CrossMark

# Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions

Abel Chemura[1,2] · Onisimo Mutanga[1] · Timothy Dube[3]

**Abstract** Coffee leaf rust (CLR) caused by the fungus *Hemileia vastarix* is a devastating disease in almost all coffee producing countries and remote sensing approaches have the potential to monitor the disease. This study evaluated the potential of Sentinel-2 band settings for discriminating CLR infection levels at leaf levels. Field spectra were resampled to the band settings of the Sentinel-2, and evaluated using the random forest (RF) and partial least squares discriminant analysis (PLS-DA) algorithms with and without variable optimization. Using all variables, Sentinel-2 Multispectral Imager (MSI)-derived vegetation indices achieved higher overall accuracy of 76.2% when compared to 69.8% obtained using raw spectral bands. Using the RF *out-of-bag* (OOB) scores, 4 spectral bands and 7 vegetation indices were identified as important variables in CLR discrimination. Using the PLS-DA Variable Importance in Projection (VIP) score, 3 Sentinel-2 spectral bands (B4, B6 and B5) and 5 vegetation indices were found to be important variables. Use of the identified variables improved the CLR discrimination accuracies to 79.4 and 82.5% for spectral bands and indices respectively when discriminated with the RF. Discrimination accuracy slightly increased through variable optimization for PLS-DA using spectral bands (63.5%) and vegetation indices (71.4%). Overall, this study showed the potential of the Sentinel 2 MSI band settings for CLR discrimination as part of crop condition assessment. Nevertheless further studies are required under field conditions.

**Keywords** *Hemileia vastatrix* · Disease discrimination · Red edge · Random forest · Variable optimization

✉ Abel Chemura
  achemura@gmail.com

1 School of Agricultural, Earth & Environmental Sciences, Geography Department, University of KwaZulu-Natal, P/Bag X01, Scottsville, Pietermaritzburg 3209, South Africa

2 Environmental Science and Technology Department, Chinhoyi University of Technology, P.Bag 7724, Chinhoyi, Zimbabwe

3 Department of Geography & Environmental Studies, University of Limpopo, P/Bag X1106, Sovenga, Polokwane 0727, South Africa

## Introduction

Coffee is one of the most important commodities in international agricultural trade, with an annual value of over ninety billion U.S dollars and employs about half a billion people in its value chain (Honorato Júnior et al. 2015). The main actors in coffee production are individual farmers, cooperatives and corporate farmers in over 70, mainly developing, countries (Baker et al. 2001). In these countries, coffee contributes significantly to foreign exchange earnings as an important export commodity (Gresser and Tickell 2002). Among the major challenges for the global coffee industry are pests of economic significance such as the coffee white stem borer (*Monochamus leuconotus* (Pascoe)), coffee leaf minor (*Leucoptera meyricki* (Ghesquiére)) and antestia bug (*Antestiopsis orbitalis* (Kirk)), and major diseases such as coffee leaf rust caused by *Hemileia vastatrix* (B & Br), Fusarium bark disease caused by *Fusarium lateritium* (Nees) var. *longum* (Wollenw.) and coffee berry disease caused by *Colletorichum kahawae* (Waller & Bridge) (Brown 2008; Chemura et al. 2015; Kutywayo et al. 2013). Coffee leaf rust (CLR), caused by the fungus *Hemileia vastatrix*, is the most destructive disease and threat to coffee production the world-over (Cressey 2013; Ghini et al. 2011).

CLR is most severe on *Coffea arabica*, which accounts for about 70% of the world coffee production and supply (Dinesh et al. 2011). CLR is regarded among the most potentially devastating of all foliar plant diseases (Suresh et al. 2012). The basidiomycete *H. vastatrix* is an obligate biotrophic fungus that is found in almost all coffee producing countries and capable of long distance dispersal (Brown and Hovmøller 2002). The colonization of living plant tissue results in small chlorotic spots which quickly become fruiting structures (uredia) rupturing the epidermal layer and releasing the uredospores (Avelino et al. 2012; Haddad et al. 2009; Silva et al. 2006). The abaxial yellow to orange uredia of *H. vastatrix,* which may be called rusty, grow and may coalesce. Unlike other fungal plant diseases, CLR is not necrotic and its symptoms appear only on the underside of the leaves (Belan et al. 2015, Honorato Júnior et al. 2015). In the absence of early detection and proper management, CLR results in up to 50% loss of leaves and 70% yield reduction in coffee through premature leaf drop, dieback and debilitation of trees, which will eventually lead to death of coffee plants (Avelino et al. 2004).

Current CLR disease monitoring methods rely on occasional field surveys by teams of specially trained and experienced personnel. However, besides being the largely adopted approach, particularly in resource limited areas, the technique is strenuous especially for large coffee plantations and is subjective. Remote sensing therefore, offers timely and spatially explicit objective assessment of plant condition throughout the growing season (Sankaran et al. 2010). The urgent adoption of these technologies is perceived to increase crop productivity through the provision of accurate, targeted and up-to-date crop information, as well as reduce costs of disease control and environmental contamination associated with excessive pesticide applications (Barbedo 2013; Laudien et al. 2004). The success of potential crop protection methods is highly dependent on early disease detection.

Previous work demonstrates that remote sensing approaches can be reliably used in the detection of plant diseases in many crops. For instance, Huang et al. (2007) demonstrated that the photochemical reflectance index (PRI) developed from field hyperspectral remote sensing data can be applied to detect and quantify yellow rust in winter wheat ($r^2 = 0.91$). They concluded that this provided a basis for development of proximal sensing approaches, and after scale and field operational issues are factored in, for airborne or spaceborne-based

monitoring of yellow rust in winter wheat. In addition, Mahlein et al. (2013) developed specific spectral disease indices (SDIs) for the detection and discrimination of healthy sugar beet leaves from those infected with *Cercospora* leaf spot, sugar beet rust and powdery mildew and achieved high accuracy and sensitivity of over 85%. It is therefore clear from these studies that species-specific disease indices derived from hyperspectral bands located in the narrow contiguous parts of the red-edge and NIR regions have the strength and capability to significantly enhance crop disease detection, identification and monitoring. Reflectance in these regions is able to identify changes in internal leaf structure, content and processes that affect absorption of radiation and thus reveal physiological stress in plants caused by the disease or other stress (Coops et al. 2003; Eitel et al. 2011). These detected changes are associated with changes in the quality and quantity of chlorophyll and chemical properties of the affected leaves, when compared to their previous stress-free condition or unstressed counterparts (Carter and Knapp 2001; Mutanga and Skidmore 2007). Many of these specific wavebands were previously only available in hyperspectral sensors, which are known to have many challenges, such as high costs, high dimensionality and above all poor spatial coverage. So far, there are no prospects that there will be global coverage of affordable hyperspectral data in the near future at sufficient temporal resolution (Rulinda et al. 2012). This therefore means that there is need to shift towards harnessing the opportunities provided by the new generation of multispectral imaging sensors.

New generation multispectral space-borne earth observation instruments, such as WorldView-2, RapidEye and the recently-launched Sentinel-2 multispectral imager have incorporated narrow wavebands including those in the red-edge position that were not available in predecessor sensors (i.e. Landsat series, MODIS, SPOT, ASTER etc.). These technological advancements therefore provide an opportunity for timely landscape or farm-based assessment of crop condition (i.e. health status and yield estimation). Unlike the hyperspectral sensors, multispectral sensors have a huge swath-width and are currently available at low or no costs for many developing countries where coffee is produced. There has been a lot of interest in the Sentinel 2 multispectral imager (MSI) data in terms of its potential to advance multispectral remote sensing applications (Clevers and Gitelson 2013). This is because it is freely available with relatively high resolution, as well as strategically positioned bands, which makes it useful for many applications, including vegetation characterization and mapping (Hedley et al. 2012; Hill 2013; Vincini et al. 2014). The recently-launched Sentinel 2 multispectral satellite capitalizes on the technology and the vast experience acquired with SPOT and Landsat series data over the past decades. Sentinel 2 multispectral imager which is a polar orbiting, super-spectral high resolution imaging mission (Frampton et al. 2013; Hansen and Loveland 2012), has a huge swath-width of about 290 km with thirteen unique spectral bands. These spectral bands range from the visible and near infrared (VNIR) to the shortwave infrared (SWIR) regions of the spectrum. Of these thirteen bands, four are provided at 10 m spatial resolution, six bands at 20 m spatial resolution and three bands at 60 m spatial resolution (Table 1) (Clevers and Gitelson 2013; D'Odorico et al. 2013; Hill 2013). Most importantly, the sensor provides data about the earth's surface every five days under cloud-free conditions, and typically every 15–30 days in cloudy areas, making it attractive for temporal feature analysis (Hedley et al. 2012).

One of the most attractive features of the Sentinel 2 multispectral imager is that it incorporates three new bands in the red-edge region, which are centered at 705, 740 and 783 nm specifically designed for vegetation characterization and quantification (Frampton et al. 2013; Hedley et al. 2012). Because of these advanced sensor characteristics, Sentinel-2

**Table 1** Specifications of the Sentinel-2 Multispectral Instrument (MSI) band settings showing center wavelengths, band width and spatial resolution

| Spectral band | Center wavelength (nm) | Band width (nm) | Spatial resolution (m) |
|---|---|---|---|
| **B1** | **443** | **20** | **60** |
| B2 | 490 | 65 | 10 |
| B3 | 560 | 35 | 10 |
| B4 | 665 | 30 | 10 |
| B5 | 705 | 15 | 20 |
| B6 | 740 | 15 | 20 |
| B7 | 783 | 20 | 20 |
| B8 | 842 | 115 | 10 |
| **B8a** | **865** | **20** | **20** |
| **B9** | **945** | **20** | **60** |
| **B10** | **1380** | **30** | **60** |
| **B11** | **1610** | **90** | **20** |
| **B12** | **2190** | **180** | **20** |

The bold bands were excluded from the analysis

multispectral imager is therefore hypothesized to be capable of providing timely data for the generation of high-level operational products. These include the generation of spatially explicit estimation and monitoring of important plant biophysical variables, (i.e. chlorophyll, LAI and leaf water content and crop health) in addition to producing generic land-cover, land-change detection and crop disease maps. The practical application of the Sentinel 2 sensor in coffee and other plantation crops may, however, be limited. This is because the effect of disease infection are not only confused by soil background effect as influenced by age but also by other confounding factors such as co-infection with other diseases, nutrient deficiency issues, water stress among many others. This therefore means that what the sensor detects may not necessarily be the effect of the crop disease but something else. It therefore becomes of paramount importance to determine the possibility of separating disease levels using the Sentinel 2 sensor in the absence of other potential stressors. The aim of this work was therefore to evaluate the potential of the Sentinel-2 multispectral imager-derived band settings and vegetation indices in discriminating coffee leaf rust infection levels using the random forest (RF) and partial least squares discriminant analysis (PLS-DA) algorithms at leaf level.

## Materials and methods

### Study area and experimental setup

The study was carried out at Coffee Research Institute, Chipinge, Zimbabwe (32°37.523′E, 20°12.474′S and altitude 1100 m above sea level). The climate in Chipinge is subtropical with two distinct seasons: the dry season and the wet season. Average total annual rainfall is 1800 mm of which 80% falls in 5 months from November to March. The mean daily maximum temperature is 20 °C and mean daily minimum temperature is 14 °C. Coffee
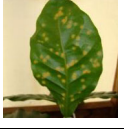
seeds (variety Yellow Catuai) were pre-germinated on damp sand covered with Hessian in October 2013. Pre-germination was done because it quickly brings the coffee seed to the stage of radicle emergence (i.e. when the root tip breaks through the parchment casing) due to the artificial medium. This technique is easier to manage and also results in relatively same aged seedlings as germination is erratic and varied under natural media (Logan and Biscoe 1987). All the plantlets that had reached two true leaf stage were transplanted into black polythene pots (290 mm height × 135 mm diameter giving a volume of 4150 cm$^3$). The substrate in the polythene pots was made of virgin soil, sand and local compost at the ratio of 2:1:1. Compound S (6 N: 17P$_2$0$_5$: 6K$_2$0) was mixed thoroughly with potting soil at a rate of 3 kg m$^{-3}$ of the soil-sand-compost mixture to achieve about 10 g per pot. All other routine nursery management activities related to irrigation and pest control were done according to the Coffee Handbook (Logan and Biscoe 1987). The seedlings were left to grow under a nursery shade (70%) until they were 6 months old (May 2014) and over 120 polythene pots were transferred to a greenhouse for further management and treatment application. After two months of acclimatization to the greenhouse conditions, CLR disease inoculations were done.

## Coffee leaf rust inoculation

Eight months old coffee seedlings were inoculated with coffee leaf rust in a greenhouse for the study. This age was selected because it is the minimum recommended age for transplanting coffee seedlings into the field (Logan and Biscoe 1987). The average number of leaves at time of inoculation was 12 (6 fully grown leaf pairs excluding the apex bud) while the average height of the plants was 232 mm. For the inoculation, CLR spores were collected from naturally infected coffee plants from a coffee field that is maintained at the station as disease reservoir in June 2015. The spores from the infected leaves were scraped into petri dishes, using a razor blade. These spores were then used to make a spore suspension of 8 × 10$^6$ spores/ml, using sterile deionized water as counted by a hemocytometer. Inoculation of spores was done by brushing the spore suspension on the underside of leaves, using pen brushes. The inoculated plants were incubated for 72 h in dark incubation chambers with ∼100% relative humidity.

After three days in the incubation chamber, seedlings were removed, and laid on benches to allow for the leaves to show disease symptoms. Eighty plants were inoculated in two batches of 40 each in a space of two weeks. The first inoculation was done with 40 plants and the success of the inoculation was evaluated after one week with the successfully inoculated plants separated while those that did not succeed had the inoculation repeated. Twenty-one days after the first inoculation, coffee plants were scored and grouped with the help of a plant pathologist, into severely infected, moderately infected and healthy seedlings (no inoculation) based on diseased area. Three weeks was used because it was considered sufficient for the *H. vastarix* to complete its life cycle and for the lesions to mature. The reason for varying the inoculation period was to achieve a gradient in CLR levels. The gradient was required because the detection is important for assessing potential for early detection which can allow for effective control. Leaves showing other signs of infection apart from CLR were excluded. Reflectance from 21 coffee plants was measured for each level of infection (healthy, moderate CLR and severely infected level (n = 63, Table 2). Only 21 samples were used for each class because of the poor success rate of inoculation associated with CLR under controlled conditions and this was the minimum number of samples that produced a balanced number of samples for all groups. The distribution of area diseased for moderate and severe leaf samples are shown in Fig. 1.

**Table 2** Description of levels of CLR infection levels, sample images and number of samples used in the study

| Disease levels | Sample picture | Description of class |
| --- | --- | --- |
| Healthy (n = 21) |  | Healthy leaves from plants that were left non-inoculated |
| Moderate infection (n = 21) |  | Infected leaves with early or sparse spores of CLR visible of the underside of the leaves. Estimated covered area less than 10% of the leaf area |
| Severe infection (n = 21) |  | Severely infected leaves with typical CLR yellowing on the underside of leaves. Covered area more than 10% of the leaf area |
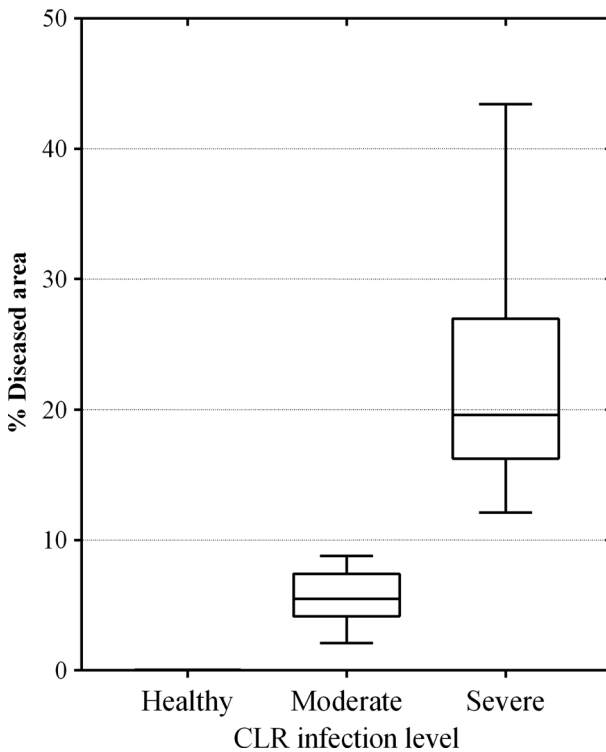


**Fig. 1** Box plots of percent diseased area of leaves for the three levels as measured on the day of reflectance measurements

## Reflectance measurements and resampling

Reflectance was measured using an Apogee VIS–NIR spectrometer (Apogee Instruments, Inc., Logan, UT, USA) with an effective spectral range of 400–900 nm and a spectral resolution of 0.5 nm. Each reading consisted of an average of three spectral scans, taken at 150 mm above the coffee leaf of interest at 30° field of view, allowing measurement from a leaf spot with a radius of 20 mm. This resulted in 1257 $mm^2$ to be measured on each coffee leaf. A white polytetrafluoroethylene (PTFE) reflectance standard was used as a reference. Reflectance by wavelength was calculated as the ratio of scene reflectance to the reflectance of the standard. A moving Savitzky–Golay filter with a frame-size of 3 data points and a 2nd order polynomial were employed to smooth the spectra (Savitzky and Golay 1964) and the reflectance was averaged to 5 nm to reduce dimensionality.

The collected reflectance measurements were resampled to simulate the Sentinel-2 satellite sensor's reflectance (Table 1). The resampling of the field spectra was done in ENVI 4.7 software (Exelis Visual Information Solutions, Boulder, CO, USA). The resampling method used applies a Gaussian model with a Full Width at Half-Maximum (FWHM) equal to the band spacing provided. The technique uses the field spectral data from the spectrometer and resamples it to the spectral width of the sensor being simulated. Only seven Sentinel-2 MSI land management bands were used, because the other bands were considered unnecessary for plant biophysical studies, had a higher spatial resolution for application in coffee or were outside the range of the spectro-radiometer used in this study.

## Vegetation indices

Seventeen spectral ratio-based vegetation indices were computed and applied in evaluating transformation of spectral bands of the Sentinel-2 MSI sensor characteristics ability to discriminate the different CLR infection levels (Table 3). These vegetation indices were selected based on their reported ability to discriminate different vegetation characteristics and conditions from remotely sensed data (Eitel et al. 2006). The three red edge bands contained in Sentinel-2 MSI were used to generate vegetation indices that were indexed according to the particular red edge band used (B5, B6 and B7).

## Modelling approaches

Two machine learning algorithms, the random forest (RF) and partial least squares-discriminant analysis (PLS-DA), were used to model CLR from both spectral bands and vegetation indices. These two classifiers were selected for the modelling task because they have been proved to be more robust and effective for vegetation condition discrimination in other studies when compared to most supervised parametric and machine learning algorithms (de Almeida et al. 2013; Lebedev et al. 2014; Nitze and Schulthess 2012; Rogan et al. 2008). These two methods also have in-built variable optimization, making them more attractive for approaches requiring selection of variables as in this study (Li et al. 2016; Pal 2005). In addition, it was required that the methods selected for the task be robust for relatively small sample sizes and compatible with cross-validation for accuracy assessment given the difficulties in getting large sample sizes for CLR.

**Table 3** Selected vegetation indices (VIs) evaluated in the study, showing their formula, bands used and source

| Name | Formula | Sentinel-2 Bands | Source |
|---|---|---|---|
| Normalized Difference Vegetation Index | $NDVI = \frac{\rho_{NIR} - \rho_R}{\rho_{NIR} + \rho_R}$ | B8, B4 | Rouse et al. (1973) |
| Simple Ratio | $SR = \frac{\rho_{NIR}}{\rho_R}$ | B8, B4 | Baret and Guyot (1991) |
| Green Chlorophyll Index | $GCI = \left(\frac{\rho_{NIR}}{\rho_{GREEN}}\right) - 1$ | B8, B3 | Gitelson et al. (2005) |
| Green Normalized Difference Vegetation Index | $GNDVI = \frac{\rho_{NIR} - \rho_{GREEN}}{\rho_{NIR} + \rho_{GREEN}}$ | B8, B3 | Gitelson et al. (1996) |
| Renormalized Normalized Difference Vegetation Index | $RNDVI = \frac{\rho_{NIR} - \rho_{RED}}{\sqrt{\rho_{NIR} + \rho_R}}$ | B8, B4 | Gitelson and Merzlyak (1994) |
| Normalized Difference Red edge Index | $NDVI.RE1 = \frac{\rho_{NIR} - \rho_{RE1}}{\rho_{NIR} + \rho_{RE1}}$ | B8, B5 | Gitelson and Merzlyak (1994) |
| | $NDVI.RE2 = \frac{\rho_{NIR} - \rho_{RE2}}{\rho_{NIR} + \rho_{RE2}}$ | B8, B6 | Gitelson and Merzlyak (1994) |
| | $NDVI.RE3 = \frac{\rho_{NIR} - \rho_{RE3}}{\rho_{NIR} + \rho_{RE3}}$ | B8, B7 | Gitelson and Merzlyak (1994) |
| Simplified Canopy Chlorophyll Content Index | $SCCCI1 = \frac{NDVI.RE1}{NDVI}$ | B8, B4, B5 | Barnes et al. (2000) |
| | $SCCCI2 = \frac{NDVI.RE2}{NDVI}$ | B8, B4, B6 | Barnes et al. (2000) |
| | $SCCCI3 = \frac{NDVI.RE3}{NDVI}$ | B8, B4, B7 | Barnes et al. (2000) |
| Red-edge Chlorophyll Index | $CI_{RE1} = \left(\frac{\rho_{NIR}}{\rho_{RE1}}\right) - 1$ | B8, B5 | Gitelson et al. (2005) |
| | $CI_{RE2} = \left(\frac{\rho_{NIR}}{\rho_{RE2}}\right) - 1$ | B8, B6 | Gitelson et al. (2005) |
| | $CI_{RE3} = \left(\frac{\rho_{NIR}}{\rho_{RE3}}\right) - 1$ | B8, B7 | Gitelson et al. (2005) |
| Normalized Red Edge Difference Index | $NREDI1 = \frac{\rho_{RE3} - \rho_{RE1}}{\rho_{RE3} + \rho_{RE1}}$ | B7, B5 | Gitelson and Merzlyak (1994) |
| | $NREDI2 = \frac{\rho_{RE3} - \rho_{RE2}}{\rho_{RE3} + \rho_{RE2}}$ | B7, B6 | Gitelson and Merzlyak (1994) |
| | $NREDI2 = \frac{\rho_{RE2} - \rho_{RE1}}{\rho_{RE2} + \rho_{RE1}}$ | B6, B5 | Gitelson and Merzlyak (1994) |

## Random forest algorithm

The RF is an ensemble algorithm developed by Breiman (2001). The classification version of RF that was developed by Breiman and Cutler (2007) together with its in-built variable optimization was used in this analysis. The random forest employs an iterative bagging operation where a number of trees (*ntree*) are independently built using a random subset of samples from the training samples. Each tree is then independently grown to a maximum size based on a bootstrap sample of about two-thirds the training dataset. Each node is then split using the best, among a subset of input variables (*mtry*) (Chemura and Mutanga 2016; Ghimire et al. 2012). The random forests ensemble classifies the data that are not in the trees as out-of-bag (OOB) data, and by averaging the OOB error rates from all trees, the RF algorithm gives an error rate called the OOB classification error for each input variable

(Breiman 2001; Lebedev et al. 2014). The default number of trees (501) and the square root of the number of variables for *mtry* were used as RF settings in the randomForest library (Liaw et al. 2009) in R (R Core Team, 2013).

## PLS-DA algorithm

In addition to the RF approach, partial least squares discriminant analysis (PLS-DA) algorithm was used in CLR discrimination using Sentinel-2 MSI sensor-derived variables. PLS-DA is the classification version of PLS regression and is a powerful multivariate supervised pattern recognition method that uses a training routine to assign class membership to variables based on their known statistical parameters projected into latent variables (Wang et al. 2011). The produced eigenvectors from spectral bands and vegetation indices explain the variance within the input data and their correlation with response variables. The PLS-DA method is performed by a PLS regression against a dummy matrix $Y$ that indicates class membership (Li et al. 2016). A regression analysis is then conducted to establish the PLS regression model between categorical variable $Y$ and calibration matrix of explanatory variable, $X$. The PLS-DA model can be represented as:

$$X = TP^T + E \tag{1}$$

$$Y = UQ^T + F \tag{2}$$

where $X$ represents the matrix of spectral data and vegetation indices, $T$ is a factor score matrix, $P$ is the $X$ loadings, $E$ is the residual or a noise term, $Y$ is a matrix of the categorical variable, $U$ is the scores for $Y$, $Q$ is the $Y$ loadings, and $F$ is the residuals (de Almeida et al. 2013; Wang et al. 2011).

The PLS-DA algorithm has been used with spectral data because it has many advantages over the linear discriminant analysis (LDA) approach. These include noise reduction, showing the probability of a sample belonging to the class being modelled and the selection of variables (Li et al. 2016). The latter is particularly attractive for this study because it is important to determine the important variables that are explaining the discrimination outcome in order to get a mechanistic understanding for interpretation of the variables. It is very important in PLS-DA to determine the appropriate number of components for the model to avoid overfitting and this was done using cross-validation (Wang et al. 2011). The appropriate number of components was identified as 2 for spectral and 3 for vegetation indices and these were used accordingly. The PLS-DA approach was implemented using the mixOmics library (Cao et al. 2015) in R (R Core Team, 2013).

## Variable optimization

It was found imperative to test the performance of the most important variables as determined by the two algorithms (RF and PLS-DA). The variables considered important for CLR discrimination were then used for discrimination excluding all other variables. The RF and PLS-DA algorithms have different internal approaches for variable optimization and these were employed accordingly to Sentinel-2 MSI-derived spectral bands and vegetation indices.

## RF and PLS-DA variable optimization

The RF variable importance measure was used to select bands and vegetation indices that were used in CLR discrimination. The RF algorithm assesses the importance of each input variable to the outcome by comparing how much the OOB error increases when a variable is removed, while all others are left unchanged (Breiman and Cutler 2007; Gislason et al. 2004). This way, the RF ranks the variables according to the mean decrease in error when that variable is included in the modelling and those variables with higher mean decreases in error are the most important variables for the modelling and should therefore be retained.

Similarly, the PLS-DA was used in variable optimization through the Variable Importance in Projection (VIP) scores produced in the discrimination. The VIP is a quantitative estimate of the discriminatory power of each individual variable used in the model and as such can be used to interpret the outcome in relation to the input variables. The VIP is a weighted sum of squares of the PLS loadings that takes into account the amount of explained Y-variance of each component. The weights are a function of the reduction of the sums of squares across the number of PLS components. A variable with a VIP Score greater than 1 is considered important in the model while those with VIP scores less than 1 are less important, and are candidates for exclusion from the model (de Almeida et al. 2013). To obtain the importance of the $j$th variable ($VIP_j$) the following formula is used:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \cdot SSY_f \cdot J}{SSY_t \cdot F}} \tag{3}$$

where $w_{if}$ is the PLS weight value of the $j$th variable and the $f$th component, $SSY_f$ is the sum of squares of the dependent variable that was obtained from the discriminant model with $f$ ($f = 1,2,\ldots,F$) components, $J$ is the number of spectral bands or vegetation indices being used, $SSY_t$ is the total sum of squares of the dependent variable, and $F$ is the number of PLS components being assessed (de Almeida et al. 2013).

## Accuracy assessment

In order to assess the performance of CLR discrimination, k-fold cross validation with 10 folds was used since the sample number was relatively small (n = 63) for sub-setting the data into training and test data. A confusion matrix, defined as a table that describes the performance of a discriminating model on a set of test data for which the true values are known, was used to evaluate the overall accuracy of the RF and PLS-DA discrimination of CLR infection levels (i.e. healthy, moderate and severe CLR infection. Overall accuracy, Kappa (k) and related class user's and producer's accuracies were calculated to evaluate the performance of Sentinel-2 MSI derived spectral bands and vegetation indices in discriminating CLR, using the RF and PLS-DA algorithms. The effect of variable optimization on discrimination accuracy was determined by the McNemar's test (Foody 2004) of confusion matrices of discrimination with all variables versus discrimination with optimized variables.

# Results

## Spectral resampling

Figure 2 shows the mean reflectance of Sentinel-2 MSI bands for healthy coffee leaves, and leaves with moderate and severe CLR infection. As expected with vegetation, the results show that coffee leaf reflectance was higher in the NIR region of the spectrum when compared to the visible spectral region. Healthy leaves produced higher reflectance in B3 (green) and the least in two of the three Sentinel-2 MSI red edge bands (B5 and B6) where the severely CLR infected leaves produced the highest reflectance. The reflectance of the severely infected CLR was least in the red-edge 3 band (B7) and the NIR band (B8).

## CLR Discrimination with all variables

The results from CLR discrimination with all variables (n = 7 for bands and n = 17 for vegetation indices) are shown in Table 4. When Sentinel-2 MSI seven bands were used with the PLS-DA, an overall accuracy of 63.5% (k = 0.45) was achieved while this increased to 69.8% (k = 0.55) from the use of the RF algorithm. The lowest producer and user's accuracies were obtained for the moderate CLR class, while the best classified was the healthy class (RF and PLS-DA producer's accuracies of 81 and 90.5% and user's accuracies of 77.3 and 76.0% respectively). Results indicate spectral confusion in discriminating between the moderate and severe CLR when compared to the discrimination between healthy and moderate CLR for both algorithms. For example, eight samples in the moderate class were classified by the PLS-DA as being in the severe CLR rust while seven in the severe class were classified by the same algorithm to be in the moderate class
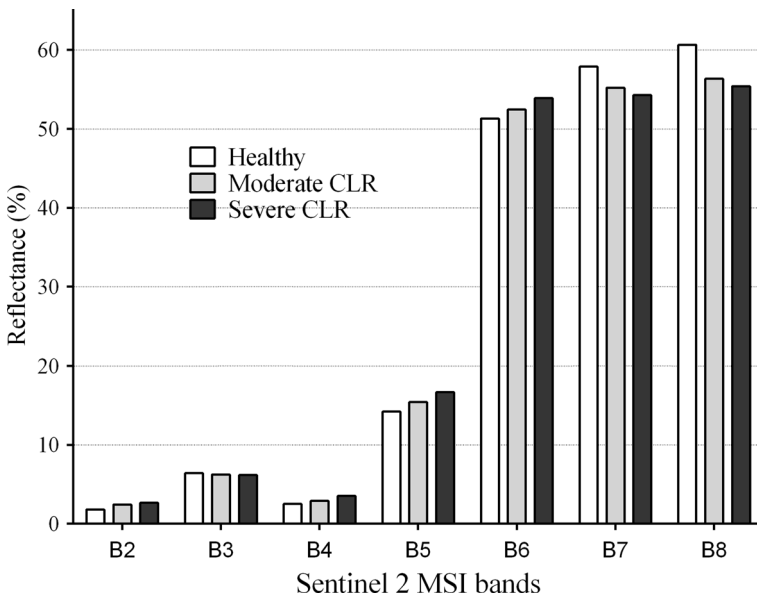


**Fig. 2** Mean spectral reflectance of CLR infection levels across the Sentinel-2 spectral band settings obtained from resampling hyperspectral imagery

Table 4 Confusion matrix showing overall, users and producers' accuracy and kappa for CLR discrimination accuracies obtained using all Sentinel 2 MSI derived spectral bands and vegetation indices

| | RF | | | | | PLS-DA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Healthy | Moderate CLR | Severe CLR | Total | Producer's accuracy | Healthy | Moderate CLR | Severe CLR | Total | Producer's accuracy |
| **All Bands (N=7)** | | | | | | | | | | |
| Healthy | 17 | 3 | 1 | 21 | 81.0 | 19 | 2 | 0 | 21 | 90.5 |
| Moderate CLR | 3 | 13 | 5 | 21 | 61.9 | 4 | 9 | 8 | 21 | 42.9 |
| Severe CLR | 2 | 5 | 14 | 21 | 66.7 | 2 | 7 | 12 | 21 | 57.1 |
| Total | 22 | 21 | 20 | 63 | | 25 | 18 | 20 | 63 | |
| User's accuracy | 77.3 | 61.9 | 70.0 | | | 76.0 | 50.0 | 60.0 | | |
| Overall accuracy | 69.8 | | | | | 63.5 | | | | |
| Kappa | 0.55 | | | | | 0.45 | | | | |
| **All VIs (N=17)** | | | | | | | | | | |
| Healthy | 19 | 0 | 2 | 21 | 90.5 | 19 | 2 | 0 | 21 | 90.5 |
| Moderate CLR | 3 | 14 | 4 | 21 | 66.7 | 4 | 10 | 7 | 21 | 47.6 |
| Severe CLR | 2 | 4 | 15 | 21 | 71.4 | 0 | 7 | 14 | 21 | 66.7 |
| Total | 24 | 18 | 21 | 63 | | 23 | 19 | 21 | 63 | |
| User's accuracy | 79.2 | 77.8 | 71.4 | | | 82.6 | 52.6 | 66.7 | | |
| Overall accuracy | 76.2 | | | | | 68.3 | | | | |
| Kappa | 0.64 | | | | | 0.52 | | | | |

(Table 4). These were reduced to five when the discrimination was done using the RF algorithm.

Higher accuracies in discriminating CLR were achieved by using all vegetation indices (PLS-DA = 68.3% and RF = 76.2%). Use of Sentinel-2 derived vegetation indices in CLR discrimination resulted in increases in user and producer accuracies of healthy, moderate and severe CLR classes, when compared to that obtained from using raw spectral bands. However, notable misclassifications were observed when vegetation indices were used between moderate CLR (producer accuracy = 66.7% for RF and 47.6% for PLS-DA) and severe CLR classes (producer accuracy = 71.4% for RF and 66.7% for PLS-DA). It was therefore, clear that a significant number of samples in the moderate CLR class were classified as severe CLR when using all Sentinel-2 spectral bands and vegetation indices in CLR discrimination (Table 5).

## Variable optimization

The results in Fig. 3 illustrate the most important selected model variables (i.e. bands and vegetation indices) derived using the RF OOB and VIP methods. The RF model identified four spectral bands (B4, B6, B3, B7) whereas PLSA-DA selected only three (B4, B6 and B5). Comparatively, both models managed to select two identical variables (i.e. B4 and B6 bands) as the most important variables for discriminating CLR levels (Fig. 3a and 3b).

When the variable optimization was implemented using the seventeen derived vegetation indices, only seven were selected as important by RF and most of these were transformations of Sentinel-2 MSI red edge bands and the NIR (Fig. 3c). The seven vegetation indices identified as important in CLR discrimination by RF were SCCCI3 (B8, B7 & B4), RNDVI (B8 & B4), SCCCI2 (B8, B6 & B4), $CI_{RE1}$ (B8 & B5), $CI_{RE3}$ (B8 & B7), NREDI3 (B6 & B5) and GCI (B8 & B3) (Fig. 3d). On the other hand, the PLS-DA algorithm identified only five vegetation indices (i.e. SCCCI3, $CI_{RE1}$, RNDVI, NREDI2, GCI) as having VIP values >1. Of all the variables selected by the two models, only four of them (vegetation indices) were selected by both the methods (SCCCI3, $CI_{RE1}$, RNDVI and GCI).

## CLR discrimination with optimized variables

Optimized Sentinel-2 MSI spectral band settings and vegetation indices were applied in CLR discrimination using RF and PLS-DA algorithms. Table 5 shows the results of CLR discrimination using optimized variables. The results indicate that there is an improvement in accuracy of CLR discrimination when optimized variables are used, with the highest magnitude of change observed in optimized vegetation indices (Table 5). Overall, optimized Sentinel-2 derived vegetation indices achieved the highest overall CLR discrimination accuracies of 82.5% (k = 0.74) and 71.4% (k = 0.57) using the RF and PLS-DA algorithms respectively. It can be noted that CLR discrimination accuracies increased by about 10% for RF algorithm. On the other hand, optimization of Sentinel-2 derived vegetation indices reduced the number of parameters from 17 to 5 variables. Optimization improved the performance of the PLS-DA model in CLR discrimination, particularly in terms of the producer accuracy of the class severe and all user accuracies.

**Table 5** Confusion matrix showing overall, users and producers' accuracy and kappa for CLR discrimination accuracies derived using the most important selected model variables (i.e. optimized spectral bands and vegetation indices)

**Optimized Bands (n = 4) — RF**

| | Healthy | Moderate CLR | Severe CLR | Total | Producer's Accuracy |
|---|---|---|---|---|---|
| Healthy | 20 | 1 | 0 | 21 | 95.2 |
| Moderate CLR | 3 | 13 | 5 | 21 | 61.9 |
| Severe CLR | 1 | 3 | 17 | 21 | 81.0 |
| Total | 24 | 17 | 22 | 63 | |
| User's Accuracy | 83.3 | 76.5 | 77.3 | | |
| Overall Accuracy | 79.4 | | | | |
| Kappa | 0.69 | | | | |

**Optimized Bands (n = 3) — PLS-DA**

| | Healthy | Moderate CLR | Severe CLR | Total | Producer's Accuracy |
|---|---|---|---|---|---|
| Healthy | 19 | 1 | 1 | 21 | 90.5 |
| Moderate CLR | 3 | 9 | 9 | 21 | 42.9 |
| Severe CLR | 3 | 6 | 12 | 21 | 57.1 |
| Total | 25 | 16 | 22 | 63 | |
| User's Accuracy | 76.0 | 56.3 | 54.6 | | |
| Overall Accuracy | 63.5 | | | | |
| Kappa | 0.45 | | | | |

**Optimized VIs (n = 7) — RF**

| | Healthy | Moderate CLR | Severe CLR | Total | Producer's Accuracy |
|---|---|---|---|---|---|
| Healthy | 19 | 2 | 0 | 21 | 90.5 |
| Moderate CLR | 1 | 14 | 6 | 21 | 66.7 |
| Severe CLR | 1 | 1 | 19 | 21 | 90.5 |
| Total | 21 | 17 | 25 | 63 | |
| User's Accuracy | 90.5 | 82.4 | 76.0 | | |
| Overall Accuracy | 82.5 | | | | |
| Kappa | 0.74 | | | | |

**Optimized VIs (n = 5) — PLS-DA**

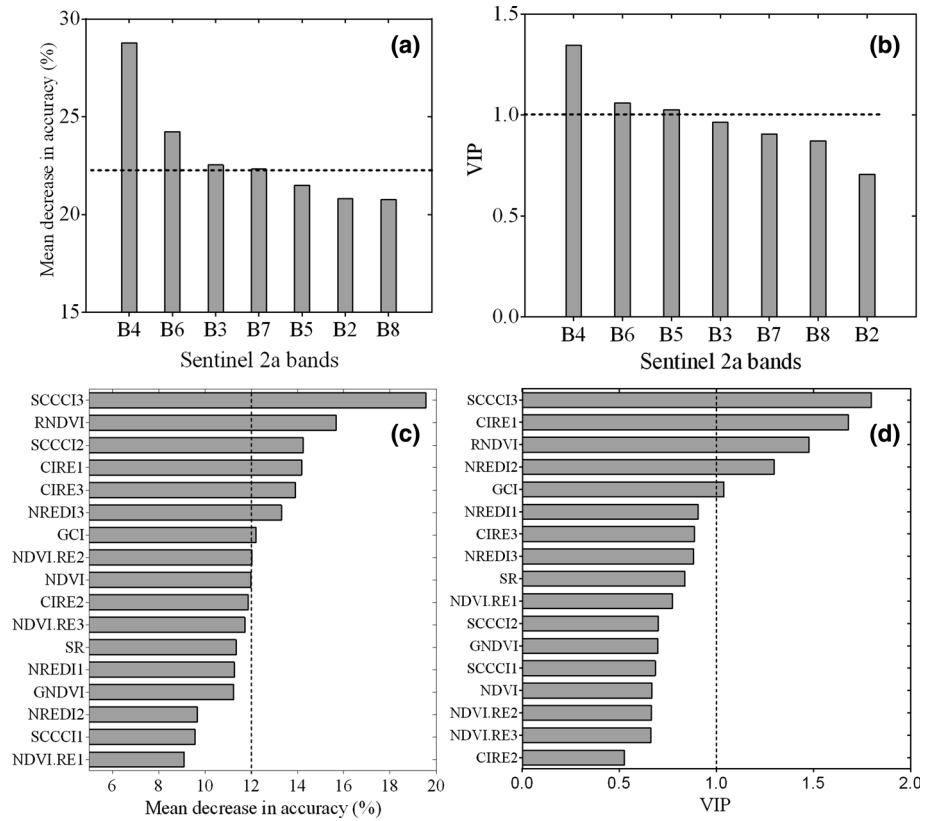| | Healthy | Moderate CLR | Severe CLR | Total | Producer's Accuracy |
|---|---|---|---|---|---|
| Healthy | 18 | 3 | 0 | 21 | 85.7 |
| Moderate CLR | 2 | 11 | 8 | 21 | 52.4 |
| Severe CLR | 1 | 4 | 16 | 21 | 76.2 |
| Total | 21 | 18 | 24 | 63 | |
| User's Accuracy | 85.7 | 61.1 | 66.7 | | |
| Overall Accuracy | 71.4 | | | | |
| Kappa | 0.57 | | | | |

**Fig. 3** Optimization of Sentinel-2 MSI variables for CLR discrimination through **a** RF-OOB error for spectral bands **b** PLS-DA for spectral bands **c** RF-OOB error for vegetation indices and **d** PLS-DA VIP for vegetation indices. The *dotted line* shows the cut-off point for variables

## Comparison of un-optimized and optimized variables in CLR discrimination

The RF outperformed the PLS-DA in CLR discrimination using Sentinel-2 MSI data with and without model optimization (Fig. 4). The differences in accuracy due to optimization were also assessed through McNemar's test (Table 6), which showed that variable optimization significantly improved the accuracy of CLR discrimination with spectral bands run through the RF algorithm ($\chi^2 = 4.17$, $p < 0.05$).

## Discussion

This study aimed to evaluate the potential of the Sentinel-2 MSI spectral band settings and vegetation indices in discriminating CLR infection levels as a basis for field disease monitoring and modelling for precision farming. Two machine-learning algorithms were applied on un-optimized and optimized Sentinel-2 MSI spectral bands and vegetation indices. Accurate discrimination of CLR infection levels provides the most required
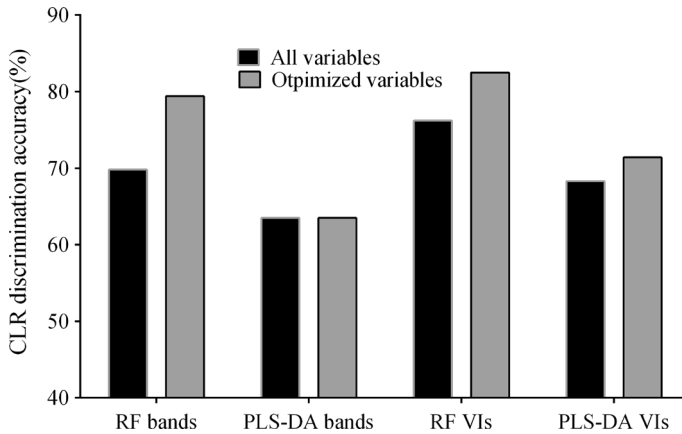
**Fig. 4** Effect of RF and PLS-DA model variable optimization on CLR discrimination accuracies using Sentinel-2 MSI bands and vegetation indices

**Table 6** McNemar's test for RF and PLS-DA models on reflectance bands and vegetation indices with and without variable optimization

| Parameters | $f_{11}$ | $f_{12}$ | $f_{21}$ | $f_{22}$ | Total | $\chi^2$ | $P$ value (95%) |
|---|---|---|---|---|---|---|---|
| RF Bands | 44 | 0 | 6 | 13 | 63 | 4.17 | 0.041 |
| RF VIs | 48 | 0 | 4 | 11 | 63 | 2.25 | 0.133 |
| PLS-DA Bands | 37 | 3 | 3 | 20 | 63 | 0.00 | 1.000 |
| PLS-DA VIs | 41 | 2 | 4 | 16 | 63 | 0.17 | 0.683 |

knowledge for precise crop diseases monitoring and controlling, especially if high crop productivity is to be realized.

## Spectral resampling

The resampled spectra demonstrated the general response of vegetation to electromagnetic radiation with lower reflectance in the visible region and higher reflectance in the NIR regions. This research confirms the general effect of different types of biotic stress on vegetation reflectance observable by the blue shift in the red-edge region of the spectrum where infected samples produced higher reflectance, when compared to healthy samples. The blue shift as indicated by the shift of the slope of stressed plants towards the blue spectrum is associated with a decline in leaf chlorophyll quantity and composition (Feng et al. 2016; Ustin et al. 2009) and can be used as an early general indicator of presence of stressors such as disease infection, allowing for further investigations. This phenomenon has been reported for plant diseases (Prabhakar et al. 2013), nutrient stress (Mutanga and Skidmore 2007), pest attacks (Stone et al. 2001) and other physio-chemical plant disorders (Carter and Knapp 2001).

## Performance of Sentinel-2 spectral bands in CLR discrimination

This study demonstrated the utility and strength of the new generation Sentinel-2 MSI sensor characteristics in discriminating disease infections in commercial crops such as coffee- a previously challenging task from broadband multispectral sensor characteristics (e.g. Landsat series). For example, the results of this study have successful managed to identify Sentinel-2 MSI's most important individual spectral bands (i.e. B4, B6 and B5) related to leaf conditions induced by CLR infection. The results from the RF OOB score show that the Sentinel-2 RE3 band (B7) provides the most significant variables for CLR discrimination. This confirms the importance of this region of the spectrum in fungal disease detection and discrimination in vegetation as reported in previous studies. For instance, Mahlein et al. (2013) observed that the greatest shift in reflectance of diseased plants compared to healthy plants occurs in the red edge region of the spectrum. Furthermore, Devadas et al. (2014) achieved a large increase in discrimination of healthy, nitrogen deficient and yellow rust infected winter wheat by inclusion of the band centered at 725 nm, indicating the role of the red edge in crop condition assessments.

The finding that the red band (B4) was ranked by both the RF OOB score and the PLS-DA VIP score as the most significant band was rather surprising as red edge bands were expected to be more significant. However, this could be because CLR is different from other diseases and plant stress conditions in that it is not necrotic and also results in distinctive brownish pustules on the underside of the leaf (Belan et al. 2015; Honorato Júnior et al. 2015). The reflectance measurements were done on the adaxial leaf side while the symptoms of the disease are evident on the abaxial leaf side. Therefore, while studies have generalized the approaches for remote sensing plant diseases, it is clear from this work that the application of spectral bands and vegetation indices in disease discrimination could be disease and plant specific, given the observed peculiarities.

## Performance of Sentinel-2 vegetation indices in CLR discrimination

The general finding from this study was that the use of spectral vegetation indices derived from the recently launched Sentinel-2 MSI band settings produced plausible CLR discrimination results, when compared to the use of the sensor's raw spectral bands. For instance, the use of the derived vegetation indices yielded higher overall, user and producer accuracies of CLR levels (i.e. healthy, moderate and severe CLR), when compared to the use of raw spectral bands as an independent dataset. This observation is in line with previous findings that concluded that the spectral transformation that occurs in converting spectral bands to vegetation indices enables more information to be obtained resulting in better performance from vegetation indices. For examples Dube & Mutanga (2015) achieved higher model fit in biomass estimation using vegetation indices ($R^2 = 0.53$) than by spectral bands ($R^2 = 0.40$) with the RF algorithm. Similarly, other studies reported better performance of vegetation indices than spectral bands in biomass estimation, plant water content estimation and plant condition assessments (Hill 2013; Wang et al. 2004).

In addition, the better performance of vegetation indices could also be attributed to the ability of Sentinel-2 MSI band settings derived vegetation indices to deal with confounding factors, such as reflectance saturation, leaf area, roughness and moisture in the leaf and canopy that reduce the performance of raw spectral bands. For example, it is known that narrow band vegetation indices like the ones obtainable from Sentinel-2 MSI sensor settings are capable of reducing the effects of asymptotic saturation common in raw

reflectance and broadband vegetation indices (Baret and Buis, 2008; Glenn et al. 2008; Mutanga and Skidmore 2004). This could be because Sentinel-2 MSI derived vegetation indices are more sensitive to plant biochemical and biophysical properties, as they are a combination of two or more strategically positioned spectral bands solely designed for vegetation condition assessment. This clearly shows that the transformations and combinations of the bands that were identified as important for CLR discrimination into vegetation indices resulted in better accuracy than expected (Mahlein et al. 2013; Sankaran et al. 2010).

Overall, this study has demonstrated the performance of the recently-launched Sentinel-2 MSI sensor characteristics in accurate crop diseases monitoring, especially in data scarce areas. The observed performance of Sentinel-2 MSI band settings is in line with the results observed from previous work that associated the good performance with its improved design that includes the presence of multiple strategically positioned bands i.e. three red edge bands centered at 703, 740 and 783 nm. Most of these are coupled with relatively narrow spectral bands (≤30 nm width) and available at high spatial resolution (≤20 m), all of which are previously missing components in multispectral sensors, such as Landsat, SPOT and ASTER (Dangwal et al. 2016; Lanfredi et al. 2015; Rembold et al. 2013; Rulinda et al. 2012). In addition, Sentinel-2 offers an unprecedented short revisit time for multispectral sensors (about weekly due to a constellation of two identical satellites). A combination of these advantages offers opportunities for CLR detection, mapping and monitoring in both space and time useful for the coffee sector.

CLR discrimination with Sentinel-2 data could inform farm managers and plant phytosanitary regulators with information on occurrence and levels of severity of plant diseases, reducing the subjectivity, costs, spatial singularity and inconveniences associated with field methods. In addition, the high temporal resolution of Sentinel-2 can provide opportunities for modelling CLR rate and direction of spread by combining analysis in both time and space over fields. Therefore, Sentinel-2 MSI has the potential to freely bring information on plantation crop condition that current multispectral sensor characteristics could not. These opportunities have to be considered with limits to practical applications of the technology under field conditions. Factors such as age of the crop for example may have significant influence on the potential application of aerial or satellite imagery in crop condition assessments (Chemura and Mutanga 2016).

## Effect of variable optimization and algorithm on CLR discrimination

Results have also shown that that variable optimization of RF and PLS-DA algorithms improves the accuracy of CLR discrimination, when compared to model implementation without variable selection. For instance, in this study, it has been observed that variable selection significantly improved CLR discrimination by about 10% for bands and 6% for vegetation indices when classified with RF. There are many reasons why variable optimization is able to achieve better results that using all available variables. These multispectral sensors have been developed for wide purposes ranging from forestry, agricultural, water and urban applications (although Sentinel-2 has a bias towards vegetation) and therefore, just a few parameters may be fit for purpose. In addition, a number of wavebands maybe correlated or do not represent useful information making optimized variables to achieve better results when redundant features are removed. The remaining few wavebands and vegetation indices therefore had important relationships with CLR infection levels and masking out these superfluous variables that were unnecessarily increasing noise and data handling challenges in machine learning algorithms improves the accuracy as obtained in

this study (Pal and Foody 2010; Saeys et al. 2007). Therefore, although the wavebands in multispectral data have been strategically selected, there remains a need to find the best model parameters from these bands. This is even despite the fact that the number of variables was small (7 for spectral bands and 17 for vegetation indices), overfitting is a known problem in PLS methods (Frampton et al. 2013; Wang et al. 2011) and therefore selecting a few important variables is expected to improve the performance of the machine learning methods.

The effect of variable selection/optimization was even more remarkable in the RF algorithm that is reported to be robust to noise, has an ability to handle non-linear data and is able to deal with large numbers of variables (Gislason et al. 2004; Mutanga et al. 2012; Rodriguez-Galiano et al. 2012). These strengths of the RF explain the observed superior performance of the RF over PLS-DA, making it a better candidate for use in CLR discrimination with recently launched Sentintel-2 MSI data. This is important because in order to improve effectiveness of remote sensing applications, retrieval algorithms should be accurate, fast, robust and sufficiently flexible to make use of the Sentinel-2 spectral bands and vegetation indices. It appears from this study that the RF algorithm is fit for this purpose with an added advantage of explaining how the output is obtained through ranking the weights of input variables. The obtained accuracies are good for application but they could have been better. The misclassification of disease levels could be because of the fact that the CLR shows disease signs on the abaxial side of the leaf and yet reflectance measurements are taken on the adaxial with difficulties for discriminating early stages of infection. In addition, the largest misclassification occurred for moderate and severe CLR possibly because the disease could be considered moderate when it has already done severe internal damage resulting in the errors.

## Conclusion

The aim of this work was to explore the utility of the Sentinel-2 MSI characteristics in detecting and discriminating low and advanced levels of coffee leaf rust caused by *Hemileia vastatrix*, using two classification algorithms (RF and PLS-DA). Results have demonstrated the ability of the recently launched Sentinel-2 MSI sensor settings for discriminating CLR infection levels (i.e. healthy, moderate and severe) with high accuracy. Specifically, this work has demonstrated that Sentinel 2 MSI derived bands and vegetation indices, computed using spectral information located at the red-edge position, are useful for disease detection or assessments of crop status. Also, the study has shown that optimized vegetation indices and spectral bands perform better in CLR discrimination when compared to the use of all variables as an independent dataset. This study therefore underpins the potential use of Sentinel-2 MSI data settings in crop and vegetation state assessment that can improve management of croplands and stewardship of the environment through reduced unnecessary use of crop protection chemical for disease control.

# References

Avelino, J., Romero-Gurdián, A., Cruz-Cuellar, H. F., & Declerck, F. A. (2012). Landscape context and scale differentially impact coffee leaf rust, coffee berry borer, and coffee root-knot nematodes. *Ecological Applications, 22*(2), 584–596.

Avelino, J., Willocquet, L., & Savary, S. (2004). Effects of crop management patterns on coffee rust epidemics. *Plant Pathology, 53*(5), 541–547.

Baker, P., Bentley, J., Charveriat, C., Dugne, H., Leftoy, T., & Munyua, H. (2001). The Coffee Smallholder. In P. Baker (Ed.), *Coffee Futures: A source book of some critical issues confronting the coffee industries*. Chinchina: Cabi-Federacafe-USDA-ICO.

Barbedo, J. G. A. (2013). Digital image processing techniques for detecting, quantifying and classifying plant diseases. *Spinger Plus, 2*(660), 1–12.

Baret, F., & Buis, S. (2008). Estimating canopy characteristics from remote sensing observations: Review of methods and associated problems. In S. Liang (Ed.), *Advances in land remote sensing: System, modeling, inversion and application* (pp. 173–201). Amsterdam, The Netherlands: Springer.

Baret, F., & Guyot, G. (1991). Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sensing of Environment, 35,* 161–173.

Barnes, E. M., Clarke, T. R., Richards, S. E., Colaizzi, P. D., Haberland, J., Kostrzewski, M., et al. (2000). Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In: P. C. Robert, R. H. Rust, W. E. Larson (Eds.) *Proceedings of the 5th international conference on precision agriculture* (pp 16–19). Madison, USA: American Society of Agronomy

Belan, L. L., Pozza, E. A., de Oliveira Freitas, M. L., Pozza, A. A. A., de Abreu, M. S., & Alves, E. (2015). Nutrients distribution in diseased coffee leaf tissue. *Australasian Plant Pathology, 44*(1), 105–111.

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

Breiman, L., & Cutler, A. (2007). *Random forests-classification description*. Berkeley, USA: Department of Statistics.

Brown, H. B. (2008). *Smallholder coffee production in Zimbabwe*. Harare, Zimbabwe: CRS Press.

Brown, J. K. M., & Hovmøller, M. S. (2002). Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. *Science, 297,* 537–541.

Cao, K.-A. L., Gonzalez, I., & Dejean, S. (2015). Package 'mixOmics'. Retrieved 1 December, 2016 from http://www.mixOmics.org.

Carter, G. A., & Knapp, A. K. (2001). Leaf optical properties in higher plants: linking spectral characteristics to stress and chlorophyll concentration. *American Journal of Botany, 88*(4), 677–684.

Chemura, A., Kutywayo, D., Chidoko, P., & Mahoya, C. (2015). Bioclimatic modelling of current and projected climatic suitability of coffee (*Coffea arabica*) production in Zimbabwe. *Regional Environmental Change, 16*(2), 473–485.

Chemura, A. & Mutanga, O. (2016). Developing detailed age-specific thematic maps for coffee (*Coffea arabica* L.) in heterogeneous agricultural landscapes using random forests applied on Landsat 8 multispectral sensor. *Geocarto International*, 1–18 (inpress).

Clevers, J. G. P. W., & Gitelson, A. A. (2013). Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *International Journal of Applied Earth Observation and Geoinformation, 23,* 344–351.

Coops, N., Stanford, M., Old, K., Dudzinski, M., Culvenor, D., & Stone, C. (2003). Assessment of Dothistroma needle blight of Pinus radiata using airborne hyperspectral imagery. *Phytopathology, 93*(12), 1524–1532.

Cressey, D. (2013). Coffee rust regains foothold. *Nature, 493,* 587.

D'Odorico, P., Gonsamo, A., Damm, A., & Schaepman, M. E. (2013). Experimental evaluation of Sentinel-2 spectral response functions for NDVI time-series continuity. *IEEE Transactions of GeoScience & Remote Sensing, 51*(3), 1336–1348.

Dangwal, N., Patel, N., Kumari, M., & Saha, S. (2016). Monitoring of water stress in wheat using multispectral indices derived from Landsat-TM. *Geocarto International, 31*(6), 682–693.

de Almeida, M. R., Correa, D. N., Rocha, W. F. C., Scafi, F. J. O., & Poppi, R. J. (2013). Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. *Microchemical Journal, 109,* 170–177.

Devadas, R., Simpfendorfer, S., Backhouse, D., & Lamb, D. W. (2014). Effect of stripe rust on the yield response of wheat to nitrogen. *The Crop Journal, 2*(4), 201–206.

Dinesh, K. P., Shivanna, P., & Santa Ram, A. (2011). Identification of RAPD (Random Amplified Polymorphic DNA) markers for Ethiopian wild *Coffea arabica* L genetic resources in the tropics. *Research: plant Genomics, 2*(11), 1–7.

Dube, T., & Mutanga, O. (2015). Evaluating the utility of the medium-spatial resolution Landsat 8 multi-spectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa. *ISPRS Journal of Photogrammetry and Remote Sensing, 101,* 36–46.

Eitel, J. U. H., Gessler, P. E., Smith, A. M. S., & Robberecht, R. (2006). Suitability of existing and novel spectral indices to remotely detect water stress in *Populus* spp. *Forest Ecology and Management, 229*(1–3), 170–182.

Eitel, J. U. H., Vierling, L. A., Litvak, M. E., Long, D. S., Schulthess, U., Ager, A. A., et al. (2011). Broadband, red-edge information from satellites improves early stress detection in a New Mexico conifer woodland. *Remote Sensing of Environment, 115*(12), 3640–3646.

Feng, W., Shen, W., He, L., Duan, J., Guo, B., Li, Y., et al. (2016). Improved remote sensing detection of wheat powdery mildew using dual-green vegetation indices. *Precision Agriculture, 17*(5), 608–627.

Foody, G. M. (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing, 70*(5), 627–633.

Frampton, W. J., Dash, J., Watmough, G., & Milton, E. J. (2013). Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing, 82,* 83–92.

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2012). An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts', USA. *GIScience & Remote Sensing, 49*(5), 623–643.

Ghini, R., Bettiol, W., & Hamada, E. (2011). Diseases in tropical and plantation crops as affected by climate change: Current knowledge and perspectives. *Plant Pathology, 60,* 122–132.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2004). Random forest classification of multisource remote sensing and geographic data. In: *Geoscience and remote sensing symposium, 2004 (IGARSS'04)* (Vol 2, pp. 1049–1052). Piscataway: IEEE International

Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment, 58,* 289–298.

Gitelson, A. A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology, 22*(3), 247–252.

Gitelson, A. A., Viña, A., Ciganda, V., Rundquist, D. C., & Arkebauer, T. J. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters, 32,* L08403.

Glenn, E. P., Huete, A. R., Nagler, P. L., & Nelson, S. G. (2008). Relationship between remotely-sensed vegetation indices, canopy attributes and plant physiological processes: What vegetation indices can and cannot tell us about the landscape. *Sensors, 8*(4), 2136–2160.

Gresser, C., & Tickell, S. (2002). *Mugged: Poverty in your coffee cup.* Boston, USA: Oxfam.

Haddad, F., Maffia, L. A., Mizubuti, E. S., & Teixeira, H. (2009). Biological control of coffee rust by antagonistic bacteria under field conditions in Brazil. *Biological Control, 49*(2), 114–119.

Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment, 122,* 66–74.

Hedley, J., Roelfsema, C., Koetz, B., & Phinn, S. (2012). Capability of the Sentinel 2 mission for tropical coral reef mapping and coral bleaching detection. *Remote Sensing of Environment, 120,* 145–155.

Hill, M. J. (2013). Vegetation index suites as indicators of vegetation state in grassland and savanna: An analysis with simulated Sentinel 2 data for a North American transect. *Remote Sensing of Environment, 137,* 94–111.

Honorato Júnior, J., Zambolim, L., Aucique-Pérez, C. E., Resende, R. S., & Rodrigues, F. A. (2015). Photosynthetic and antioxidative alterations in coffee leaves caused by epoxiconazole and pyraclostrobin sprays and *Hemileia vastatrix* infection. *Pesticide Biochemistry and Physiology, 123,* 31–39.

Huang, W., Lamb, D., Niu, Z., Zhang, Y., Liu, L., & Wang, J. (2007). Identification of yellow rust in wheat using in situ spectral reflectance measurements and airborne hyperspectral imaging. *Precision Agriculture, 8*(4–5), 187–197.

Kutywayo, D., Chemura, A., Kusena, W., Chidoko, P., & Mahoya, C. (2013). The impact of climate change on the potential distribution of agricultural pests: The case of the coffee white stem borer (*Monochamus leuconotus* P.) in Zimbabwe. *PLOS ONE, 8*(8), e73432.

Lanfredi, M., Coppola, R., Simoniello, T., Coluzzi, R., Imbrenda, V., & Macchiato, M. (2015). Early identification of land degradation hotspots in complex bio-geographic regions. *Remote Sensing, 7*(6), 8154–8179.

Laudien, R., Bareth, G., & Doluschitz, R. (2004). Comparison of remote sensing based analysis of crop diseases by using high resolution multispectral and hyperspectral data: Case study: Rhizoctonia solani in sugar beet. In: S.A. Brandt (Ed.) *Proceedings of the 12th International Conference on Geoinformatics, Gävle* (pp. 670-676). Sweden: Gävle University Press.

Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., et al. (2014). Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical, 6,* 115–125.

Li, X.-L., Yi, S.-L., He, S.-L., Lv, Q., Xie, R.-J., Zheng, Y.-Q., et al. (2016). Identification of pummelo cultivars by using Vis/NIR spectra and pattern recognition methods. *Precision Agriculture, 17*(3), 365–374.

Liaw, A., Wiener, M., Breiman, L., & Cutler, A. (2009). Package "randomForest". Retrieved 1 December, 2016 from https://cran.r-project.org/package=randomForest.

Logan, W. J. C., & Biscoe, J. (1987). *Coffee handbook.* Harare: Zimbabwe Coffee Growers' Association.

Mahlein, A. K., Rumpf, T., Welke, P., Dehne, H.-W., Plümer, L., Steiner, U., et al. (2013). Development of spectral indices for detecting and identifying plant diseases. *Remote Sensing of Environment, 128,* 21–30.

Mutanga, O., Adam, E., & Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation, 18,* 399–406.

Mutanga, O., & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing, 25*(19), 3999–4014.

Mutanga, O., & Skidmore, A. K. (2007). Red edge shift and biochemical content in grass canopies. *ISPRS Journal of Photogrammetry and Remote Sensing, 62*(1), 34–42.

Nitze, I., Schulthess, U., & H., A. (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. In: R.Q. Feitosa, G.A.O.P. Costa, C. M. Almeida, L. M. G. Fonseca, H.J.H. Kux (Eds.) *Proceedings of the 4th GEOBIA* (pp. 35–40). Rio de Janeiro, Brazil: Brazilian National Institute for Space Research.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing, 48*(5), 2297–2307.

Prabhakar, M., Prasad, Y. G., Desai, S., Thirupathi, M., Gopika, K., Rao, G. R., et al. (2013). Hyperspectral remote sensing of yellow mosaic severity and associated pigment losses in *Vigna mungo* using multinomial logistic regression models. *Crop Protection, 45,* 132–140.

R Core Team. (2013). R: a language and environment for statistical computing. Vienna, Austria. Retrieved from 31 Novemer 2016 http://www.R-project.org/.

Rembold, F., Atzberger, C., Savin, I., & Rojas, O. (2013). Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sensing, 5,* 1704–1733. doi:10.3390/rs5041704.

Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 67,* 93–104.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., & Roberts, D. (2008). Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. *Remote Sensing of Environment, 112*(5), 2272–2283.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the Great Plains with ERTS. In: S.C. Freden, E.P. Mercanti, M.A. Becker (Eds.) *Proceedings of the third earth resources technology satellite-1 symposium, Greenbelt, Maryland, USA* (pp 309).

Rulinda, C. M., Dilo, A., Bijker, W., & Stein, A. (2012). Characterising and quantifying vegetative drought in East Africa using fuzzy modelling and NDVI data. *Journal of Arid Environments, 78,* 169–178.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517.

Sankaran, S., Mishra, A., Ehsani, R., & Davis, C. (2010). A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture, 72,* 1–13.

Savitzky, A., & Golay, M. (1964). Smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry, 36*(8), 1627–1638.

Silva, M. C., Várzea, V., Guerra-Guimarães, L., Azinheira, H. G., Fernandez, D., Petitot, A.-S., et al. (2006). Coffee resistance to the main diseases: Leaf rust and coffee berry disease. *Brazilian Journal of Plant Physiology, 18*(1), 119–147.

Stone, C., Chisholm, L., & Coops, N. (2001). Spectral reflectance characteristics of eucalypt foliage damaged by insects. *Australian Journal of Botany, 49*(6), 687–698.

Suresh, N., Santa, R. A., & Shivanna, M. B. (2012). Coffee leaf rust(CLR) and disease triangle: A case study. *International Journal of Food, Agriculture and Veterinary Sciences, 2*(2), 50–55.

Ustin, S. L., Gitelson, A. A., Jacquemoud, S., Schaepman, M., Asner, G. P., Gamon, J. A., et al. (2009). Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment, 113,* S67–S77.

Vincini, M., Amaducci, S., & Frazzi, E. (2014). Empirical estimation of leaf chlorophyll density in winter wheat canopies using Sentinel-2 spectral resolution. *IEEE Transactions of GeoScience & Remote Sensing, 52*(6), 3220–3235.

Wang, C., Fritschi, F. B., Stacey, G., & Yang, Z. W. (2011). Phenology-based assessment of perennial energy crops in North American tallgrass prairie. *Annals of the Association of American Geographers, 101*(4), 742–751.

Wang, Q., Tenhunen, J., Dinh, N. Q., Reichstein, M., Vesala, T., & Keronen, P. (2004). Similarities in ground- and satellite-based NDVI time series and their relationship to physiological activity of a scots pine forest in Finland. *Remote Sensing of Environment, 93*(1–2), 225–237.