Data driven ensemble learning for soybean yield prediction

Souradeep Chattopadhyay*, Matthew Carroll, Baskar Ganapathysubramanian, Asheesh K. Singh , Soumik Sarkar

Iowa State University
*soura@iastate.edu

Abstract

Soybean yield prediction is a challenging problem in plant breeding that is often affected by many different factors simultaneously. Hyperspectral reflectance data from plants provide breeders with useful data about the health of soybean plants and using this data for yield prediction is an active area of research. Often breeding programs suffer from issues such as data imbalance and several external factors such as genotype variablility in different environments which can pose a serious challenge for developing yield prediction models for large scale breeding programs. In this work we demonstrate a cluster based ensemble approach for yield prediction that can perform well for large scale breeding programs by efficiently harnessing useful information from data through an unsupervised approach.

1 Introduction

Plant breeders strive to select superior genetic lines to meet the needs of farmers, industry and consumers, with seed yield as one of the most important traits under selection. Traditional methods of estimating seed yield consist of machine harvesting plots at the end of the growing season, and using these data to make breeding decisions to select or discard. Furthermore, large population sizes are included, necessitating phenotyping for multiple traits, across numerous locations and several years. Therefore, the plant breeding process is resource and time intensive, and requires significant labor, land and capital to develop new varieties (Singh, Singh, and Singh 2021). To overcome these challenges, plant breeders and scientists have proposed new methods of assessing lines by using the power of remote sensing combined with machine learning to predict yield in season and to reduce the time and labor requirements at harvest (Li et al. 2022; Yoosefzadeh-Najafabadi et al. 2021; Chiozza et al. 2021; Shook et al. 2021; Riera et al. 2021; Guo et al. 2021; Singh et al. 2021). Accurate in season yield prediction is a difficult task due to the complexity of the differences in genotype, variability in macro and micro environments, and other factors which can be difficult to measure and quantify across the growing season. Hyperspectral reflectance data has been shown to be able to capture plant health indicators not observable in the visible light spectrum (Nagasubramanian et al. 2018) and in-season seed yield prediction with rank performance (Parmley et al. 2019b). Hyperspectral data and vegetation indices have a large number of predictors and typically few observed data points due to the complexity of the data machine learning (ML) has shown to be a valuable tool to deal with these complex and often nonlinear prediction problems (Parmley et al. 2019a). However, this is an evolving field and continual improvement are desired to achieve a high level of success in yield prediction for improved decision making by breeders.

Ensemble learning is a powerful ML concept that works by constructing a set of individual models and then combines the output of all the constructed models using a decision fusion strategy to produce a single answer for a given problem. The two main challenges of ensemble learning are: (1) how to select appropriate training data and learning algorithms to make the individual learning tasks as diverse as possible, and (2) how to finalize the learning results by proper fusion of the individual learning tasks (Huang, Xie, and Xiao 2009). Essentially, ensemble learning is able to take into account the effects of the individual learning tasks to provide an efficient solution to the main problem. Since large scale breeding programs are affected by experimental plot design, ensemble learning can help in developing efficient prediction algorithms that can factor in the effect from the plot design to provide yield prediction for large scale scenarios. With this motivation, we develop two ensemble approaches that are able to identify and factor in potential experimental design effects in the modelling paradigm. The first approach called a field ensemble model tries to leverage the experimental design information specific to a breeding program to perform predictions, while the second method called cluster ensemble approach tries to infer the possible effects within the dataset through a unsupervised learning approach and then build a ensemble strategy based on the derived information.

2 Description of Data

Data was collected across two years (2020 and 2021), in preliminary yield trials (PYT), and the advanced yield trials (AYT) each having multiple experiments in each year. Hyperspectral reflectance was collected for each plot using a Thorlabs CCS200 spectrometer (Newton, NJ) using a system similar to that described in (Bai et al. 2016). Each plot

Table 1: Number of datapoints in each of the four fields.

Field No.	Number of observations
1	770
2	912
3	800
4	679

had data from 200 to 1000 nm and was collected at three timepoints each year. Plot level seed yield data was collected for each plot with an Almaco small plot combine (Nevada, IA), and was adjusted to 13% moisture and converted to Kg/ha. Data was collected in blocks based on the layout of the breeding program, which grouped lines based on similar genetic backgrounds and their maturity groups. We report on the third timepoint because these measurments have the highest feature importance values, and as measurments get closer to physiological maturity there is a greater relationship between reflectance data and yield. We grouped data from all the experiments within each test (PYT and AYT) for each year (2020 and 2021) to create four datasets, hereon referred to as fields.

Using the hyperspectral reflectance values we computed fifty-two health indices taken from Li et al. (2022) details of which are given in table 4 in appendix A. The number of data points in each of the four fields are given in table 1.

Before computing the health indices we performed preprocessing on the data to remove outliers from our analysis. The steps taken for preprocessing are as follows:

1. All observations for band values less than 400 nm were removed. This was done since we noticed many anomalies in the readings in those bands.
2. All datapoints which had negative hyperspectral values in any of the bands ranging from 400 nm to 1000 nm were removed.
3. All datapoints which had negative seed yield values were removed.

3 Methodology

In this section, we describe two ensemble approaches: (a) field weighted ensemble model that used four fields described in section 2, and then (b) cluster based ensemble approach, which is developed on the combined data from the four fields.

3.1 Field weighted ensemble model

The ensemble framework described in this section aims to leverage field information in the modelling paradigm by fitting separate models to each of the four fields and then combine predictions from each of the four tasks to arrive at the final prediction. The steps to build the ensemble framework are given below

1. Let $\mathcal{D}_k = (\mathcal{X}_k, \mathcal{Y}_k)$ denote the dataset for the kth field F_k (dataset for kth task), where $\mathcal{X}_k$ denotes the predictors (health indices) and $\mathcal{Y}_k$ the response (seed yield), $k = 1, 2, 3, 4$. Divide $\mathcal{D}_k$ into $\mathcal{D}_k^{Tr} = (\mathcal{X}_k^{Tr}, \mathcal{Y}_k^{Tr})$ and $\mathcal{D}_k^{Te} = (\mathcal{X}_k^{Te}, \mathcal{Y}_k^{Te})$, where $\mathcal{D}_k^{Tr}$ and $\mathcal{D}_k^{Te}$ denote the training and test set for $\mathcal{D}_k$ respectively.
2. Fit a random forest regression model (Breiman 2001) to every $\mathcal{D}_k^{Tr}$.
3. Let $\mathcal{W}_K = (\mathcal{X}_k^{Tr}, Z_k)$, where $Z_k = k$ for every data point in $\mathcal{X}_k^{Tr}$ denote a new dataset. Here Z_k denotes the field membership of $\mathcal{X}_k^{Tr}$. Combine $\mathcal{W}_k$ row wise to create a combined dataset $\mathcal{W}$.
4. Fit a multinomial logistic regression classifier (Abdillah et al. 2020) on $\mathcal{W}$. This classifer provides the ensemble weights for new observations.

Now, let $\mathcal{D}^{Te} = (\mathcal{X}^{Te}, \mathcal{Y}^{Te})$ denote the dataset obtained by combining $\mathcal{D}_k^{Te}$, $k = 1, 2, 3, 4$ row wise.. Then for any datapoint X_i from $\mathcal{X}^{Te}$, let $\hat{y}_{ik}$ denote the prediction obtained from the kth task. Let $(\omega_{i1}, \omega_{i2}, \omega_{i3}, \omega_{i4})$ denote the classification weights for X_i obtained using the multinomial classifier, where ω_{ik} is the probability that X_i belongs to F_k $\left(\sum_{k=1}^{4} \omega_{ik} = 1\right)$. The final prediction of X_i is then obtained as

$$\hat{y}_i = \sum_{k=1}^{4} \omega_{ik} \hat{y}_{ik} \tag{1}$$

3.2 Cluster ensemble model

The cluster ensemble method uses a unsupervised approach to find clusters in the data that serve as datasets for the indivitual tasks. The appeal of this method is that it does not require any additional field information separately, but is able to determine it from the full dataset. The steps to build the cluster ensemble framework are given below:

Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ be a dataset where $\mathcal{X}$ are the predictors (health indices) and $\mathcal{Y}$ is the response (seed yield). In this setting we do not assume any field information about $\mathcal{D}$. The aim here is to first identify possible clusters within $\mathcal{D}$, and then use the obtained cluster information to build our ensemble approach for prediction. For this setup we split the data $\mathcal{D}$ into train and test parts namely $\mathcal{D}^{Tr} = (\mathcal{X}^{Tr}, \mathcal{Y}^{Tr})$ and $\mathcal{D}^{Te} = (\mathcal{X}^{Te}, \mathcal{Y}^{Te})$

1. Divide $\mathcal{X}^{Tr}$ into K homogeneous groups using k-means (MacQueen 1967) and elbow method. Let $\mathcal{C}$ be a variable denoting the cluster memberships of $\mathcal{X}^{Tr}$.
2. Based on $\mathcal{C}$ divide $\mathcal{D}^{Tr}$ into K groups $\mathcal{D}_k^{Tr} = (\mathcal{X}_k^{Tr}, \mathcal{Y}_k^{Tr})$, $k = 1, 2, \ldots, K$. $\mathcal{D}_k^{Tr}$ is hence the dataset for the kth task.
3. Fit a random forest regression model to every $\mathcal{D}_k^{Tr}$, $k = 1, 2, \ldots, K$.
4. Fit a logistic regression classifier to $(\mathcal{X}^{Tr}, \mathcal{C})$ to determine the ensemble weights.

Now for any observation X_i from $\mathcal{X}^{Te}$ let $\hat{y}_{ik}$ denote the prediction from the kth task, $k = 1, 2, \ldots, K$. Let $(\omega_{i1}, \omega_{i2}, \ldots, \omega_{iK})$ be the ensemble weights for X_i obtained using the classifier where ω_{ik} is the probability that the X_i belongs to the kth group $(\sum_{k=1}^{K} \omega_{ik} = 1)$. Then the

final predicted value for X_i is obtained as

$$\hat{y}_i = \sum_{k=1}^{K} \omega_{ik} \hat{y}_{ik} \quad (2)$$

Choice of model for individual tasks: Choice of the appropriate model for each individual task is an important aspect of this framework. We fitted several models for each task. These included linear regression, Ridge regression and Gaussian Process, but random forest regression yielded the best accuracy in terms of R-squared.

Our framework was implemented using the python library scikit-learn (Pedregosa et al. 2011)

4 Results and Discussions

In this section we compare the accuracy of the two approaches, and discuss the usefulness of the cluster ensemble approach for large scale scenarios and potential use in variety development.

4.1 Field weighted ensemble model

For this experiment we performed a eighty-twenty train-test split on each of the four fields. Each individual task was fitted using a 3-fold cross validation and eighty-twenty train-test split. To make sure we have fitted a good classifier to our combined training data we use a eighty-twenty train-test split and a 3-fold cross validation on our combined data (denoted by $\mathcal{W}$ under section 3.1). Our fitted classifier achieved an R^2 of 0.96 indicating that our fitted classifier will be able to generate appropriate ensemble weights for the weighted average prediction

The accuracy of each of the four tasks and the ensemble model for the combined test data are given in table 2. It is evident that the individual field models do not achieve a good accuracy, with R-squared values ranging from -0.19 to 0.20. The field ensemble model is able to achieve a higher accuracy than all of the individual fields. This is because the ensemble weights are able to leverage the contributions of the four individual models for every test data to give an increase in accuracy. Although the increase in this case is not significant, this example is a clear indication that ensembling using the classification weights tends to provide better prediction. The main reason behind this low accuracy of the overall framework is because each of the individual fields have a poor fit and that is mainly because of the low amount of data in each of the fields. We hypothesize that where fields have a sufficient amount of data to provide a good fit the ensemble can result in increased accuracy, and will be future research direction.

4.2 Cluster ensemble model

The first step in implementing the cluster ensemble model is creation of data for the individual tasks, i.e. clustering the predictors in our training set (we used a eight-twenty train-test split on the whole data). Figure 1 shows the values of the sum of squared distances for k (number of clusters) values one to ten. The elbow plot suggest the optimum number of groups in the combined data to be two. The first cluster has

Table 2: Test accuracy (measured by R^2) for fitted random forest model for each of the four fields and the field ensemble model.

Field No.	Accuracy
1	0.20
2	0.02
3	-0.19
4	0.04
Field ensemble	0.26

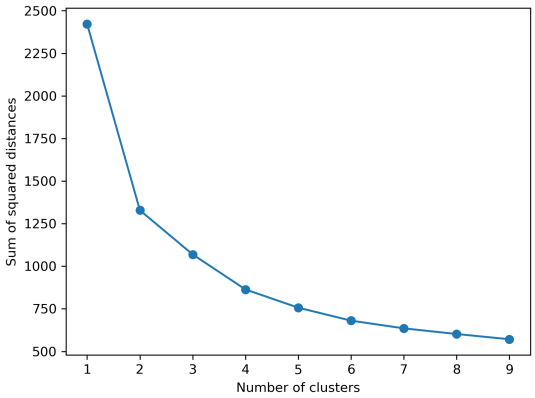
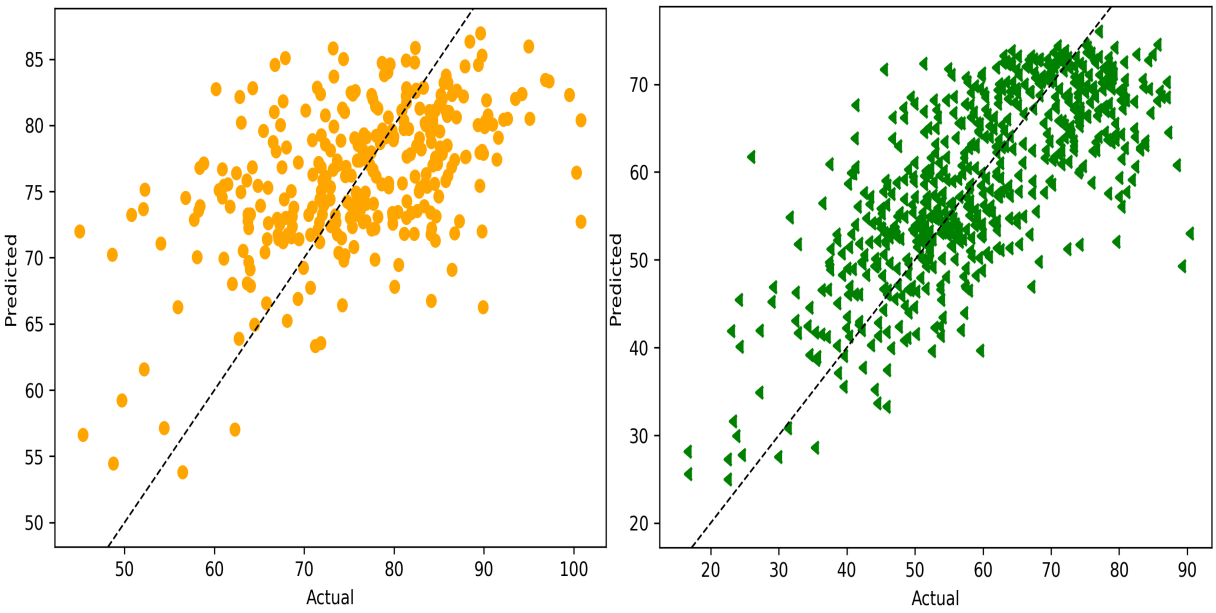


Figure 1: Sum of squared distances for k values 1-9.

2100 observations and the second 428 observations.. To train our classifier and cluster ensemble model we use the same principle as section 4.1. The obtained test accuracy for the classifier is 0.95. In each of the obtained clusters we fitted a random forest regression using a 3-fold cross validation. The test accuracy of the individual models and the task ensemble model are given in table 3. The results of cluster ensemble model points to an important fact that combining data from the different fields appropriately leads to much better accuracy of the individual tasks hence pointing to the fact that the different fields are indeed related in some way. The cluster ensemble model is able to identify such types of combinations efficiently and leverage the individual models through the ensemble weights which enables it to provide predictions with much higher accuracy compared to the field ensemble model. The potential for this type of learning task in a large breeding program could help to overcome problems with class imbalance that can be common in breeding pro-

Table 3: Test accuracy (measured by R^2) for fitted random forest model on the two clusters, the cluster ensemble model.

Cluster No.	Accuracy
1	0.41
2	0.47
Cluster ensemble	0.54



(a) (b)

Figure 2: From left to right: (a) The actual vs predicted plot for field ensemble model, (b) The actual vs predicted plot for cluster ensemble model.

grams due to skewed representation of genetic backgrounds. Genomic and pheomic prediction can be biased with improved prediction performance for large classes, but can fail on smaller subsets of data. Data driven ensemble approaches can help to increase overall accuracy and confidence in predictions made across breeding programs and be more precise in all genetic backgrounds.

To investigate the accuracy of the cluster ensemble model even further we fitted a random forest regression to the full data with the same train-test split as the cluster ensemble model and 3-fold cross validation. The R^2 for the fitted model turned out to be 0.53 which indicates that even though the cluster ensemble approach is a significant improvement over the field ensemble approach it is not significantly better than a simple random forest on the whole data. This we believe happens because the dataset used here is not sufficiently large to produce a significant increase in accuracy for the cluster ensemble model. Larger breeding programs giving rise to large scale datasets can give rise to more clusters each with sufficient amounts of data which can increase the accuracy's of the individual tasks which in turn will increase the accuracy of the ensemble model.

5 Conclusions

In this study we demonstrated the effectiveness of ensemble methods, particularly the cluster ensemble methods for soybean yield prediction, using known vegetation indicies. Initial prediction modeling used the raw reflectance data, but yielded poor results. Significant improvements in model accuracies were obtained by using the vegetation indices described in table A. This aligns with previous work that has shown significant improvements by using known indices over raw reflectance data(Parmley et al. 2019a; Li et al. 2022). We find that the cluster ensemble model performs significantly better compared to the field ensemble model even though we do not see any significant improvement over a simple random forest approach over the full dataset. This happens most probably due to the size and nature of the data used for this study but larger datasets can potentially give rise to better accuracy compared to a simple model. Future work in this area includes analyzing if blocking the data based on layout of the breeding program by grouping lines based on similar genetic backgrounds can result in models with better accuracy. We also aim to include other external factors like soil health and weather data along with hyperspectral reflectance to create robust models for yield prediction (Shook et al. 2021).

6 Acknowledgments

The authors thank staff and student members of SinghSoybean group at Iowa State University, particularly Brian Scott, Will Doepke, Jennifer Hicks, Ryan Dunn, and Sam Blair for their assistance with field experiments and phenotyping.

The authors sincerely appreciate the funding support from the Iowa Soybean Association (A.K.S.), North Central Soybean Research Program (A.K.S.), USDA CRIS project IOW04717 (A.K.S., A.S.), AI Institute for Resilient Agriculture (USDA-NIFA #2021-67021-35329) (B.G., S.S., A.K.S., A.S.), COALESCE: COntext Aware LEarning for Sustainable CybEr-Agricultural Systems (CPS Frontier # 1954556) (S.S., A.K.S.), Smart Integrated Farm Networt for Rural Agricultural Communities (SIRAC) (NSF S&CC

1952045) (A.K.S., S.S.), RF Baker Center for Plant Breeding (A.K.S.), and Plant Sciences Institute (A.K.S., S.S.). M.E.C. was partly supported by a graduate assistantship through NSF NRT Predictive Plant Phenomics project.

References

Abdillah, A.; Sutisna, A.; Tarjiah, I.; Fitria, D.; and Widiyarto, T. 2020. Application of Multinomial Logistic Regression to analyze learning difficulties in statistics courses. *Journal of Physics: Conference Series*, 1490(1): 012012.

Bai, G.; Ge, Y.; Hussain, W.; Baenziger, P. S.; and Graef, G. 2016. A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding. *Computers and Electronics in Agriculture*, 128: 181–192.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Chiozza, M. V.; Parmley, K. A.; Higgins, R. H.; Singh, A. K.; and Miguez, F. E. 2021. Comparative prediction accuracy of hyperspectral bands for different soybean crop variables: From leaf area to seed composition. *Field Crops Research*, 271: 108260.

Guo, W.; Carroll, M. E.; Singh, A.; Swetnam, T. L.; Merchant, N.; Sarkar, S.; Singh, A. K.; and Ganapathysubramanian, B. 2021. UAS-Based Plant Phenotyping for Research and Breeding Applications. *Plant Phenomics*, 2021: 9840192.

Huang, F.; Xie, G.; and Xiao, R. 2009. Research on Ensemble Learning. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 3, 249–252.

Li, Z.; Chen, Z.; Cheng, Q.; Duan, F.; Sui, R.; Huang, X.; and Xu, H. 2022. UAV-Based Hyperspectral and Ensemble Machine Learning for Predicting Yield in Winter Wheat. *Agronomy*, 12(1).

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297. Berkeley, California: University of California Press.

Nagasubramanian, K.; Jones, S.; Sarkar, S.; Singh, A. K.; Singh, A.; and Ganapathysubramanian, B. 2018. Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant methods*, 14(1): 86.

Parmley, K.; Nagasubramanian, K.; Sarkar, S.; Ganapathysubramanian, B.; Singh, A. K.; et al. 2019a. Development of Optimized Phenomic Predictors for Efficient Plant Breeding Decisions Using Phenomic-Assisted Selection in Soybean. *Plant Phenomics*, 2019: 5809404.

Parmley, K. A.; Higgins, R. H.; Ganapathysubramanian, B.; Sarkar, S.; and Singh, A. K. 2019b. Machine Learning Approach for Prescriptive Plant Breeding. *Scientific Reports*, 9(1): 17132.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Riera, L. G.; Carroll, M. E.; Zhang, Z.; Shook, J. M.; Ghosal, S.; Gao, T.; Singh, A.; Bhattacharya, S.; Ganapathysubramanian, B.; Singh, A. K.; and Sarkar, S. 2021. Deep Multiview Image Fusion for Soybean Yield Estimation in Breeding Applications. *Plant Phenomics*, 2021: 9846470.

Shook, J.; Gangopadhyay, T.; Wu, L.; Ganapathysubramanian, B.; Sarkar, S.; and Singh, A. K. 2021. Crop yield prediction integrating genotype and weather variables using deep learning. *PLOS ONE*, 16(6): 1–19.

Singh, A. K.; Singh, A.; Sarkar, S.; Ganapathysubramanian, B.; Schapaugh, W.; Miguez, F. E.; Carley, C. N.; Carroll, M. E.; Chiozza, M. V.; Chiteri, K. O.; Falk, K. G.; Jones, S. E.; Jubery, T. Z.; Mirnezami, S. V.; Nagasubramanian, K.; Parmley, K. A.; Rairdin, A. M.; Shook, J. M.; Van der Laan, L.; Young, T. J.; and Zhang, J. 2021. *High-Throughput Phenotyping in Soybean*, 129–163. Cham: Springer International Publishing. ISBN 978-3-030-73734-4.

Singh, D. P.; Singh, A. K.; and Singh, A. 2021. *Plant breeding and cultivar development*. Academic Press.

Yoosefzadeh-Najafabadi, M.; Torabi, S.; Tulpan, D.; Rajcan, I.; and Eskandari, M. 2021. Genome-wide association studies of soybean yield-related hyperspectral reflectance bands using machine learning-mediated data integration methods. *Frontiers in plant science*, 2555.

A Table of health indices

Table 4: Summary of the 52 health indices. Here Tx denote the hyperspectral reflectance value at x nm.

Full form	Spectral Index/Ratio	Formula
Curvature index	Cl	$T675 \times T690/T683^2$
Chlorophyll Index red-edge	Clre	$T750/T710 - 1$
	Datt1	$(T850 - T710)/(T850 - T680)$
	Datt4	$T672/(T550 \times T708)$
	Datt6	$T860/(T550 \times T708)$
Double difference index	DDI	$(T749 - T720) - (T701 - T672)$
Double peak index	DPI	$(T688 + T710)/T697^2$
Gitelson2		$(T750 - T800)/(T695 - T740) - 1$
Green normalized difference vegetation index	GNDVI	$(T750 - T550)/(T750 + T550)$
Modified chlorphyll absorption ratio index	MCARI	$[(T700 - T670) - 0.2(T700 - T550)](T700/T670)$
	MCARI3	$[(T750 - T710) - 0.2(T750 - T550)](T750/T715)$
Modified normalized difference	MND1	$(T800 - T680)/(T800 + T680 - 2 \times T445)$
	MND2	$(T750 - T705)/(T750 + T705 - 2 \times T445)$
Modified simple ratio	mSR	$(T800 - T445)/(T680 - T445)$
Modified simple ratio 2	mSR2	$(T750/T705 - 1)/(\sqrt{T750/T705 + 1})$
MERIS terrestrail cholrophyll index	MTCI	$(T754 - T709)/(T709 - T681)$
Modified traingular vegetation index 1	MTVI1	$1.2[1.2(T800 - T550) - 2.5(T670 - T550)]$
Normalized difference 550/531	ND1	$(T550 - T531)/(T550 + T531)$
Normalized difference 682/553	ND2	$(T682 - T553)/(T682 + T553)$
Normalized difference chlorophyll	NDchl	$(T925 - T710)/(T925 + T710)$
Normalized difference red edge	NDRE	$(T790 - T720)/(T790 + T720)$
Normalized difference vegetation index	NDVI1	$(T750 - T650)/(T750 + T650)$
	NDVI2	$(T750 - T550)/(T750 + T550)$
	NDVI3	$(T750 - T710)/(T750 + T710)$
Normalized pigment cholrophyll index	NPCL	$(T680 - T430)/(T680 + T430)$
Normalized difference pigment index	NPQI	$(T415 - T435)/(T415 + T435)$
Optimized soil-adjusted vegetation index	OSAVI	$(1 + 0.16)(T800 - T670)(T800 + T670 - 0.16)$
Plant biochemical index	PBI	$T810/T560$
Plant pigment ratio	PPR	$(T550 - T450)/(T550 + T450)$
Physiological reference index	PRI	$(T550 - T530)/(T550 + T530)$
Pigment-specific normalized difference	PSNDb1	$(T800 - T650)/(T800 + T650)$
	PSNDc1	$(T800 - T500)/(T800 + T500)$
	PSNDc2	$(T800 - T470)/(T800 + T470)$
Plant senescence reflectance index	PSRI	$(T678 - T500)/T750$
Pigment-specific simple ratio	PSSRc1	$T[800]/T[500]$
	PSSRc2	$T[800]/T[740]$
Photosynthetic vigor ratio	PVR	$T([550] - T[650])/(T[550] + T[650])$
Plant water index	PWI	$T970/T900$
Renormalized difference vegetation index	RDVI	$(T800 - T670)/\sqrt{(T800 + T670)}$
Red-edge stress vegatation index	RVSI	$((T718 + T748)/2) - T733$
Soil-adjusted vegatation index	SAVI	$1.16((T800 - T670)/(T800 + T670 + 0.16))$
Structure intensive pigment index	SIPI	$(T800 - T445)/(T800 + T680)$
Simple ratio	SR1	$T430/T680$
	SR2	$T440/T740$
	SR3	$T550/T672$
	SR4	$T550/T750$
Disease -water stress index 4	DSWI-4	$T550/T680$
Simple ratio pigment index	SRPI	$T430/T680$
Transformed chlorophyll absorption ratio	TCARI	$3((T700 - T670) - 0.2(T700 - T550)(T700/T670))$
Traingular cholrophyll index	TCI	$1.2(T700 - T550) - 1.5(T670 - T550) \times \sqrt{T700/T670}$
Triangular vegetation index	TVI	$0.5(120(T750 - T550) - 200(T670 - T550))$
Water band index	WBI	$T970/T902$