next, the performance of an SMLR with features selected from existing VIs was compared with the SLR models with AutoVI indices. Results showed that although SMLR did significantly improve performance for both chlorophyll (R^2 = 0.7136) and sugar (R^2 = 0.7387) estimations with published VIs, these were still inferior compared with SLR results achieved using the AutoVI indices. Stepwise multiple linear regression was also performed using the AutoVI indices, with performance enhancement observed for sugar (R^2 = 0.8587) but not chlorophyll (R^2 = 0.7989) estimations. Finally, model performance for chlorophyll and sugar estimations using the AutoVI indices were compared with results produced using PLSR, a well established method for plant trait modelling using hyperspectral data (Ely *et al.*, 2019; Wu *et al.*, 2019; Burnett *et al.*, 2021). Partial least squares regression represents a different modelling approach, as it projects the entire spectrum of reflectance values (or the predictor variables) into a smaller number of variables (or components), whilst simultaneously maximising the correlation between the response and the variables (Geladi & Kowalski, 1986; Wold *et al.*, 2001). With PLSR, the quality of feature representation or learning is being compared with the models with AutoVI indices, as depicted using the projected components in PLSR, in contrast with the functions encoded in the AutoVI indices. For both chlorophyll and sugar estimations, SLR and SMLR with AutoVI indices outperformed PLSR (R^2 = 0.7379 for chlorophyll; R^2 = 0.8322 for sugar). Impressively, SLR with the best AutoVI-derived index (AutoVI-Chl or AutoVI-Sgr) produced better results and outperformed more complex approaches such as SMLR (except for sugar estimation) and PLSR. These results provided strong support for AutoVI-derived indices as high-quality features and affirmed the AutoVI as a high-performing system for novel trait-specific hyperspectral VI derivation.

However, AutoVI indices, including AutoVI-Chl and AutoVI-Sgr, are not perfected VIs and will benefit from further studies incorporating more data collected from multiple genotypes across different environments and growth stages. Particularly for field-based phenotyping, AutoVI indices should ideally be derived using canopy spectra, as these are known to differ from leaf spectra (Croft *et al.*, 2014). Depending on the underlying trait, it may be worthwhile collecting samples from different plant tissues or organs. For example, studies have shown that genotypes that accumulate more sugar in leaf and stem tissues are more heat and/or drought tolerant (Piaskowski *et al.*, 2016; Zhou *et al.*, 2017). Fortunately, the AutoVI does not impose any size or dimensional constraint on the input data and is expected to work with data derived from different hyperspectral sensors and/or spectra sources. Depending on the data provided, the AutoVI can be customised to deliver trait-specific VIs according to species, genotype, growth stage and environment, making it a powerful and versatile tool for both novice and expert users alike. In addition, compared with machine learning and statistical modelling approaches, in which model optimisation and deployment remains technically challenging, AutoVI-derived indices can be easily computed and readily deployed for HTP without requiring complex hardware or software resources.

In conclusion, results in our study demonstrated that the AutoVI is an efficient and powerful tool for deriving high-quality hyperspectral VIs for plant phenotyping. Automated hyperspectral vegetation index-derived VIs outperformed existing VIs and delivered strong performance in SLR and SMLR modelling for chlorophyll and sugar content estimations, producing results superior to PLSR. The AutoVI system is expected to accelerate the development of novel VIs for plant/crop traits that will find wide application in HTP and agriculture remote sensing, vital to improving breeding programme and crop management efficiencies.

Author contributions

JCOK and BPB designed the AutoVI system and analysed the data. JCOK implemented the AutoVI system and GUI version in Python. BPB researched and collated the list of index models used in the AutoVI. JCOK, BPB, GS and SK conceived and designed the study. JCOK, BPB, GS and SK wrote and edited the paper. All authors read and approved the final manuscript. JCOK and BPB contributed equally to this work.

ORCID

Bikram P. Banerjee https://orcid.org/0000-0002-5542-3751
Surya Kant https://orcid.org/0000-0001-6178-7036
Joshua C.O. Koh https://orcid.org/0000-0002-3678-4789
German Spangenberg https://orcid.org/0000-0002-1656-3364

Data availability

All relevant source codes and datasets, including a GUI implementation of the AutoVI in Python for Windows 10 64-bit operating system, required to reproduce the results reported in this study are available at www.github.com/AVR-SmartSense/AutoVI.

References

Aasen H, Gnyp ML, Miao Y, Bareth G. 2014. Automated hyperspectral vegetation index retrieval from multiple correlation matrices with hypercor. *Photogrammetric Engineering & Remote Sensing* **80**: 785–795.

Aasen H, Honkavaara E, Lucieer A, Zarco-Tejada PJ. 2018. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: a review of sensor technology, measurement procedures, and data correction workflows. *Remote Sensing* **10**: 1091.

Adams HD, Germino MJ, Breshears DD, Barron-Gafford GA, Guardiola-Claramonte M, Zou CB, Huxman TE. 2013. Nonstructural leaf carbohydrate dynamics of *Pinus edulis* during drought-induced tree mortality reveal role for carbon metabolism in mortality mechanism. *New Phytologist* **197**: 1142–1151.

Adão T, Hruška J, Pádua L, Bessa J, Peres E, Morais R, Sousa JJ. 2017. Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing* **9**: 1110.

Akiba T, Sano S, Yanase T, Ohta T, Koyama M. 2019. Optuna: a next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

Bajwa SG, Kulkarni SS. 2011. *Hyperspectral data mining*. Boca Raton, FL, USA; London, UK; New York, NY, USA: CRC Press/Taylor and Francis Group.

Banerjee BP, Joshi S, Thoday-Kennedy E, Pasam RK, Tibbits J, Hayden M, Spangenberg G, Kant S. 2020. High-throughput phenotyping using digital and hyperspectral imaging-derived biomarkers for genotypic nitrogen response. *Journal of Experimental Botany* **71**: 4604–4615.

Bergstra J, Bardenet R, Bengio Y, Kégl B. 2011. Algorithms for hyper-parameter optimization. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, eds. *Proceedings of the 24th international conference on neural information processing systems*. Granada, Spain: Curran Associates, 2546–2554.

Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. 2015. Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science & Discovery* **8**: 014008.

Blackburn GA. 2006. Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany* **58**: 855–867.

Burger J, Gowen A. 2011. Data handling in hyperspectral image analysis. *Chemometrics and Intelligent Laboratory Systems* **108**: 13–22.

Burnett AC, Anderson J, Davidson KJ, Ely KS, Lamour J, Li Q, Morrison BD, Yang D, Rogers A, Serbin SP. 2021. A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression. *Journal of Experimental Botany* **72**: 6175–6189.

Chauhan HJ, Mohan BK. 2013. Development of agricultural crops spectral library and classification of crops using hyperion hyperspectral data. *Journal of Remote Sensing Technology* **1**: 9.

Chein IC. 1999. *Spectral information divergence for hyperspectral image analysis. IEEE 1999 international geoscience and remote sensing symposium, vol. 501. IGARSS'99* (cat. no.99CH36293), 509–511. doi: 10.1109/IGARSS.1999.773549.

Croft H, Chen JM, Zhang Y. 2014. The applicability of empirical vegetation indices for determining leaf chlorophyll content over different leaf and canopy structures. *Ecological Complexity* **17**: 119–130.

Curran PJ, Dungan JL, Macler BA, Plummer SE. 1991. The effect of a red leaf pigment on the relationship between red edge and chlorophyll concentration. *Remote Sensing of Environment* **35**: 69–76.

Das B, Sahoo RN, Pargal S, Krishna G, Verma R, Chinnusamy V, Sehgal VK, Gupta VK, Dash SK, Swain P. 2018. Quantitative monitoring of sucrose, reducing sugar and total sugar dynamics for phenotyping of water-deficit stress tolerance in rice through spectroscopy and chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **192**: 41–51.

Donoho DL. 2000. *High-dimensional data analysis: the curses and blessings of dimensionality*. [WWW document] URL https://fdocuments.in/document/high-dimensional-data-analysis-the-curses-and-blessings-of-.html [accessed 29 September 2021].

DuBois M, Gilles KA, Hamilton JK, Rebers PA, Smith F. 1956. Colorimetric method for determination of sugars and related substances. *Analytical Chemistry* **28**: 350–356.

Elvidge CD. 1990. Visible and near infrared reflectance characteristics of dry plant materials. *International Journal of Remote Sensing* **11**: 1775–1795.

Ely KS, Burnett AC, Lieberman-Cribbin W, Serbin SP, Rogers A. 2019. Spectroscopy can predict key leaf traits associated with source–sink balance and carbon–nitrogen status. *Journal of Experimental Botany* **70**: 1789–1799.

Geladi P, Kowalski BR. 1986. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* **185**: 1–17.

Gitelson AA, Merzlyak MN, Lichtenthaler HK. 1996. Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm. *Journal of Plant Physiology* **148**: 501–508.

Gitelson AA, Viña A, Ciganda V, Rundquist DC, Arkebauer TJ. 2005. Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters* **32**: doi: 10.1029/2005GL022688.

Henrich V, Krauss G, Götze C, Sandow C. 2017. *Index database: a database for remote sensing indices*. [WWW document] URL https://www.indexdatabase.de/db/ias.php [accessed 29 September 2021].

Hinneburg A, Keim DA. 1999. Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering. *Proceedings of the 25th VLDB Conference*, Edinburgh, UK. [WWW document] URL http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-70410 [accessed 29 September 2021].

Julius BT, Leach KA, Tran TM, Mertz RA, Braun DM. 2017. Sugar transporters in plants: new insights and discoveries. *Plant and Cell Physiology* **58**: 1442–1460.

Kaur H, Manna M, Thakur T, Gautam V, Salvi P. 2021. Imperative role of sugar signaling and transport during drought stress responses in plants. *Physiologia Plantarum* **171**: 833–848.

Kokaly RF, Asner GP, Ollinger SV, Martin ME, Wessman CA. 2009. Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sensing of Environment* **113**: S78–S91.

Lemoine R, Camera SL, Atanassova R, Dédaldéchamp F, Allario T, Pourtau N, Bonnemain J-L, Laloi M, Coutos-Thévenot P, Maurousset L *et al.* 2013. Source-to-sink transport of sugar and regulation by environmental factors. *Frontiers in Plant Science* **4**: 272.

Lewitt RM. 1990. Multidimensional digital image representations using generalized Kaiser–Bessel window functions. *Journal of the Optical Society of America* **7**: 1834–1846.

Lichtenthaler HK. 1987. Chlorophylls and carotenoids: pigments of photosynthetic biomembranes. *Methods in Enzymology* **148**: 350–382.

Lichtenthaler HK, Gitelson A, Lang M. 1996. Non-destructive determination of chlorophyll content of leaves of a green and an aurea mutant of tobacco by reflectance measurements. *Journal of Plant Physiology* **148**: 483–493.

Lu B, Dao PD, Liu J, He Y, Shang J. 2020. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing* **12**: 2659.

Mir RR, Reynolds M, Pinto F, Khan MA, Bhat MA. 2019. High-throughput phenotyping for crop improvement in the genomics era. *Plant Science* **282**: 60–72.

Mishra P, Asaari MSM, Herrero-Langreo A, Lohumi S, Diezma B, Scheunders P. 2017. Close range hyperspectral imaging of plants: a review. *Biosystems Engineering* **164**: 49–67.

Murchie EH, Lawson T. 2013. Chlorophyll fluorescence analysis: a guide to good practice and understanding some new applications. *Journal of Experimental Botany* **64**: 3983–3998.

Pearson RL, Miller LD. 1972. Remote mapping of standing crop biomass for estimation of the productivity of the shortgrass prairie, Pawnee National Grasslands, Colorado. *Proceedings of the eighth international symposium on*

remote sensing of environment, Ann Arbor, MI, USA. [WWW document] URL https://eurekamag.com/research/000/179/000179997.php [accessed 29 September 2021].

Piaskowski JL, Brown D, Campbell KG. 2016. Near-infrared calibration of soluble stem carbohydrates for predicting drought tolerance in spring wheat. *Agronomy Journal* **108**: 285–293.

Rao NR, Garg PK, Ghosh SK. 2007. Development of an agricultural crops spectral library and classification of crops at cultivar level using hyperspectral data. *Precision Agriculture* **8**: 173–185.

Richardson AD, Duigan SP, Berlyn GP. 2002. An evaluation of noninvasive methods to estimate foliar chlorophyll content. *New Phytologist* **153**: 185–194.

Silleos NG, Alexandridis TK, Gitas IZ, Perakis K. 2006. Vegetation indices: advances made in biomass estimation and vegetation monitoring in the last 30 years. *Geocarto International* **21**: 21–28.

Sun W, Du Q. 2019. Hyperspectral band selection: a review. *IEEE Geoscience and Remote Sensing Magazine* **7**: 118–139.

Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. 2017. Plant phenomics, from sensors to knowledge. *Current Biology* **27**: R770–R783.

Thenkabail PS, Enclona EA, Ashton MS, Van Der Meer B. 2004. Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. *Remote Sensing of Environment* **91**: 354–376.

Winston PH. 1992. *Artificial intelligence, 3rd edn*. Reading, MA, USA: Addison-Wesley Longman.

Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**: 109–130.

Wu C, Niu Z, Tang Q, Huang W. 2008. Estimating chlorophyll content from hyperspectral vegetation indices: modeling and validation. *Agricultural and Forest Meteorology* **148**: 1230–1241.

Wu J, Rogers A, Albert LP, Ely K, Prohaska N, Wolfe BT, Oliveira RC Jr, Saleska SR, Serbin SP. 2019. Leaf reflectance spectroscopy captures variation in carboxylation capacity across species, canopy environment and leaf age in lowland moist tropical forests. *New Phytologist* **224**: 663–674.

Xu M, Liu R, Chen JM, Liu Y, Shang R, Ju W, Wu C, Huang W. 2019. Retrieving leaf chlorophyll content using a matrix-based vegetation index combination approach. *Remote Sensing of Environment* **224**: 60–73.

Xue J, Su B. 2017. Significant remote sensing vegetation indices: a review of developments and applications. *Journal of Sensors* **2017**: 1353691.

Yao X, Liu Y. 1997. Fast evolution strategies. In: Angeline PJ, Reynolds RG, McDonnell JR, Eberhart R, eds. *International conference on evolutionary programming, lecture notes in computer science, vol 1213*. Berlin/Heidelberg, Germany: Springer, 149–161.

Yu T, Zhu H. 2020. *Hyper-parameter optimization: a review of algorithms and applications*. [WWW document] URL https://arxiv.org/abs/2003.05689 [accessed 29 September 2021].

Zhou R, Kjær KH, Rosenqvist E, Yu X, Wu Z, Ottosen C-O. 2017. Physiological response to heat stress during seedling and anthesis stage in tomato genotypes differing in heat tolerance. *Journal of Agronomy and Crop Science* **203**: 68–80.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Optimisation of number of components in partial least squares regression (PLSR) for chlorophyll and sugar estimation.

Table S1 List of 47 published vegetation indices.

Table S2 Index models for automated hyperspectral vegetation index (AutoVI) selection.

Table S3 Performance of 47 published vegetation indices for chlorophyll content estimation.

Table S4 Stepwise multiple linear regression with 47 published vegetation indices for chlorophyll estimation.

Table S5 Stepwise multiple linear regression with automated hyperspectral vegetation index (AutoVI)-derived indices for chlorophyll estimation.

Table S6 Performance of 47 published vegetation indices for sugar content estimation.

Table S7 Stepwise multiple linear regression with 47 published vegetation indices for sugar estimation.

Table S8 Stepwise multiple linear regression with automated hyperspectral vegetation index (AutoVI)-derived indices for sugar estimation.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.