ISPRS Journal of Photogrammetry and Remote Sensing 167 (2020) 85–103

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

The potential of in-situ hyperspectral remote sensing for differentiating 12 banana genotypes grown in Uganda

Priyakant Sinha[a,*], Andrew Robson[a], Derek Schneider[a], Talip Kilic[b], Harriet Kasidi Mugera[e], John Ilukor[b,d], Jimmy Moses Tindamanyire[c]

[a] Applied Agricultural Remote Sensing Centre (AARSC), University of New England, Armidale 2351, NSW, Australia
[b] Living Standards Measurement Study (LSMS), Development Data Group, The World Bank, Via Labicana 110, 00184 Rome, Italy
[c] National Banana Research Programme, National Agricultural Research Laboratories (NARL), Kawanda, P.O. Box 7065, Kampala, Uganda
[d] CGIAR Standing Panel on Impact Assessment (SPIA), P.O. Box 24384, Naguru, Kampala, Uganda
[e] Development Data Group, World Bank, 1818 H Street, N.W. Washington, DC 20433, USA

ARTICLE INFO

Keywords:
Banana
Hyperspectral remote sensing
Prediction modelling
Agriculture productivity
Survey design

ABSTRACT

Bananas and plantains provide food and income for more than 50 million smallholder farmers in East and Central African (ECA) countries. However, banana productivity generally achieves less than optimal yield potential (< 30%) in most regions, including Uganda. Numerous studies have been undertaken to identify the key challenges that smallholder banana growers face at different stages of the banana value chain, with one of the main constraints being a lack of policy-relevant agricultural data. The World Bank (WB) initiated a methodological survey design aimed at identifying the distribution of banana varieties across a number of key Ugandan growing regions, at the individual household scale. To achieve this outcome a number of approaches including ground-based surveys, DNA tissue collection of selected banana plants and remote sensing were evaluated. For the remote sensing component, the set objectives were to develop statistical models from the hyperspectral reflectance properties of individual leaves that could differentiate typical ECA banana varieties, as well as their parentage (usage). The study also explored the potential of extrapolating the ground-based hyperspectral measures to high-resolution WorldView-3 (WV3) satellite imagery, therefore creating the potential of mapping the distribution of banana varieties at a regional scale. The DNA testing of 43 banana varieties propagated at the National Banana Research Program site at National Agricultural Research Organization (NARO) research station in Kampala, Uganda, identified 12 genetically different varieties. A canonical powered partial least square (CPPLS) model developed from hyperspectral reflectance properties of the sampled banana leaves successfully differentiated BLU, BOG, GON, GRO and KAY genotypes. The Random Forest (RF) algorithm was also evaluated to determine if spectral bands coinciding with those provided by WV3 data could segregate banana varieties. The results suggested that this was achievable and as such presents an opportunity to extrapolate the hyperspectral classifications to broader areas of land. The ability to spectrally differentiate these five genotypes has merit as they are not typical east African varieties. As such, identifying the distribution and density of these varieties across Uganda provides vital information to the banana breeders of NARO of where their new varieties are being disseminated too, data that has been previously difficult to obtain. Although the results from this pilot study indicated that not all banana varieties could be spectrally differentiated, the methodology developed and the positive results that were achieved do present remote sensing as a complimentary technology to the ongoing surveying of banana and other crop types grown within Ugandan household farming systems.

1. Introduction

The continued development of agricultural systems is essential for combating poverty and food security. However, establishing a baseline of agricultural statistics to assist government policy making and investment decisions is somewhat difficult. The Food and Agriculture Organization (FAO) has expended considerable effort deriving worldwide agricultural statistics on a range of crops which has proven invaluable to researchers and practitioners in the field of agriculture (Atzberger, 2013; Carfagna and Gallego, 2005; FAO, 2017a,b).

* Corresponding author.
E-mail address: psinha2@une.edu.au (P. Sinha).

https://doi.org/10.1016/j.isprsjprs.2020.06.023
Received 4 February 2020; Received in revised form 26 June 2020; Accepted 29 June 2020
Available online 17 July 2020

The Living Standards Measurement Study Integrated Surveys on Agriculture (LSMS-ISA) of the World Bank has been developing and implementing innovative survey methods to generate policy-relevant agricultural data in support of governments in Sub-Saharan Africa (Christiaensen and Demery, 2017; Kosmowski et al., 2019). Whilst, the collection of data in most cases is specifically focused on improving agricultural productivity, the WB-LSMS, in collaboration with the CGIAR (Consultative Group for International Agricultural Research) Standing Panel on Impact Assessment (SPIA), has been supporting the design, implementation and analysis of household survey data for crop variety identification (Christiaensen and Demery, 2017; Lobell et al., 2018). This methodology assesses the relative accuracy of subjective data collection approaches that are typically part of household and farm surveys in relation to the DNA fingerprinting of crop material sampled from the farmers' fields. The LSMS-ISA emphasizes the design and validation of innovative survey methods, through the use of technology such as remote sensing for improving survey data quality, and the development of analytical tools to facilitate the analysis of data collected.

Bananas and plantains (Musa spp.) are important economic resources for rural farmers in Uganda with a total annual estimated production of ~10MT (FAO, 2011; UBos, 2010). Banana production is essential for both ongoing food and income security because of its all-year-round fruiting and ability to grow in a wide range of environments and farming systems (Tinzaara et al., 2018; Tripathi et al., 2007). In Uganda, banana is used for food, beverages, snacks, livestock feed, industrial spirits and for several crafts and medicinal use (Anyasi et al., 2013). However, like any other East-Central African (ECA) country, Uganda's banana productivity and market potential within and outside the region are grossly underutilised (Kiiza et al., 2004; Van Asten et al., 2003). Few studies have highlighted the challenges and scopes for small farm holdings in banana value chain (e.g., Kiiza et al., 2004; Nyombi, 2013; PARAM, 2015; Tinzaara et al., 2018). Carletto et al. (2015) identified that the policy-relevant smallholder agricultural data was inconsistent, confined to sectors or institution (lack of data sharing), and methodologically weak. Thus, there was a need for a standard methodological survey design to provide quality of smallholder agriculture data for effective decision-making. This includes the validation of innovative survey methods such as from remotely sensing data.

Use of plant leaf-reflectance to predict plant parameters including crop variety, bio-chemical and physiological status, etc. is well established (Jacquemoud et al., 2009; Jay et al., 2017; Martínez-Martínez et al., 2018; Shi et al., 2015; Silva-Perez et al., 2017). Spectral reflectance in the broad, multispectral visible/near-infrared portion of the electromagnetic radiation (EMR) have been related to plant chlorophyll, plant health or vigour and water content, while the Red-Edge has been commonly related to photosynthesis and foliar nitrogen content (Féret et al., 2017; Jay et al., 2017; Peñuelas and Filella, 1998; Silva-Perez et al., 2017; Verrelst et al., 2015). Recently, the increased availability of hyperspectral cameras and field spectrometers, offering the full spectrum (i.e. 350–2500 nm) (NIR: 770–1300, SWIR1: 1300–1900 nm and SWIR2: 1900–2500 nm), have provided measures of an increasing range of crop biophysical traits (Ajayi et al., 2016; Duan et al., 2014; Martínez-Martínez et al., 2018; Mishra et al., 2017; Sahoo et al., 2015). For example, leaf-level reflectance has been correlated with photosynthetic parameters for wheat, cotton, potato, sunflower, common beans and maize (Duan et al., 2014; Martínez-Martínez et al., 2018; Silva-Perez et al., 2017; Chivasa et al., 2019); nitrogen content (Shi et al., 2015); and leaf dry mass (Wang et al., 2010). Vijaya Kumar et al. (2005) used hyperspectral radiometer data for evaluation of four varieties of castor beans (VP-1, 48-1, GCH-4, and Aruna) towards their tolerance to drought, characterized by higher NIR reflectance. Garriga et al. (2017) investigated non-invasively measure wheat traits and differentiated their genotypes. In another study, Ajayi et al. (2016) analysed spectral behaviour of 20 wheat genotypes of wide genetic background in relation to crop growth parameters, leaf area index (LAI) and yield, and found the MIR/SWIR regions to be the most sensitive. These studies indicate the potential of discriminating crop variety via their spectral properties. The visible portions of the spectrum are sensitive to the colour of the foliage; the reflectance in the NIR sensitive to plant structure; and the mid NIR to the presence, absence or variation in the quantity of leaf constituents. The leaf constitutes such as oil, fats, cellulose and lignins, all of which can be cultivar-specific. It is important to note that local and environmental factors can significantly affect cultivar discrimination, and that the studies on variety discrimination are often conducted for crop types grown on similar environmental and management conditions, i.e. controlled environments. Mishra et al. (2017) presented a review on hyperspectral imaging of plants, including challenges and complexities related to external and plant-related factors and the technical challenges in the assessment of plant traits.

Hyperspectral sensors provide very high spectral resolution both in terms of the spectral extent covered, and the number of narrow band wavelengths available. As such these sensors are especially sensitive to very small variations in plant varietal composition (structure, leaf constituents, colour, etc.). The discrimination of crop varieties rely on the processing and analysis techniques that are capable of isolating such smaller physiological variations (predictees or response) through determination of least collinear wavelengths (predictors) (Rapaport et al., 2015; Suarez et al., 2016; Silva-Perez et al., 2017), and then classification or regression analysis to predict biophysical properties in space and time. Several machine learning (ML) algorithms such as Random forests (RF) and Partial least squares regression (PLSR)) were found powerful tools for such tasks (e.g., Chlingaryan et al., 2018; Fu et al., 2019; Heckmann et al., 2017; Mountrakis et al., 2011; Schwieder et al., 2014; Silva-Perez et al., 2017). These algorithms generate adaptive, robust relationships and, with optimal experimental design and training, are fast to apply. The RF models the relationship between explanatory variables and response variables by a set of decision rules (Breiman, 2001). The RF classifiers potentially resolve the overfitting problem and have been used in crop parameter and type classifications (Belgiu and Drăguţ, 2016; Crabbe et al., 2020; Fu et al., 2019; Zhao et al., 2016). The use of RF is an advantage, especially in situation of small sample size and its characteristics to produce a variable importance ranking in the classification (e.g., Fletcher and Reddy, 2016; Fletcher, 2016). This is particularly useful to user in selecting variables to design simpler and effective models (Liaw and Wiener, 2002; Strobl et al., 2008). Despite an increase in the use of ML algorithms for different applications, the techniques have some fundamental limitations and require expert knowledge in parametrizations. Ali et al. (2015) listed the advantages and shortcomings of some of these techniques.

The PLSR model has been used to estimate photosynthetic capacity at the leaf level from leaf-clip reflectance spectra (Serbin et al., 2011; Silva-Perez et al., 2017). The model effectively deals with large number, multicollinear variables, where the number of explanatory variables is greater than the number of observations (Wold et al., 2001). The PLSR constructs predictive models to identify and extract the variables or components that mostly explain the variability (covariance) of both Y (response) and X (predictor) variables while employing a stopping rule to find the optimal number of components (ONC). For each latent variable, the regression coefficient is estimated through a cross validation approach. The ONC is determined by minimizing RMSE between predicted and observed response variable (Esbensen et al., 2002; Mevik and Wehrens, 2007; Suarez et al., 2016). The PLSR is commonly used for the interrogation of large continuous datasets, such as hyperspectral data, when determining those regions most sensitive to specific parametric variations (Barnes et al., 2017; Mevik and Wehrens, 2007) including those related to crop variety (e.g., Heckmann et al., 2017; Silva-Perez et al., 2017; Suarez et al., 2016). It also avoids the potential overfitting problems typical with other predictive models (Hansen and Schjoerring, 2003; Mevik and Wehrens, 2007). However, the performance of PLSR was found inconsistent with different plant species, regions, and growth environments (Fu et al., 2019). Indahl et al. (2009)

proposed a new data compression method, a Canonical Partial Least Square (CPLS), for estimating optimal latent variables when more than one response variable is available. The latent variables are found by combining PLS and canonical correlation analysis (CCA). The model predicts information more effectively than ordinary PLS approaches as it incorporates information from additional variables to improve predictions (Indahl et al., 2009). A canonical powered PLS (CPPLS) is an extension to CPLS that incorporates additional responses, individual weighting of observations and power methodology to further improve the predictive performance of the model (Indahl et al., 2009; Liland and Indahl, 2009). Numerous studies have identified the CPPLS approach as being more effective in the prediction of crop biophysical properties and yield from hyperspectral reflectance than the traditional PLS method (e.g., Øvergaard et al., 2013a,b; Suarez et al., 2017).

Previous remote sensing research on banana has included Rajkumar et al. (2012) who evaluated a lab-based visible-NIR imaging technique for banana maturity prediction and industrial sorting of banana fruit quality based on the chlorophyll characteristic and PLSR; and Johansen et al. (2009, 2014) who undertook satellite based mapping studies to determine the location of individual banana plants in peri-urban regions of Australia. However, there is presently, no prior reporting of using hyperspectral reflectance data as an input into ML algorithms such as CPPLS and RF for banana genotype discrimination. This research aims to fill this information gap by developing statistical models from the hyperspectral reflectance properties of banana leaves to differentiate typical ECA banana varieties and their parentage (usage, old or local and new or improved). The main hypothesis is that the genotype-related differences (e.g. leaf constituents and canopy architecture) produce different spectral responses that can be detected with hyperspectral sensors. In a novel approach, this study also attempts to extrapolate the in-situ measures of banana plants to the regional scale, and explore the feasibility of pairing point source hyperspectral measures (leaf-level) collected with an ASD field spectrometer with high-resolution satellite imagery WV3 (canopy-level).

2. Objectives

This study examines the potential of remote sensing for differentiating banana varieties in Uganda and as such contribute vital information to the methodological survey experiment. The specific objectives are:

(a) Generation of hyperspectral calibration dataset for differentiating banana genotypes commonly grown in the smallholder-based agricultural systems of Uganda;
(b) Determining the relative accuracy of spectral prediction of banana varieties based on DNA fingerprinting;
(c) Determining the potential of extrapolating varietal classifications of individual banana plants achieved from on-ground hyperspectral measures to 16-spectral bands of WV3 satellite imagery.

3. Materials and methods

3.1. Site selection

To determine if banana varieties could be spectrally differentiated, a calibration set of 43 banana varieties were selected from the National Banana Research Program site at National Agricultural Research Organization (NARO) research station in Kampala, Uganda (Fig. 1). The selection of banana varieties were based on information provided by the NARO staff using variety names, types and morphological attributes. The 43 selected banana varieties comprised of both Native or Old varieties, which were commonly grown across the different Ugandan regions, as well as those of introduced or improved varieties (New). The term ‘Parentage’ is synonymously used for ‘Old/New’ varieties or ‘Usage’, all of them essentially meaning the same. The NARO research station presented a strong candidate for the development of a spectral calibration dataset as all cultivars were grown in close proximity and were managed by NARO staff. Therefore, the risk of varied abiotic and biotic factors influencing the spectral responses including management, soil type or moisture variability and crop age was minimized. In addition, for the establishment of a validation data set, spectral measurements of banana plants were done from a number of varieties grown across three Ugandan districts: Isingiro, Masaka and Mbarara (results not presented here). The geographic coordinates of study site are between 32°30′E to 32°35′E longitude and 0°40′N to 0°45′N latitude. The climate is tropical savanna, with the average annual temperature and precipitation are 26 °C and 356 mm, respectively. Most of the rainfall occurs during April-May and Oct-Nov months.

3.2. Field spectroscopy

Field sampling was conducted during the late January and early February of 2018. To establish a spectral data set for each of the banana varieties, the first fully extended leaf of 30 plants per banana variety (where available) was selected and manually removed with a knife (Fig. 2a). This leaf was identified to be optimal as it was considered to be photosynthetically active, unlike the pale green cigar leaf (leaf developmental stage), and less likely to be influenced by disease, pest, weather damage and senescence as displayed by the older leaves. Each manually removed leaf was spectrally measured immediately after removal to minimise the potential influences of desiccation and the breaking down of internal leaf structures (Fig. 2b). The manual harvesting of leaves to undertake in situ measurements was found more practical due to two main reasons: (a) with instrument in hand, reaching out different parts of a leaf was difficult and dangerous due to the height of the plants; and (b) as a large number of leaves were used for sampling it was logistically more feasible to manually remove the third leaves and assemble them in combined sample set for rapid measurement, than traversing around the plantation with the equipment. The spectral reflectance measures were undertaken with an ASD FieldSpec 4 spectrometer Hi-Res (350–2500 nm) fitted with a Leaf clip attachment with a sampling resolution of 1 nm. This attachment provides a target area of interest of 2 cm and as it creates its own light source from a 4.25 V 4.5 Watt Halogen lamp (MR6), hence measures were not subject to the influences of differing ambient light conditions from cloud cover, dust, smoke etc. This was an important consideration as varying levels of cloud cover; haze and dust were persistent throughout the entirety of the field sampling campaign. An instrument was set to internal averaging of 10 raw scans for a single measurement. Preventive protocols such as uniform sensor-target distance, sampling at lamina portion of leaf and not over veins and midrib, were taken to avoid any abnormal spectral reflectance during the raw data collection. The leaf sampling and scanning were time-consuming and for each variety generally extended over many hours. Unlike seasonal influence on leaf scale spectra (Wong and Gamon, 2015), the variation in photosynthetic process during the day of sampling was not considered to influence the resultant spectra between the samples due to following reasons – (a) any influence of this dynamic process may only occur in certain wavelengths and not on the entire range – also all of the samples will have same impacts irrespective of their varieties and parentage; (b) the 1 nm spectral resolution of highly correlated bands may not vary much due to this subtle change in spectral response; and (c) the designed model may be robust enough to work on any type of spectra irrespective of their time of collection. Overall, the spectral data collected can be related to the parameters which are relatively stable during the day and right after leaf cutting. The real time visual inspection of spectra was assessed during collection to ensure its integrity and in case of any fault, was re-collected. The GPS location of each spectroscopy measurement was recorded with a differential Trimble dGPS as close as possible to the main stem of the banana plant to take a measurement (Fig. 1).

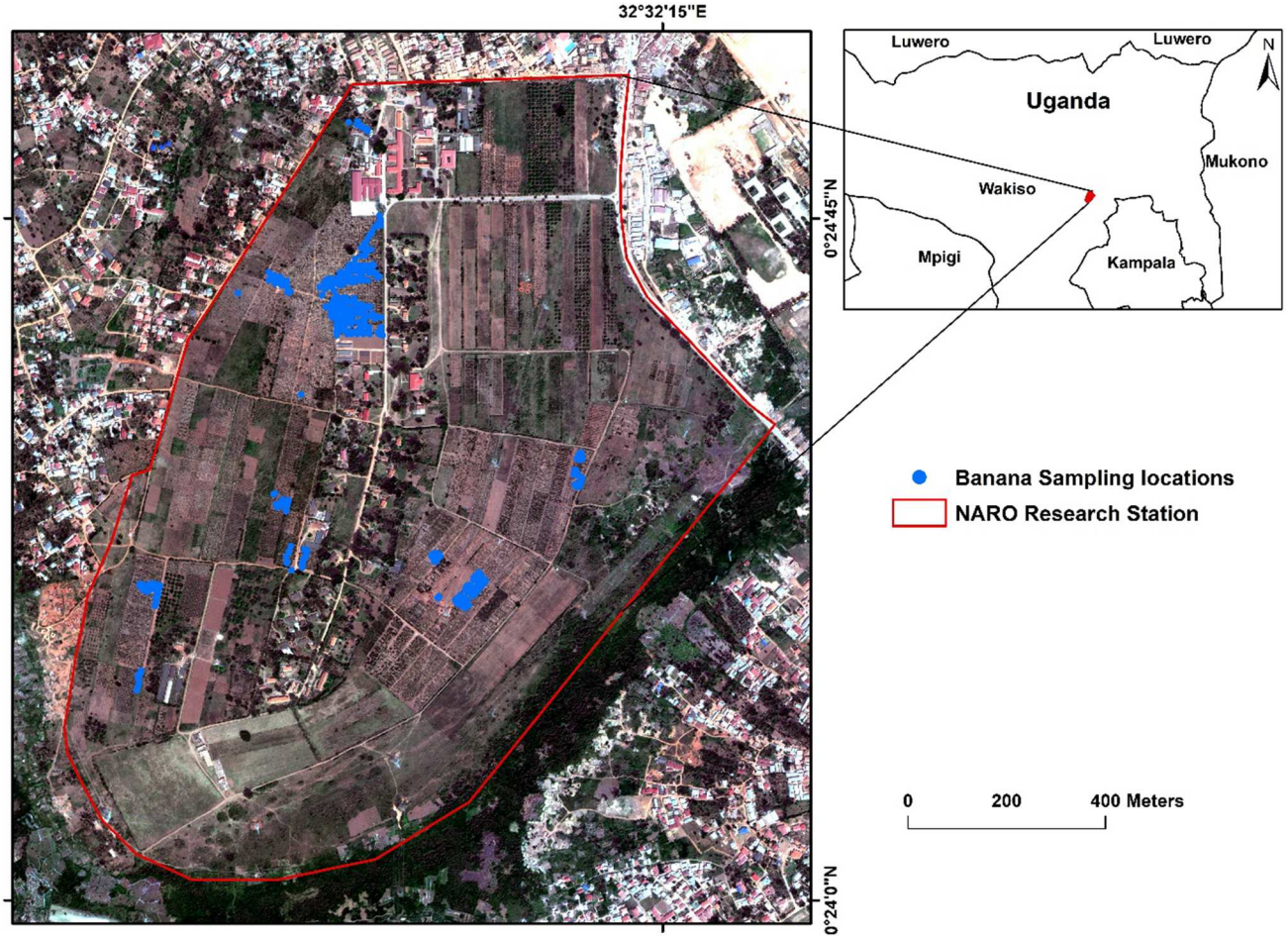

Fig. 1. Worldview 3 imagery (acquired 18 Feb 2018) of the NARO research station with the location of each individual banana plant spectrally measured.

For each leaf, five locations along the 'top surface' of each leaf blade were spectrally measured (Fig. 3). For all leaves, the spectral measures were performed along the adaxial side of leaf near the midrib in five separate locations. The multiple measurement points minimized the chance of sampling bias from non-representative variations in plant constituents from pest or disease, even though visible locations of these occurrences were avoided. Further, we were concerned about variability between many plants and not within a single leaf, thus random scanning on leaves to account for in-leaf variability was not found necessary here. Each sample was provided a unique barcode, so that the ASD spectral measures and DNA fingerprinting, could be matched. The ASD field computer was also used to enter other site information including plant age, number of tillers, tree density score, tree height, soil characteristics, plant health score, plantation management level, leaf number measured and parentage (usage).

3.3. DNA extraction and sequencing

As well as the physical identification and spectral measurements, the tissue samples were also collected for each variety. These samples were processed on site following strict protocol and then sent for DNA extraction and sequencing. Genomic DNA (gDNA) was isolated from banana leaf tissue using a modified CTAB method (Stewart Jr and Via, 1993). Approximately 50 mg of leaf dried using Silica gel was transferred into 2 mL Conical microtube containing a sterile stainless steel shot and milled into powder using a Mini-BeadBeater-8 (Biospec Products) for 30 sec at 30 oscillations per second (OPS). To the sample, 800 µL of pre-warmed (65 °C) CTAB extraction buffer was added and incubated at 65 °C for 30 min with occasional shaking. Chloroform extraction was performed by adding 800 µL of chloroform: isoamylalcohol (24:1, v:v), mixing thoroughly by vortexing, and then centrifugation of the mixture at 18,000*g* for 5 min in an Eppendorf 5415 D centrifuge. The aqueous supernatant was then transferred to a fresh 2 mL tube. Another chloroform extraction was performed on the supernatant and 2 µL of RNase A (1 mg/mL) was added to the resulting supernatant followed by incubation at 37 °C for 1 h. The resulting supernatant was transferred to a 1.5 mL Eppendorf tube. The gDNA was then precipitated by addition of an equal volume of isopropanol. The mixture was thoroughly mixed by inversion and then centrifuged at 18,000*g* for 10 min. The supernatant was then discarded and the pellet washed by adding 1 mL of ice-cold 80% (v:v) ethanol followed by centrifugation for 5 min at 18,000*g*. The supernatant was discarded and pellet dried under vacuum for 10 min at room temperature. The pellet was re-suspended overnight at 4 °C in 20 µL of nuclease free water. Purity and concentration of the gDNA was then determined by spectrophotometry.

From the 43 banana varieties initially sampled, the DNA tissue testing identified 12 varieties to be genetically different. These are also called 'Genotypes'. The two terms 'variety' and 'genotype' are used in context but essentially meant the same (i.e., banana variety). The in-field spectrometer readings for remaining verities were removed, and spectra for 12 genotypes were used in further analysis.

3.4. Pre-processing

The pre-processing of hyperspectral data involved the removal of outlier spectra and 'noisy' wavelengths from known regions of atmospheric water absorption (1350–1420 nm, 1770–1965 nm and

Fig. 2. The manual cutting of sample leaves from each plant (a), followed by the near immediate spectral measure with the ASD field spectrometer fitted with the leaf clip (b).

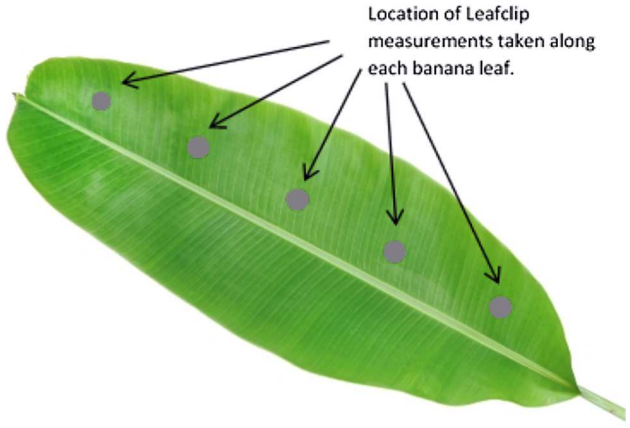

Fig. 3. Location of the 5 hyperspectral measures taken along each sampled leaf with an ASD field spectrometer fitted with a leaf clip.

2450–2500 nm) (Hennessy et al., 2020). Additional 'noisy' regions below 450 nm were also removed. The spectra for each genotype was plotted individually to allow the outliers to be identified. The Savitzky-Golay smoothing filter (parameters: order = 2, length = 11 and no derivative m = 0) was used to smooth any irregularities of the signal. The resultant 1231 spectra ($ASD_{genotype}$) corresponding to the 12 genotypes were further split into 620 and 611 samples representing the 'Old' and 'New' banana varieties, respectively ($ASD_{parentage}$). Each of the parentage class contained 6 genotypes. This was done to simplify the modelling and potentially increase the likelihood of achieving a successful model that offered useful results for plant breeders, who wanted to know the distribution and therefore adoption of new varieties. Table 1 shows the banana varieties, their assigned bin number and sample size of hyperspectral data used for modelling. Figs. 4 and 5 respectively show mean reflectance spectra of 12-binned varieties, and the mean reflectance of 'Old' and 'New' varieties within the 12 uniquely assigned bins.

Table 1
Banana varieties and their assigned bin number, parentage and sample size of hyperspectral data used for modelling. (Please see Appendix A for variety code details).

Variety Code	Parentage	Assigned Bin	No of sample spectra	Calibration samples	Validation samples
BLU	Old	Bin173	87	63	24
BOG	Old	Bin175	117	87	30
FHI17	New	Bin159	68	49	19
FH25	New	Bin167	58	43	15
GRO	Old	Bin225	87	58	29
GON	Old	Bin219	69	50	19
KAY	Old	Bin19	125	93	32
KM5	New	Bin161	119	85	34
M2	New	Bin191	127	92	35
NA31	New	Bin187	109	81	28
NAR7	New	Bin171	130	96	34
SUK	Old	Bin43	135	100	35
		Total	1231	897	334

A separate 'filtered' set of reflectance spectra was produced to match those wavelengths provided by the WV3 satellite (called $WV3_{ASD\text{-}genotype}$). The 16 bands include : Coastal Blue (400–450 nm), Blue (450–510 nm), green (510–580 nm), yellow (585–625 nm), red (630–690 nm), RedEdge (705–745 nm), NIR1 (770–895 nm), NIR2 (860–1040 nm), SWIR1 (1195–1225 nm), SWIR2 (1550–1590 nm), SWIR3 (1640–1680 nm), SWIR 4 (1710–1750 nm), SWIR5

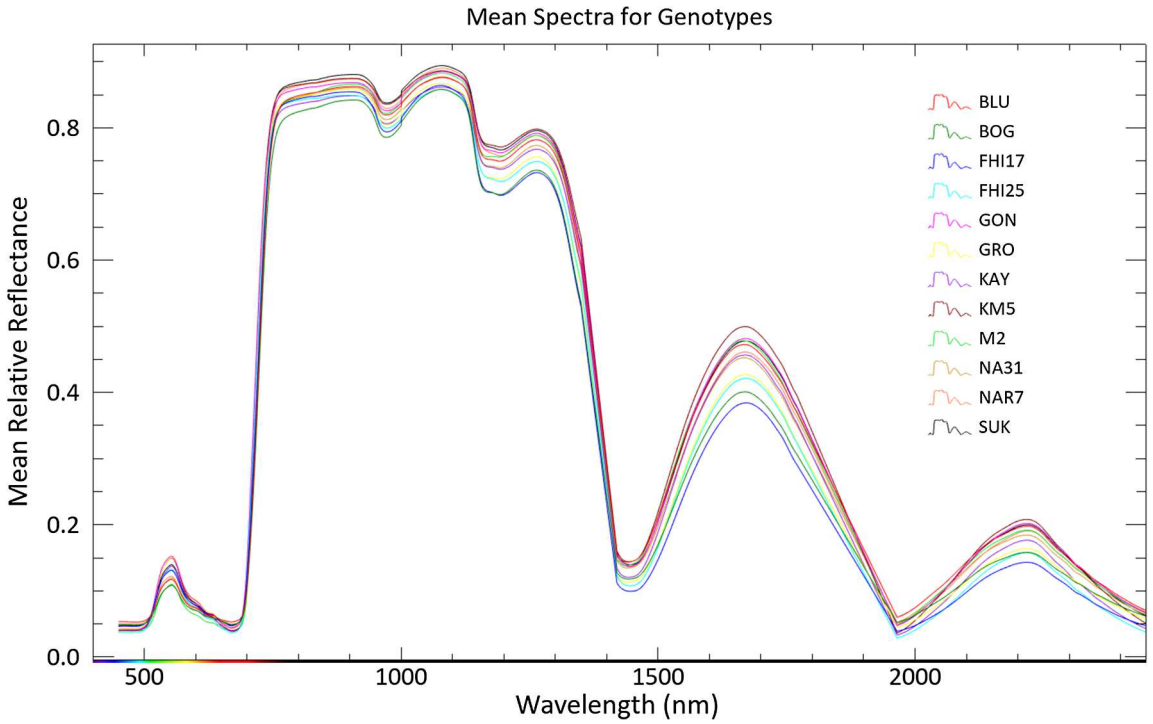

Fig. 4. Mean reflectance spectra of 12-binned banana varieties (ASDgenotype).

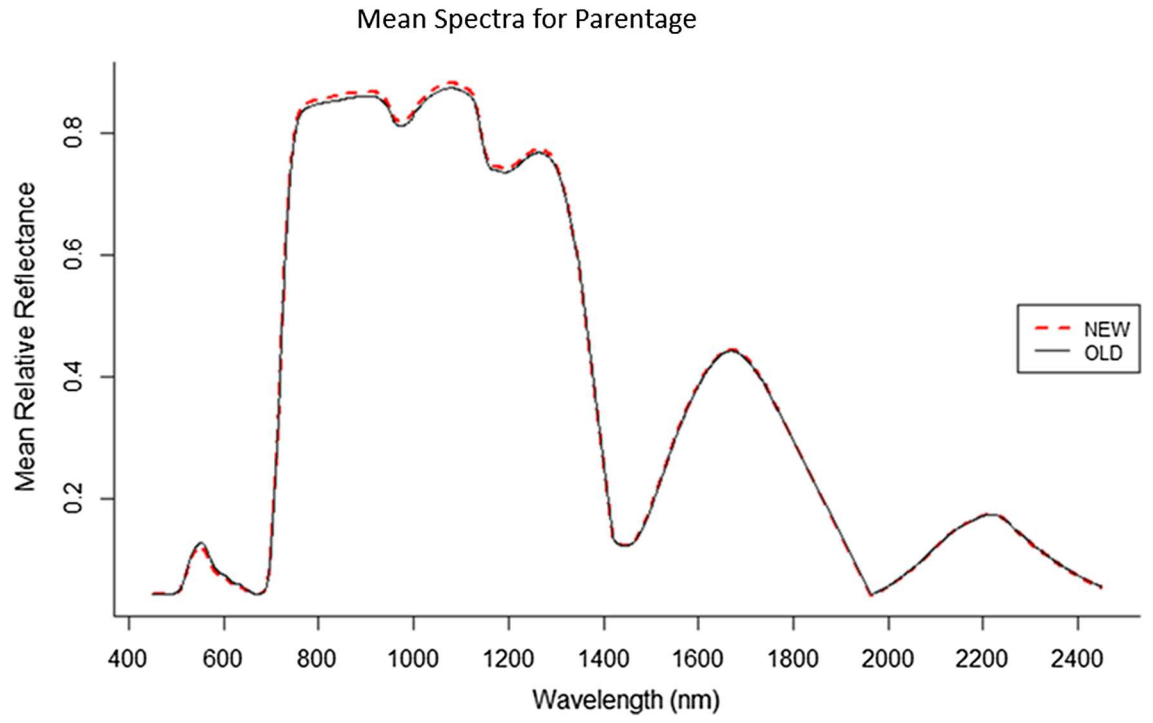

Fig. 5. Mean reflectance of 'Old' and 'New' banana varieties within the 12 uniquely assigned bins ($ASD_{parentage}$).

(2145–2185 nm), SWIR6 (2185–2225 nm), SWIR7 (2235–2285 nm), and SWIR 8 (2295–2365 nm) (Digital Globe, 2014). This was achieved by convolving the field ASD reflectance spectra to the spectral resolution of WV3 sensor using the built in resampling functions in the ENVI 5.5 software. The $WV3_{ASD\text{-}genotyepe}$ spectra was further grouped as banana parentage to get $WV3_{ASD\text{-}parentage}$ spectra for 'Old' and 'New' varieties. The WV3 satellite offers the highest spectral (16-bands) and spatial resolution (1.2 m Multispectral (MS) and 0.31 Pan Sharpened (PS)) available from a satellite platform with a temporal resolution of < 1 day. The platform is being more extensively used in research and commercial applications (e.g., Fletcher and Reddy, 2016; Robson et al., 2017). To map banana plantation in Queensland, Australia, Johansen et al. (2009), determined pixel size of ≤ 2.5 m to be sufficient for accurate identification of banana plantation row structure and object-separation from other crops. Thus, the use of WV3 data in this study was found appropriate for banana mapping at Uganda study sites. However, use of WV3 data does incur additional cost to the industry. Free to use satellite data, e.g., from Sentinel2, provide few spectral

resolutions matching with WV3, but with a spatial resolution of 10 m, do not have required high spatial resolution required for this application.

3.5. Remote sensing data

It was hypothesized that if equivalent $WV3_{ASD\text{-}genotype}$ spectra obtained from the ground based hyperspectral measurements could achieve strong varietal segregation, then there would be greater confidence in extrapolating those results to the data from a satellite platform. A 16-bands WV3 image was acquired for this study, encompassing the NARO research station in Kampala (acquired on 18th February 2018) (Supplementary Figure, SFig. 1). The timing of image acquisition coincided with the field sampling providing both radiometer and WV3 data with similar plant age and environmental conditions. The WV3 digital number (DN) was converted to surface reflectance using FLAASH algorithm in ENVI followed by the Dark object Subtraction (DOS) to minimize atmospheric effects. However, the imagery suffered from some haze, which partially degraded the image quality and posed difficulty in identifying individual banana crowns at 1.2 m resolution. The WV3 multispectral imagery (MS) (8-bands) was Pansharpened (PS) to generate PS image of 0.30 m spatial resolution, which was used to delineate banana plant boundaries using object-based approach suggested by Johansen et al. (2014) in eCognition software. The SWIR 8-bands were resampled to 1.2 m to match with the spatial resolution of MS bands, and then staked together to make 16-bands imagery for further processing. The banana boundary layer was superimposed on stacked WV3 data to extract the reflectance for 12-gneotypes ($WV3_{reflectance}$), and used for banana classification. SFig. 2 shows the comparison of ASD and WV3 reflectance for few genotypes.

3.6. Derivation and evaluating of classification models

3.6.1. Canonical powered partial least squares (CPPLS) classification algorithm

For analysis of full range reflectance spectra ($ASD_{genotpes}$ and $ASD_{parentage}$), the supervised canonical powered partial least squares (CPPLS) classification algorithm was selected due to its feature extraction and data inference abilities and that the analysis required was qualitative i.e. Y variable was banana genotype. The model integrates canonical correlation analysis and the parameterization of loading weights optimized over a given interval, and has ability to extract predictive information for the latent variables more effectively than ordinary PLS approaches (Indahl et al., 2009; Mevik and Wehrens, 2007). The CPPLS algorithm fits PLSR model, where relation between predictor matrix X (wavelenths) and response vector Y (genotypes) are found thorough the latent variables (components) iteratively (Indahl et al., 2009; Mahmood et al., 2011). The model compresses numerous collinear variables into a few orthogonal components (PCs) which explain variance–covariance structures and optimize the explained power of the response variables (Wold et al., 2001). For each latent variable, the regression coefficient is estimated by a cross validation approach, and the ONC is determined by minimizing RMSE between predicted and observed response variable (Esbensen et al., 2002).

The CPPLS classification algorithm provided in the statistical package R was used to analyse the reflectance spectra (Mevik and Wehrens, 2007). R-Packages used included: 'hyperspec' for generating hyperspectral objects and exploratory plotting for visual/manual outlier removal, 'cppls.fit' for CPPLS and 'caret' for data segregation (Mevik and Wehrens, 2007). The $ASD_{genotpes}$ and $ASD_{parentage}$ spectra were used as input in CPPLS. The model was developed from pre-processed spectra and not on derivative spectra. The derivative spectra are commonly used to remove noise and should be done with caution due to a high correlation between discrimination capabilities and bandwidth (Schmidt and Skidmore, 2004). Studies have shown that PLSR models that utilizes the full spectrum can predict photosynthetic capacity through time (Barnes et al., 2017), while noise removal demonstrated negative effects on the subsequent statistical analysis of spectral characteristics (Vaiphasa, 2006). Since, it was hypothesized that each banana variety would provide a unique spectral signature; it was unknown how plant age, location, management practice, etc. would influence this. Thus, analysis was done on pre-processed spectra to capture a subtle change in banana variety spectral reflectance. This analysis determined whether the hyperspectral reflectance data could statistically differentiate each banana genotype (including old and new) as well as indicate how accurately the derived spectral models could predict the genotypes and parentage of samples not included in the calibration set. To achieve this, the CPPLS model was built by randomly splitting the response variable into two sets: ~75% of dataset to calibrate (train) the model and remaining ~25% to validate the model's prediction accuracy. The CPPLS was initiated with a large number of PCs (40) to inspect the model fit by plotting, extracting and summarising model components (Mevik and Wehrens, 2007). The model was calibrated by k-fold cross-validation (k = 10) to evaluate the root mean square error of prediction (RMSEP) as a function of the number of components from 1 until ONC. The RMSE between the actual and predicted values calculated over all cross-validation calibrations. The best calibration equation and the number of latent variables were selected based on a low RMSEP calculated as shown in Eq. (1) (Esbensen et al., 2002).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}} \quad (1)$$

The model was then validated on the test set and its quality was evaluated with the standard error of prediction corrected of the bias (SEPc) calculated as shown in Eq. (2) (Esbensen et al., 2002).

$$SEPc = \sqrt{\frac{\sum (\hat{y_i} - y_i - Bias)^2}{n-1}} \quad (2)$$

where n is the number of sampling of test set, y_i the actual value of the sampling i and $\hat{y_i}$the predicted value for the sampling i. The bias is the mean value of the difference between actual and predicted values. Following completion of the calibration, the model was validated using prediction (validation) dataset to determine the accuracy of prediction.

The overall classification performance was also expressed in terms of sensitivity and specificity. These are statistical measures of the model performance of a binary classification according to the confusion matrix which classifies selection decisions as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (Ballabio and Consonni, 2013). 'Sensitivity' measures the proportion of actual positives that are correctly identified as positive (i.e., percentage of each genotype which are correctly identified as belonging to true genotype), while 'Specificity' measures the proportion of negatives that are correctly identified (i.e., percentage of each genotype correctly identified as not belonging to other type). These are computed as:

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

Here, TP and TN are the number of instances where a class of interest are correctly classified as well as correctly classified as not being observed, respectively. FN is the number of instances where a class is visually observed but is incorrectly classified as some other class, while FP is the number of instances where a class is incorrectly classified as the class of interest.

3.6.2. Random forest (RF) based classification

The RF classification was performed on $WV3_{ASD\text{-}genotype}$ and $WV3_{ASD\text{-}parentage}$ spectra (ASD spectra resampled to 16-bands matching those of the WV3 satellite bands). The $WV3_{ASD\text{-}genotype}$ spectra was split

into 3:1 ratio, where 75% of spectra used to train and build RF model (calibration), which also provided a general prediction accuracy for each behavior. The remaining 25% was used to validate the model's prediction accuracy. The RF is an ensemble classifier that generates multiple decision trees from a randomly selected subset of training samples and variables (Breiman, 2001). In recent years, the classifier has gained popularity due to the accuracy of its classifications. The CPPLS algorithm was not used in this case because it is more suitable in reducing the large number of measured collinear spectral variables to a few non-correlated latent variables. The variable importance ranking (VIR) in RF was performed to systematically assess the usefulness and identification of most important WV3 bands for discriminating banana genotype. The R libraries 'randomForest' (Liaw and Wiener, 2002) and 'Caret' (Classification and Regression Training) were used for RF modelling. The VIR of WV3 bands was determined based on 'Gini index', which is used to measure the error across the RF ensemble of trees (Breiman, 2001). The process was repeated for $WV3_{ASD\text{-}parentage}$ spectra.

Initially, a RF model was developed with default parameters for '*mtry*' = 7 (the number of variables tried at each split, which is approximately equal to the square root of the number of variables for classification) and *'ntree' = 500* (the number of trees to grow) (Belgiu and Drăguţ, 2016). The optimization of two parameters was done through a '*random search*' strategy within a given range of '*mtry*' and *'ntree'*. The 10-fold cross-validation with 3 repeats, were used to limit and reduce overfitting on the training set (Schratz et al., 2019). A graph of model accuracy with different parameters are shown in the SFig. 3. The optimal *'mtry'* = 6 and *'ntree'* = 2000 were used for genotype classification; and 4 and 1500, respectively, for parentage classification. A confusion matrix was computed, from which the accuracy was calculated using the following equations:

$$Accuracy = (TP + TN) / (TP+FP + TN+FN) \quad (5)$$

The RF model was also built on $WV3_{reflectance}$ that were extracted for 12 genotypes (Section 2.5). The prediction accuracy for each genotype and VIR of WV3 bands were compared with those from $WV3_{ASD\text{-}genotype}$ RF modeling outputs to determine if similar results can be obtained from canopy-level reflectance from the satellite platforms.

4. Results

4.1. CPPLS model calibration

4.1.1. Genotype

The explanatory analysis results from the PCA applied to full range hyperspectral dataset ($ASD_{genotype}$) produced a six PC model that explained > 99% of data variance. However, the score plot of first two PCs did not show strong clustering for all genotypes (SFig. 4). The initial CPPLS model fit with large number of 40 PCs explained a maximum of 72.7% of data variance, out of which first 12 PCs explained 69% of data variance, while the remaining PCs contributed only ~4% mostly as noise (SFig. 5). Fig. 6 shows the RMSEP plots as a function of the number of PCs, which demonstrated increasing prediction accuracy with the model complexity as the RMSEP values decreased with increasing number of PCs.

Utilizing virtually all of the variance observed in the dataset indicated over parameterization of the model and the subsequent reduced performance with new dataset. 12 PCs was determined as optimal as the RMSEP values did not greatly improve with increasing number of PCs (Fig. 6). The RMSEP plots for each genotype suggested that the prediction accuracy is variable and required a complex model. With 12 PCs an overall accuracy of 78% was achieved.

Fig. 7 shows the random split of $ASD_{genotyepe}$ spectra for each genotype; 75% for calibration and development of CPPLS model from 12 PCs, and 25% for validation of model to determine model's performance and accuracy. Both calibration and validation datasets showed a high 'goodness of fit' for all genotypes indicating better prediction from the two datasets. The proportions of genotypes sampled were identical between training and testing datasets as all of the samples were within the 1:1 line (SFig. 6). The 'pairwise plot' of score values for first 10 PCs shown in Fig. 8 explained the pattern, groupings and outliers in the 12-binned spectra ($ASD_{genotyepe}$) and the experimental design of data. The PC1 (7.44%), PC2 (9.59%), PC4 (22.14%), PC5 (19.65%) and PC9 (2.19%) explained relatively higher amount of X-variance (wavelengths) in genotype predictions.

The loadings plots (regression coefficients) of the PC4 and PC5 (highest data variance) show the contribution of specific wavelength regions in genotype separation (Fig. 9). The plots indicated the model's complexity as spectral peaks varied between the components and algebraic expressions (positive and negative loadings). Thus, determination of a specific wavelength contribution based on any one component was difficult. The 12 PCs regression coefficient was difficult to interpret and hence plots of PC4 and PC5 were used to find the position of spectral peaks or profile pattern. The model considered all of these variabilities to deal with the complexities of spectrally similar banana varieties.

4.1.2. Parentage

The parentage of the 12-binned banana varieties is given in Table 1. The analysis $ASD_{parentage}$ initially with 40 PCs CPPLS model fit for parentage classification indicated the first 7 PCs explained ~99% of data variance. The RMSEP plots for 'New' and 'Old' varieties were found similar (SFig. 7), and together required a complex model to attain a maximum accuracy of 0.81 with 40 PCs (Fig. 10). However, as not much data variance was explained by higher PCs, the gradual decrease in RMSEP indicating change of overfitting of model (Fig. 10a). Thus, 7 PCs were found optimal and used to build the CPPLS model for parentage prediction.

The 'pairwise plot' of score values for first 7 PCs is shown in Fig. 10b, indicating PC1, PC2 & PC3 contributing most of the data variances (> 82%). The predictions plots of 'New' and 'Old' varieties (SFig. 8) show significant overlap between the two samples, showing spectral similarities between the two. The plot of regression coefficients in Fig. 11 indicated the model's complexity where the spectral peaks varied between the components (positions and algebraic expressions). The spectral bands that produced the highest significance in the spectral models where Green, Red, RE, and few NIR and SWIR bands.

4.2. Model validation and prediction accuracy

The analysis of model performance on validation data demonstrated that the prediction accuracy increased with the model complexity for 12-binned varieties and also for the parentage (i.e., with the number of PCs) (Fig. 12a and b). With the optimal number of PCs, producing average prediction accuracies of 60% and 64%, respectively, for genotype and parentage classifications. For 12-binned genotypes, BLU and GON produced 80% of prediction accuracy, followed by NAR7 and KAY with ~70% of accuracy. For parentage, the accuracy of 'New' variety was slightly higher than the 'Old' variety.

The decision matrix in the Table 2 represents the classification accuracy for each genotype expressed as user accuracy (UA), producer accuracy (PA) and overall accuracy (OA). The OA of 64.2% suggests an above average performance of the model in general. The higher PA's (> 76%) and UA's (> 64%) for BLU, GON, NAR7 and GRO suggesting their correct identifications in both classifications and also in the field, in most occasions. KAY has shown marginally low PA (71.0%) and UA (57.9%) due to misclassification of samples from other types. The remaining genotypes showed spectral intermixing, particularly between FH17, KM5, FH25, and BOG, resulting in poor classification accuracies.

These findings were further assessed through the sensitivity and specificity analysis for the prediction of genotypes from the validation datasets (Table 3). In general, the higher specificity (≥93%) indicated

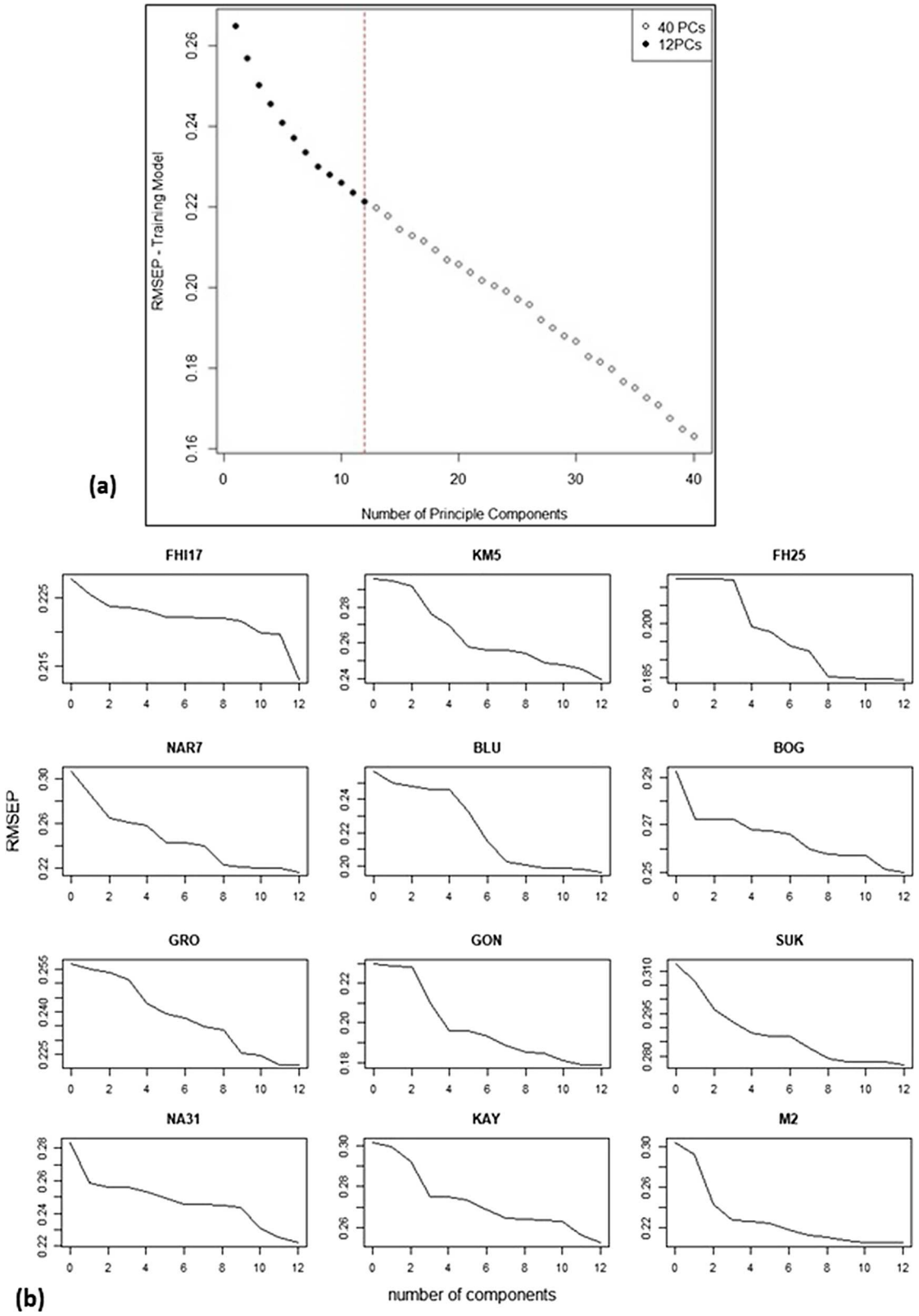

Fig. 6. Optimization of number of PCs against RMSEP as a measure of model prediction performance for genotype (a). The RMSEP plots for each genotype showing variable prediction accuracy with 12 PCs (b).

that the model performed very well in differentiating one genotype from the other. However, the relatively lower sensitivity for few genotypes (e.g., FHI17, KM5, FH25) indicated that the model did not successfully predict these as true variety (i.e., reliably ruling out of a genotype classified as a particular variety). Higher sensitivity values for BLU, GON, NAR7 and GRO (> 75%), and to some extent for KAY (71%) showed greater reliability of these genotypes when classified. These parameters suggested that the current model may not be very accurate in all banana genotype identification, except for few, but can be used to reject genotypes that do not belong to a true class. This is important for

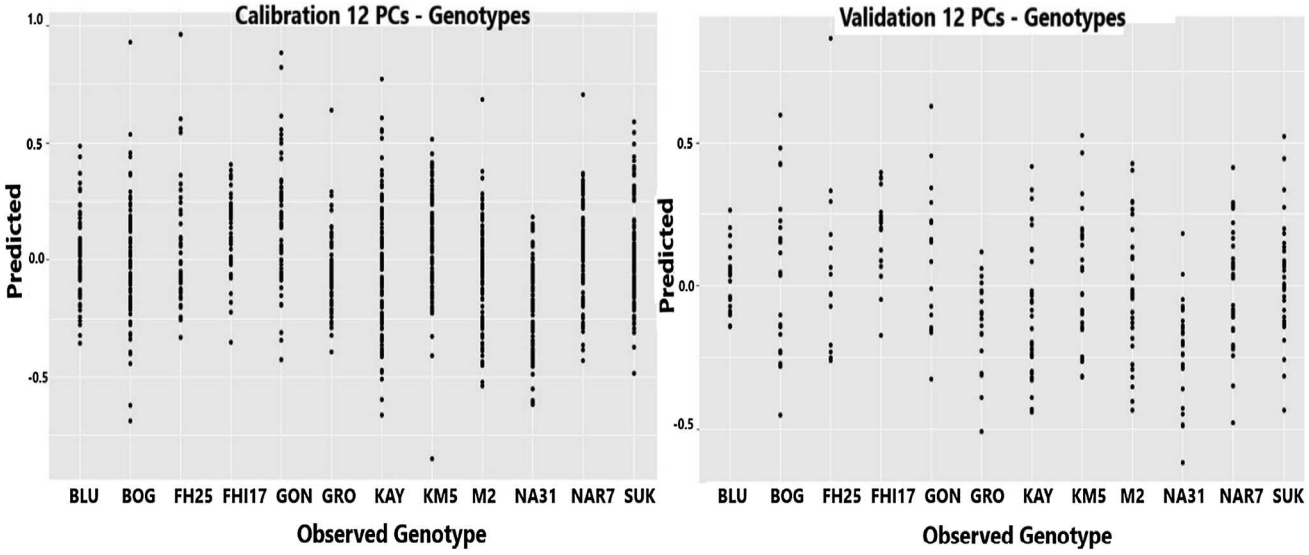

Fig. 7. Calibration and validation datasets used in CPPLS model showing high goodness of fit for all genotypes.

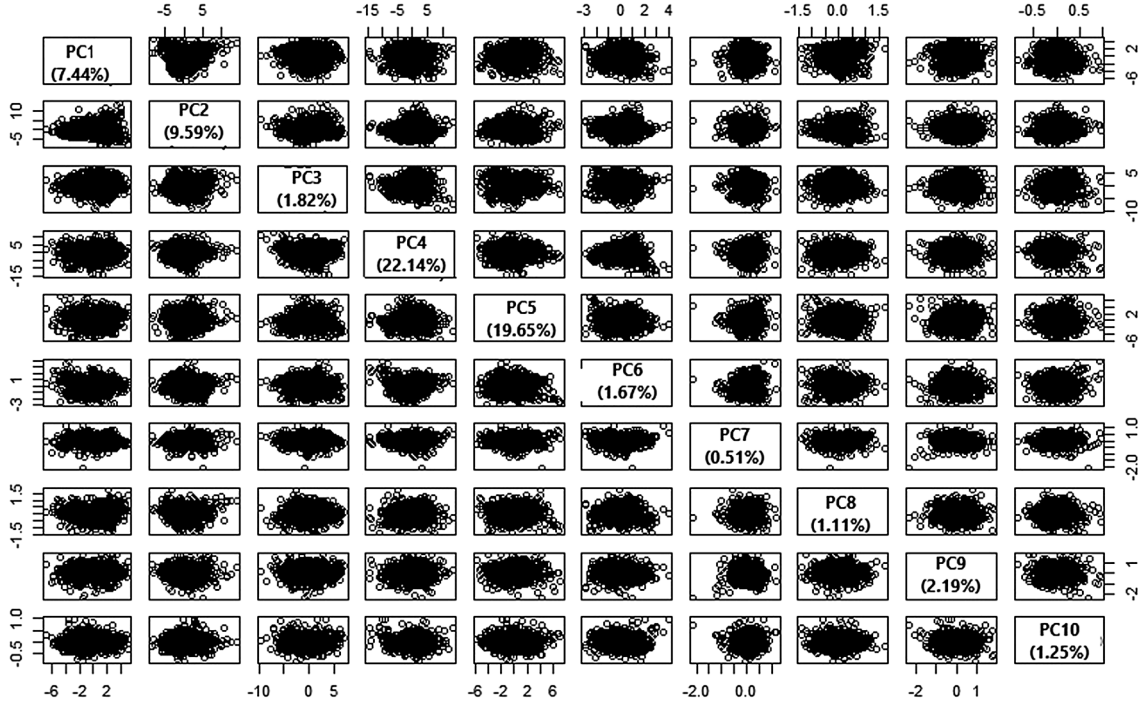

Fig. 8. Score plot for first 10 PCs of CPPLS model showing pattern and grouping of data in genotype prediction.

this study as a positive result with high specificity (high probability of true classification) indicates a correct classification, rather than a misclassification (low type I error rate) (Ballabio and Consonni, 2013). Thus, the higher accuracy (> 89%) is achieved mainly because of the true negative predictions (NPV = TN/ (TN + FN)) for all genotypes. With the current datasets the classifications of genotypes such as BLU, GON, NAR7, GRO and KAY (≥91% accuracy) are encouraging, although further validation over more samples is recommended.

4.3. Extrapolating the hyperspectral results to the WV3 satellite data bandwidths

To determine if the accuracies for varietal segregation achieved from the full hyperspectral data set could be extended to the WV3 satellite wavelengths, a sub-set of spectral bands matching those of WV3 wavelengths ($WV_{ASD\text{-}genotypes}$ and $WV3_{ASD\text{-}parentage}$) were undertaken for further analysis. The RF algorithm was used to determine the classification accuracies for genotypes and parentage.

4.3.1. Genotype

Table 4 shows the RF classification accuracies for 12-binned genotypes. The mean of model and prediction accuracies were 44.8% and 52.8%, respectively. Genotypes such as BLU and BOG, produced higher predictions accuracies (≥70%), followed by GON, GRO and KM5 (> 61%). The prediction accuracies for remaining genotypes were relatively low, with NA31 being the lowest (32.2%). From both full range hyperspectral data in CPPLS and RF classifications of spectra of WV3 wavelengths, the genotypes such as BLU, GON and GRO produced higher prediction accuracies. The results were found in consistent with other similar studies (e.g., Fletcher and Reddy, 2016; Fletcher, 2016).

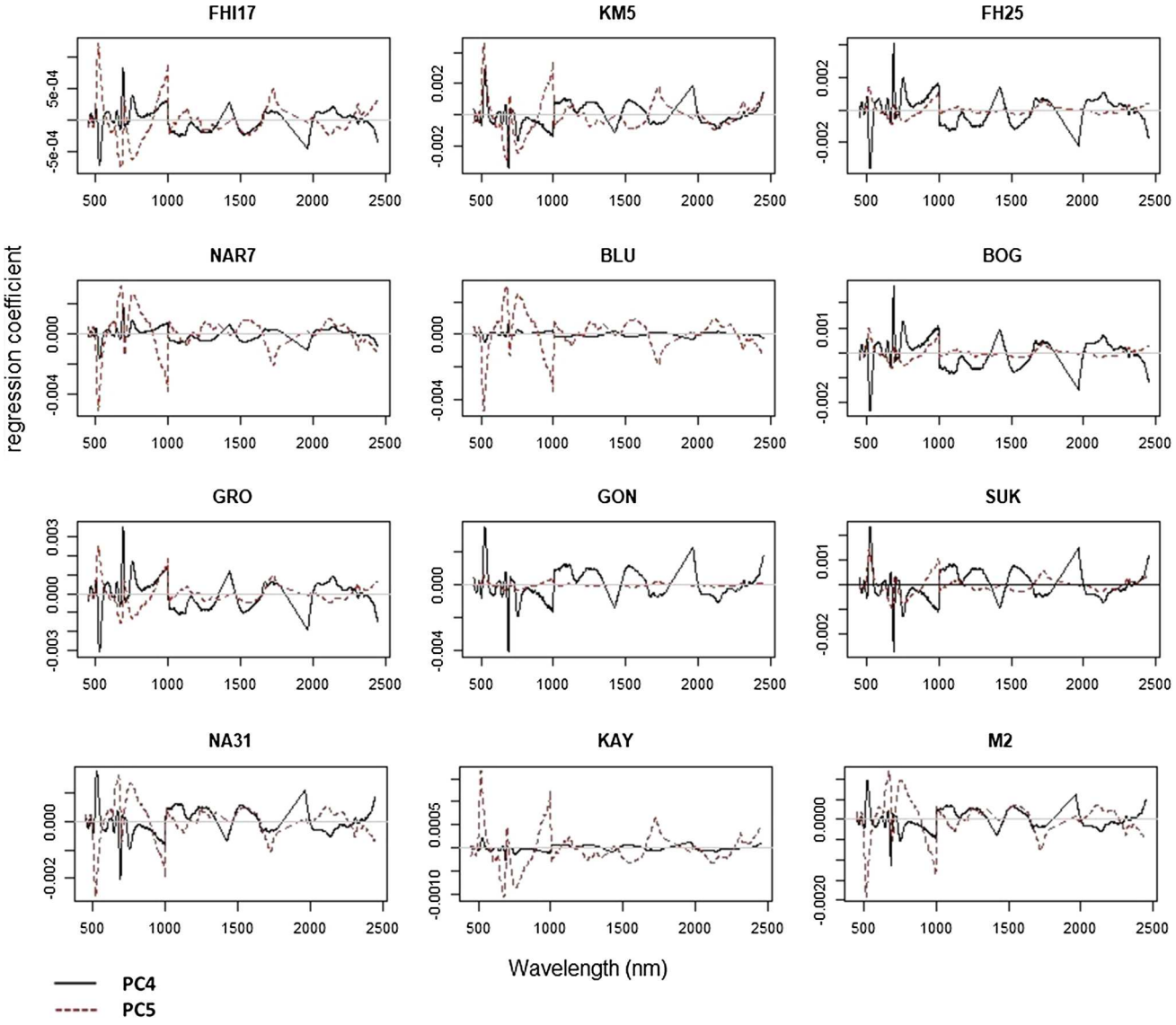

Fig. 9. Regression coefficients of PC4 and PC5 to explain the contribution of wavelengths in genotype separation.

However, the accuracies in this study were relatively low due to use of spectra of banana genotypes with nearly similar characteristics; whilst others classified two different types of vegetation.

The classification of the remaining genotypes NAR7 and KAY produced higher sensitivities (> 0.71) (i.e., measure to predict true positive) in the CPPLS; but their RF prediction accuracies were relatively poor. This result indicates that these genotypes may be differentiated by using only the wavelengths that corresponded with WV3. Whilst, BOG and KM5 had higher accuracies in RF classifications as compared to CPPLS modelling, thereby indicating the usefulness WV3 wavelengths in their spectral separation. Overall, the results from two models suggested BLU, GON and GRO, and to some extent, BOG and KM5 genotypes are spectrally distinguishable from both on-ground hyperspectral measurements and from the 16 spectral wavelengths that correspond to the WV3 satellite. These varieties pose as strong candidates for classifications from WV3 satellite data. This is a major outcome of this research. Further research is required to determine if similar outcomes can be achieved across other regions in Uganda.

Fig. 13a illustrates the variable importance rankings (VIR) of WV3 spectral bands (y-axis) expressed as Gini Index (x-axis) in genotypes classification. The ranking is in top-to-bottom as most- to least-important. The RedEdge band was found most important to the model. The spectral bands Blue, Green, SWIR8, and SWIR1 were found of equal importance, followed by SWIR1 and NIR2, as these bands could further decrease in Gini index to 60. The SWIR5 band was found to be of the lowest importance. The top ranked variables resulted in nodes with higher purity with a higher decrease in Gini coefficient, with an overall difference of ~40 found between top and bottom ranked variables.

4.3.2. Parentage

Table 5 shows the confusion matrix for RF classification accuracies for banana parentage from the resampled spectra to WV3 bands. The overall prediction accuracies achieved was > 68%. The majority of samples were correctly classified as 'New' (TP = 113) and as an 'Old' (TN = 116), FN = 52 indicates the number of instances 'New' was visually observed but was incorrectly classified as 'Old', while FP = 53 indicates the number of instances when 'Old' was incorrectly classified as 'New'. The terms *'Condition Positive'* implies the total number of reference samples for 'New' (TP + FN), and *'Condition Negative'* for 'Old' variety (TN + FP). Thus, the *'Predictive Condition Positive'* of 'New' is equal to TP and FP respectively, when the number of samples in 'New' is correctly classified as 'New', and number of samples in 'Old' are misclassified as 'New'. Similar interpretation can be made for *''Predictive Condition Negative'*.

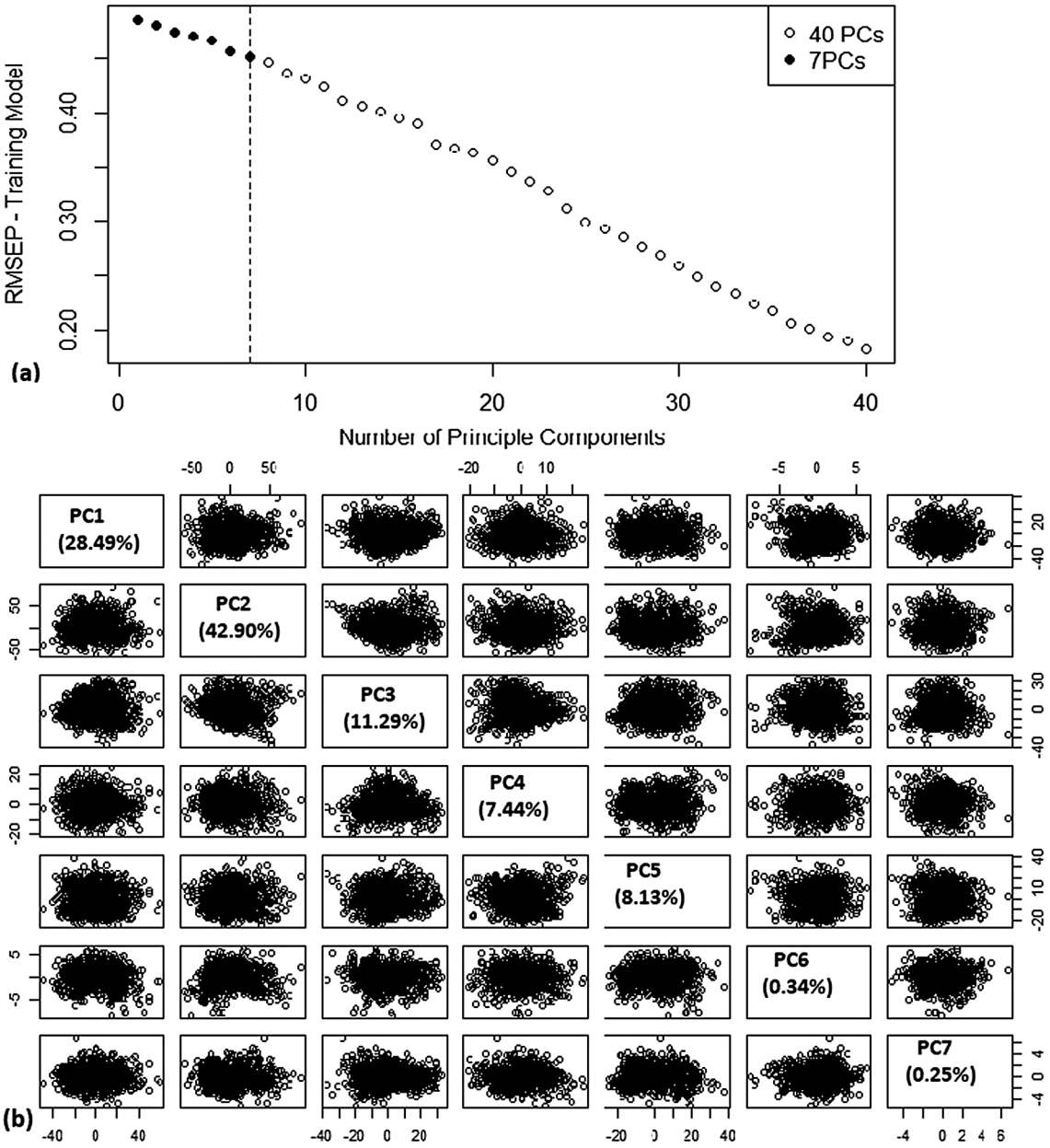

Fig. 10. Optimization of number of PCs against RMSEP as a measure of prediction performance of CPPLS model for parentage (a). Score plots for first 7 PCs showing pattern and grouping of data in parentage prediction (b).

Fig. 13b shows the VIR of WV3 bands for RF classification for banana parentage. The Green, RedEdge and NIR2 bands have shown higher contributions to the homogeneity of the nodes and leaves in the resulting RF. Overall, all of visible and NIR bands were considered important as compared to SWIR bands. This result was found different from genotype classifications, where wavelengths in all part of the spectrum, were found crucial in RF classification.

4.4. RF classification of WV3 surface reflectance for genotype

The mean ASD hyperspectral and WV3 satellite reflectance extracted for BLU, BOG, GON, GRO and KAY genotypes are shown in SFig. 2. These varieties have shown higher predictions from the RF and CPPLS models and hence were used for the comparison. The major differences between the two reflectance measures were observed in the visible-NIR regions (Bands 2–8) and also in three SWIR regions (SWIR1, 3–4). The higher WV3 reflectance in the visible bands could be attributed to high back-scattering of electromagnetic radiation in shorter wavelengths from atmospheric haze (Kaufman, 1993). Table 6 shows the RF prediction accuracy for each genotypes, and Fig. 14 shows the most important WV3 bands (satellite) for discriminating banana genotype in this case.

5. Discussion

This novel study explored the use of hyperspectral reflectance measurements of banana leaves as a means of differentiating variety and parental origin. The study also evaluated the accuracies of extrapolating hyperspectral ground based reflectance measurements of banana plants to the Worldview-3.satellite imagery. Whilst some previous research has extrapolated leaf-level hyperspectral measurements of different vegetation types to satellite spectral bands (Fletcher and Reddy, 2016; Fletcher, 2016) these studies compared the spectral behaviour of two different vegetation types (soybean with pigweed), and

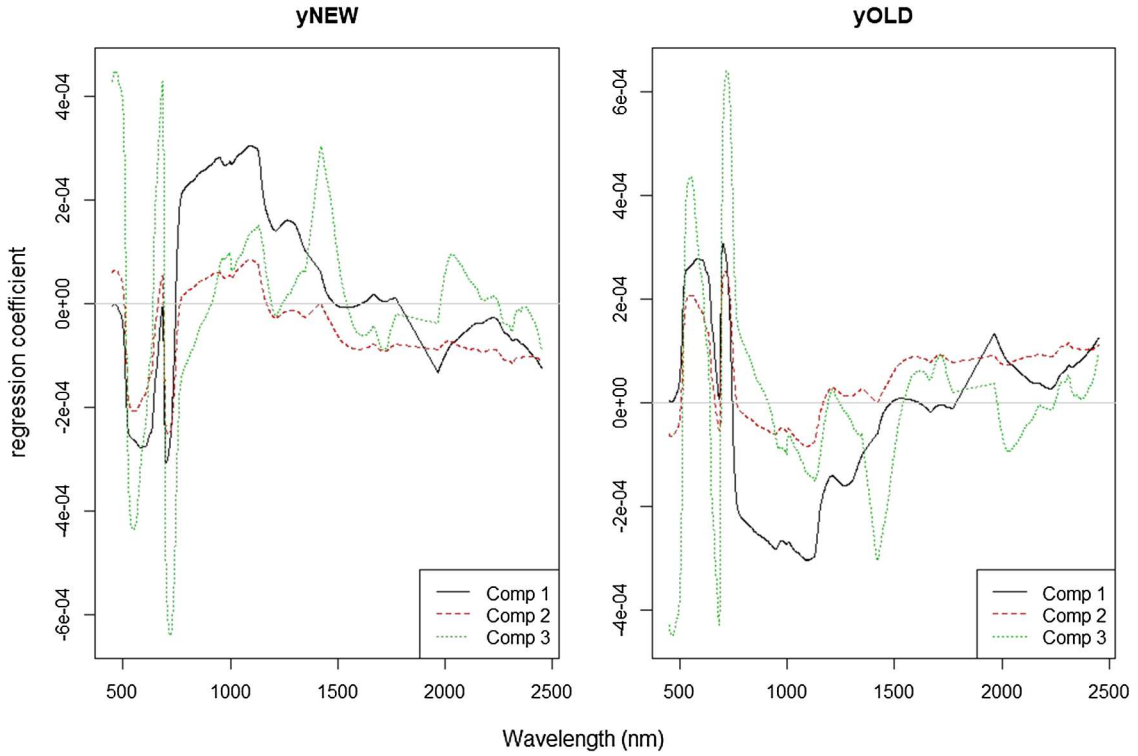

Fig. 11. Regression coefficient showing contribution of a specific wavelength for 7 PCs CPPLS model.

not between the genotypes and did not validate the accuracies from the actual satellite based measurements. The research presented in this study addresses some of these research gaps.

The CPPLS model was found appropriate for the analysis of the hyperspectral reflectance measures due to its feature extraction and data inference abilities, and suitabilty for analysing categorical data (banana genotypes) (e.g., Øvergaard et al., 2013a,b; Suarez et al., 2017). Using leaf-level hyperspectral reflectance data, the CPPLS algorithm successfully discriminated banana genotypes BLU and GON (prediction accuracy of 80%), NAR7 and KAY (70%), and for GRO and SUK genotypes (> 60%). Additionally the CPPLS modelling successfully categorised 'New' and 'Old' varieties (prediction accuracies of 65% and 63%, respectively). These results were consistent with other similar studies to predict crop biophysical properties and yield from hyperspectral reflectance (Meacham-Hensold et al., 2019; Øvergaard et al., 2013a,b). The CPPLS analysis of a high dimensional ASD hyperspectral data analysis offers dimension reduction, model validation, and tuning of model complexity as described by Lee et al. (2018). The initiation of the model with a large number of PCs (40), and the subsequent evaluation of the model fit (RMSEP, loading etc.) identified the ONC required for the discrimination of the banana genotypes and parentage, being assessed. The determination of ONC is usually done by

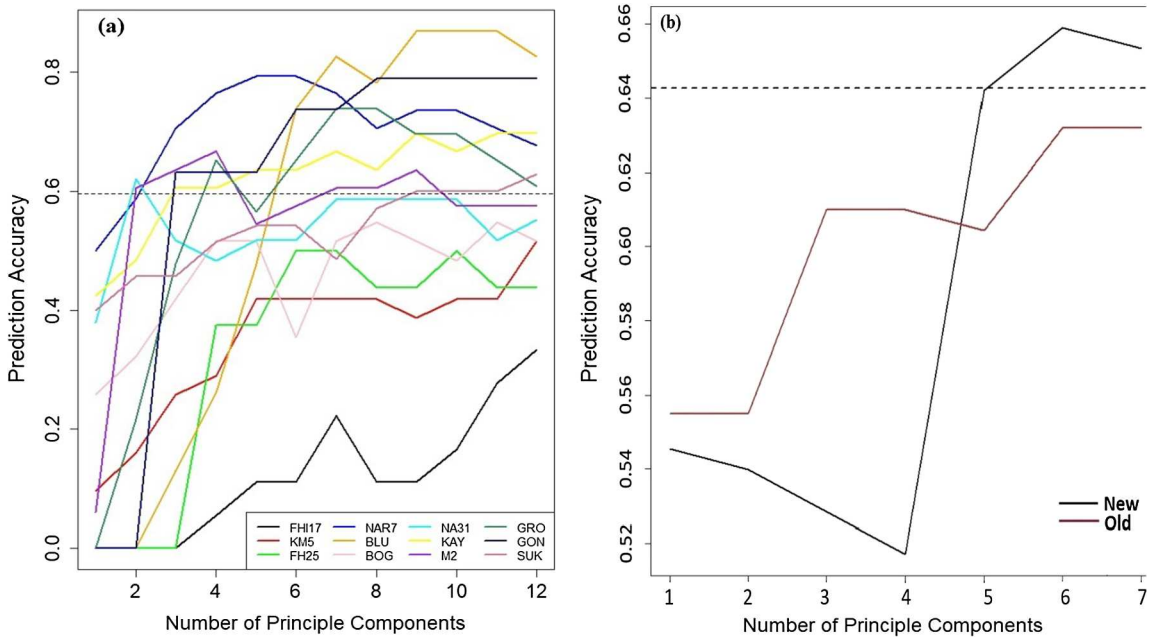

Fig. 12. Overall modelling accuracy with optimal number of principle components for 12 Genotypes (a) and Parentage (i.e., new and old varieties) (b).

Table 2
CPPLS based banana variety classification accuracy (please see Appendix A for variety code details).

Classification	Reference												
	Genotype	FHI17	KM5	FH25	NAR7	BLU	BOG	NA31	KAY	M2	GRO	GON	SUK	UA (%)
	FHI17	3	1	2	2	0	0	0	0	5	1	0	0	21.4
	KM5	2	13	0	0	1	1	1	0	1	1	0	3	56.5
	FH25	1	1	8	0	0	0	0	0	0	2	0	0	66.7
	NAR7	2	4	1	25	0	0	0	0	0	0	0	7	64.1
	BLU	0	0	0	0	20	6	3	0	0	0	0	0	69.0
	BOG	0	2	0	0	0	15	1	2	0	0	0	0	75.0
	NA31	0	1	0	0	0	3	17	0	0	0	0	0	81.0
	KAY	1	2	0	3	0	3	4	22	1	0	1	1	57.9
	M2	0	0	0	1	0	0	0	1	19	0	0	1	86.4
	GRO	3	0	1	0	0	1	0	1	1	16	0	0	69.6
	GON	0	1	0	0	0	0	0	5	0	0	15	0	71.4
	SUK	5	4	2	1	0	0	1	0	4	1	1	21	52.5
	PA (%)	17.6	44.8	57.1	78.1	95.2	51.7	63.0	71.0	61.3	76.2	88.2	63.6	
	OA (%)	64.2												

taking the first local minimum in the RMSEP curve (Mevik and Wehrens, 2007). Although subjective, the process was found crucial in this study as the prediction accuracy for each genotype was variable (Fig. 6). Wold et al. (2001) discussed the importance of upper and lower bounds (RMSE) with several Y variables, and the PLSR's ability to model and analyse several Y's together. Thus 12 PCs model fit with RMSEP of 0.24 addressed the issue of any over parametrization. The 40 PCs model indicated the likely over parameterization by utilising virtually all of the variance observed in the data set. It is therefore likely that it would have resulted in poor model performance when applied to novel datasets. The variable selection techniques (e.g., regression coefficient, Fig. 9) identified particular spectral regions (wavelengths) more sensitive to varietal segregation. The method was found consistent with other similar studies (e.g., Mahmood et al., 2011; Peerbhay et al., 2013; Schratz et al., 2019). Assuming that this model/technique is deployed for the discrimination of banana varietal population across other regions in Uganda, the results from 12 PCs for genotypes and 7 PCs for parentage seemed justified. This is particularly important considering high spectral similarities between banana genotypes. With the ONC, the model considered all of these variabilities to deal with the complexities of spectrally similar banana varieties.

The random forest (RF) algorithm applied to the resampled hyperspectral data also showed high potential for discriminating some of the genotypes and their parentage. The optimization of RF parameters '*mtry*' and '*ntree*' was considered important to reach a robust assessment of the model's predictive power (e.g., Belgiu and Drăguţ, 2016; Schratz et al., 2019). The 'random search strategy' and '10-fold cross validation' was found effective and supported findings from Schratz et al. (2019). The variable importance ranking (VIR) helped determine those WV3 bands that were crucial in the RF classifications of both banana genotypes and their parentage. Similar studies conducted on species discrimination from hyperspectral data using RF also highlighted the usefulness of VIR in identifying crucial bands for species classification (Fletcher and Reddy; Fletcher, 2016; Peerbhay et al., 2013). The RF model built on the WV3 satellite reflectance also identified BLU, BOG, GON, GRO genotypes with higher prediction accuracies. The performance of RF model slightly improved in this case (56%) as compared to ASD reflectance resampled to WV3 wavelengths (~53%) (Table 4). All of the SWIR bands were found to be of higher importance, followed by the NIR and then visible bands (Fig. 14). Because of a thin haze layer

Table 4
ASD based RF model and prediction accuracy (%) for each genotype. The overall percentage accuracy is average of each genotype accuracy.

Genotype	Bin No.	No. of samples	Model Accuracy	Prediction Accuracy
BLU	Bin173	24	**46.7**	**70.0**
BOG	Bin175	30	**66.6**	**71.4**
FHI17	Bin159	19	29.1	46.6
FH25	Bin167	15	34.2	36.1
GON	Bin225	19	**62.5**	**65.6**
GRO	Bin219	29	**43.5**	**61.5**
KAY	Bin19	32	31.1	46.6
KM5	Bin161	34	**66.6**	**65.7**
M2	Bin191	35	46.8	52.7
NA31	Bin187	28	42.7	32.2
NAR7	Bin171	34	33.3	47.3
SUK	Bin43	35	34.5	38.8
	Mean		44.8	52.8

Table 3
Sensitivity-Specificity results for 12-binned genotypes predictions.

Genotype	TP	FN	FP	TN	Sensitivity	Specificity	PPV	NPV	Accuracy
FHI17	3	16	2	313	0.16	0.99	60.00	95.14	94.61
KM5	17	17	17	283	0.50	0.94	50.00	94.33	89.82
FH25	8	7	2	317	0.53	0.99	80.00	97.84	97.31
NAR7	29	5	20	280	**0.85**	**0.93**	59.18	98.25	92.51
BLU	21	3	9	301	**0.88**	**0.97**	70.00	99.01	96.41
BOG	18	12	8	296	0.60	0.97	69.23	96.10	94.01
NA31	18	10	7	299	0.64	0.98	72.00	96.76	94.91
KAY	22	10	21	281	0.69	0.93	51.16	96.56	90.72
M2	19	16	4	295	0.54	0.99	82.61	94.86	94.01
GRO	22	7	3	302	**0.76**	**0.99**	88.00	97.73	97.01
GON	15	4	9	306	**0.79**	**0.97**	62.50	98.71	96.11
SUK	18	17	25	274	0.51	0.92	41.86	94.16	87.43

TP = true positive; FN = false negative; FP = false positive; TN = true negative; PPV = positive predictive value [TP/(TP + FP)) * 100]; NPV = negative predictive value [(TN/(TN + FN))*100].

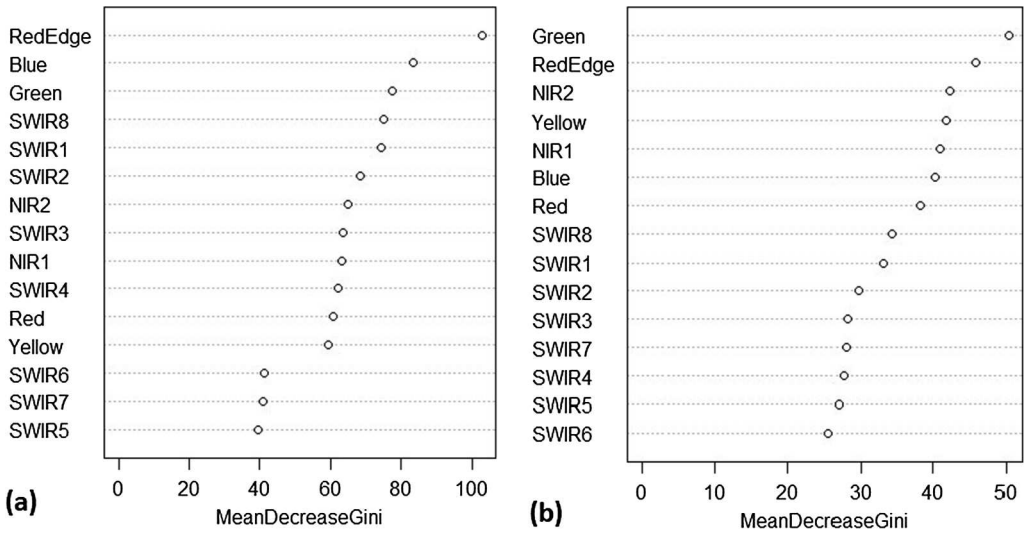

Fig. 13. ASD resampled WV3 bands importance rankings for genotype classifications from the RF in terms of mean decrease Gini index (a) and for parentage (Old and New) (b). WV3 bands refer to spectra aggregated to Woldview-3 wavelengths.

present on WV3 imagery, the higher wavelengths suffer lesser impacts. Thus the SWIR wavelengths were ranked higher than visible bands. Though the results were in consistent with Fletcher and Reddy (2016) and Fletcher (2016), the analysis of resampled spectra ($WV3_{ASD\text{-}genotype}$) identified the RedEdge, Green and NIR bands, along with few SWIR bands of higher importance (Fig. 13). Thus, the VIR of WV3 bands seemed inconclusive here and further investigation is required from a higher quality satellite data.

The statistical analysis of the data identified a number of specific wavelengths more sensitive to the differentiation of the 12 genotypes and their parentage. The model contained at least one wavelength in the Green spectral region (~510 nm) which indicated high correlation between healthy banana plants and chlorophyll/xanthophyll content (e.g., Rapaport et al., 2015). The Red range (~630 nm), RedEdge (RE), specifically at ~720–750 nm, and NIR spectral range (~875 nm, ~915 nm and ~1010 nm) (Fig. 9). The Green, Red, RedEdge, and NIR bands have been identified in other vegetation types as being sensitive to changes in plant physiological conditions and constraints (e.g., Fletcher and Reddy, 2016; Shapira et al., 2013; Suarez et al., 2016; Zhao et al., 2007; Mutanga and Skidmore (2007; 2004). The SWIR regions found significant at ~1225 nm (C—H 2nd OT-CH or oil/ lipid), ~1475 nm (N—H stretch 1st OT-CONHR), ~1520 nm (N—H stretch 1st OT- Urea), ~1750 nm (possibly fatty acid), ~2010 nm (C═O stretch 2nd OT-Urea), and ~2250 nm (fatty acid/ amino acid) were also identified in previous research (e.g., Hansen and Schjoerring, 2003; Rodriguez et al., 2006; Shi et al., 2015). These results indicate that spectral differences between the banana varieties are likely the result of variations in the composition (presence, concentration) of internal leaf constituents. Whilst this is highly plausible, care should be taken when extrapolating these results to other growing locations where external abiotic and biotic conditions (e.g. water stress, pest, disease, etc.) may also influence these particular constituents.

The field measures determined BLU, BOG, GON, GRO and KAY genotypes to be distinguishable from both full spectral range and also from the equivalent WV3 band widths of hyperspectral data. Validation from NARO researchers confirmed these identified cultivars as triploids and not of typical East African highland bananas varieties. BLU (ABB) and KAY (ABB) are used for cooking, while BOG (AAA) and GRO (AAA) are used as dessert bananas and GON (AAB) as plantain (Anyasi et al., 2013; Daniells et al., 2001; Tripathi et al., 2007) (please see Appendix A for banana variety name). This could be potentially be reason for their spectral variability and their prevalence in Uganda. All East African highland bananas are cooking type and triploid (AAA) with common ancestry and are therefore different to the above-mentioned varieties. The fact that the other cultivars have shared parentage may be the reason that they are harder to spectrally segregate. Nevertheless, the ability to differentiate these varieties from the others (and from each other) does have some merit as accurately mapping their distribution

Table 6
WV3 reflectance based RF model and prediction accuracy (%) for each genotype. The overall percentage accuracy is average of each genotype accuracy.

Genotype	Bin No.	Model Accuracy	Prediction Accuracy
BLU	Bin173	**57.1**	**66.6**
BOG	Bin175	**78.3**	**85.0**
FHI17	Bin159	50.0	25.0
FH25	Bin167	62.5	50.0
GON	Bin225	**83.3**	**74.8**
GRO	Bin219	**66.6**	**77.2**
KAY	Bin19	50.0	50.0
KM5	Bin161	40.0	50.0
M2	Bin191	40.0	30.0
NA31	Bin187	**65.3**	**62.4**
NAR7	Bin171	16.6	48.3
SUK	Bin43	50.0	52.3
	Mean	55.0	56.0

Table 5
Confusion matrix for ASD based RF classification accuracies for 12-binned banana parentage (new and old).

		Condition Positive New	Condition Negative Old
Predictive Condition positive	New	TP = 113	FP = 53
Predictive Condition negative	Old	FN = 52	TN = 116
	Prediction accuracy (%)	68.5	68.7
	Overall Accuracy (%)	68.6	

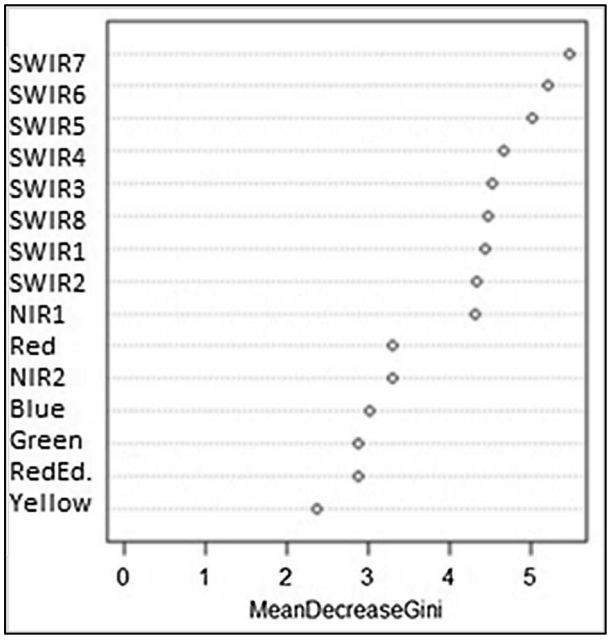

Fig. 14. Variable importance ranking of WV3 bands (satellite) for discriminating banana genotype in RF model.

can indicate how successfully current banana breeding programs are. A follow up study will check the robustness of developed method by applying it on classification of similar banana varieties grown in other districts of Uganda.

The translation of the full hyperspectral measures of banana plants to the spectral resolution of WV3 and then validation of results from actual WV3 satellite data, was found particularly crucial in this study. This provided a greater confidence in extrapolating these findings to satellite platforms. As WV3 imagery could be considered as expensive to purchase over large areas and is subject to spectral degradation from atmospheric scattering, haze, cloud and viewing geometry, this initial evaluation under 'controlled' conditions was deemed worthwhile. The results achieved and the additional benefit that high-resolution satellite imagery can provide information of plant structure (crown diameter, leaf orientation, canopy and sucker density etc.), all parameters that can be influenced by variety, does justify the further evaluation of the satellite platform itself. Particularly, the extrapolation of leaf-level reflectance of banana genotypes to satellite data is novel, given that the leaf angle and whole plant morphology as potential difference between the leaf-level and canopy-level reflectance profile (Knyazikhin et al., 2013). Gara et al. (2019) discussed two means of extrapolating leaf-level reflectance to canopy-level: direct extrapolation and canopy integrated method, both require samples to be taken from the sunlit top-of-the canopy layer. However, banana plants have different structure, configuration and leaf shape and as such detection of sunlit apex is difficult on satellite data. In this study, the pre-processing of WV3 data helped reducing the atmospheric impact by converting pixel value to surface reflectance. The 30 cm PAN sharpened WV3 imagery was used to delineate banana plant boundaries using object-based approach suggested by Johansen et al. (2014), from which canopy-level reflectance was extracted. This allowed the direct comparisons of two reflectances possible by accounting for banana canopy structures. Additionally, satellite (or airborne data) can provide essential insights into the spatial and temporal distribution of banana plants, which itself offers significant benefit for better predicting regional production, for biosecurity preparedness and for post natural disaster monitoring. Perry et al. (2018) mentioned the advantage of canopy-level reflectance measured from UAVs over leaf-level reflectance in characterizing %N variability across an orchard, and the canopy N maps generated from the UAV imagery could be highly valuable as related technologies. As the presence of environmental haze and overcast condition at this time of year (Feb-Mar) is a common phenomenon of tropical regions, such as Uganda, getting a high quality data will be a challenging task. Although, a few atmospheric correction methods were applied to reduce the influence of haze, a more efficient haze removal algorithm is required to reduce the impacts in visible and NIR regions (e.g., Zha et al., 2012; Sun et al., 2017). For future evaluation of remote sensing in this region, some consideration would be required to determine the optimal timing of capture and even platform used. Lower altitude airborne or even UAV based platforms may be useful. Freely available satellite data, such as Sentinel 2, provide spectral resolutions matching some of the WV3 bands and hence can be used for further investigation in banana genotype classification. Gara et al. (2019) explored the upscaling of leaf-level measurements to simulated Sentinel2 data. The additional measure of banana plant health, including incidences of pest, disease and management can significantly improve Uganda's ongoing surveillance and response to high risk constraints.

This study as such established several sampling protocols that combined DNA tissue testing with the hyperspectral measurement of individual banana leaves. Whilst the results achieved only partial success in differentiating all 12 varieties, it was successful in determining the spectral differences (and similarities) between typical ECA banana varieties, and in their parentage (usage). The information will be used in building spectral library for banana verities by following spectral library building protocols as suggested by Rao et al. (2007) and Jiménez and Díaz-Delgado (2015). The method can be extended to other crop variety predictions at other locations. It is noteworthy that the sampling protocols followed in this study provided the best quality data (limiting disease, developmental differences, etc.) that when implemented into models stand the best chance of achieving positive results under, real-world scenarios. This part is proposed in a follow up study on prediction of same varietal population in the other districts (Isingiro, Mbarara) of Uganda, to assess the modelling performance under varying locational and seasonal conditions. The hyperspectral measurements and WV3 data have already been obtained for these two districts.

The outcomes from this scoping study demonstrate some potential of remote sensing as a beneficial technology for differentiating the types of banana (cooking type, plantain, dessert) and also for the discrimination of some varieties. In the 2008/09 agricultural census, UBoS (Uganda Bureau of Statistics) has used this categorization to report planted, production and productivity (PARAM, 2015; UBoS, 2010). Thus further evaluating of the accuracies of remote sensing data for the elicitation of banana type at the household level would offer significant benefit to future LSMS-ISA and WB survey missions in Sub-Saharan Africa.

6. Conclusions

The study advanced the potential for using in-situ hyperspectral remote sensing data to identify banana genotypes under Ugandan growing conditions. A controlled experiment undertaken within the National Banana Research Program site at National Agricultural Research Organization (NARO) research station in Kampala, Uganda, ensured the influence of non-varietal parameters were minimised and as such any spectral variation was genotype specific. The statistical analysis of reflectance spectra measured from 12 banana genotypes (identified after DNA tissue testing of 43 banana varieties), achieved higher prediction accuracies for BLU, BOG, GON, GRO and KAY genotypes. The further extrapolation of the hyperspectral measures to bandwidths consistent with the WV3 satellite also achieved encouraging results. This outcome not only presents as a novel approach for transitioning ground based leaf measures to satellite based sensors, but also supports the increased scalability of banana varietal mapping across Uganda, through remote sensing.

The accurate discrimination of banana usage from both the hyperspectral and satellite based sensors also offers significant benefit. By mapping of distribution of old and new varieties, NARO can form a better understanding of the extent of adoption of newly developed genotypes. This information can better inform of market acceptance as well as future sustainability to pest and disease as well as resilience to food shortages.

Whilst the results presented in this study support the integration of DNA tissue testing and remote sensing for discriminating some banana cultivars, further research is required to better understand the influence of growing location and season on the spectral responses. Also the evaluation of additional platforms such as airborne and UAV may reduce the influences of atmospheric haze that is prevalent in the Ugandan region.

This study has been a collaborative partnership between the Uganda Bureau of Statistics under the World Bank LSMS-CGIAR SPIA partnership and the University of New England.

Funding

This research was funded by the World Bank, contract number 7185845.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the support given by NARO staff members in conducting field work in Uganda. The funding from the World Bank LSMS unit is well acknowledged that facilitated this research work. The valuable suggestions received from three anonymous reviewer are highly commendable that improved the manuscript significantly.

Appendix A

Banana variety Code and Name sampled at NARO Research Station after DNA fingerprinting results

BLU	Bin173	Bluggoe/Kivuvu
BOG	Bin175	Bogoya cavendishes/ williams/grandnaine/lakatan/robusta/ dwarf/ chinese cavendish
FHI17	Bin159	FHIA17/KABANA 3H
FH25	Bin167	FHIA25/KABANA 7H
GON	Bin225	Gonja/Nakatansese/Manjaya/Kakira/Nakakongo/Mukono/Obinolewayi/Nig-erian Agbaba
GRO	Bin219	Bogoya local bogoya/Gros Michel
KAY	Bin19	Kayinja
KM5	Bin161	KM5/KABANA 6H
M2	Bin191	M2
NA31	Bin187	NARITA 31 – Pisang Ceylan
NAR7	Bin171	NARITA 7 -M9/Kiwangazi
SUK	Bin43	Sukari Ndizi/Kabaragara

Appendix B

Abbreviation code description

ASD	Analytical Spectral Devices
CGIAR	Consultative Group for International Agricultural Research
CPPLS	Canonical Powered Partial Least Squares
ECA	East and Central African
FAO	Food and Agriculture Organization
LSMS-ISA	Living Standards Measurement Study – Integrated Surveys on Agriculture
ML	Machine Learning
NARO	National Agricultural Research Organization
PCA	Principal Component Analysis
PLSR	Partial Least Squares Regression
RF	Random Forests
RMSEP	Root Mean Square Error of Prediction
SPIA	Standing Panel on Impact Assessment
SWIR	Short-Wave Infrared
UBoS	Uganda Bureau of Statistics
NIR	Near-Infrared
WB	World Bank
WV3	Worldview 3

Appendix C. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2020.06.023.

References

Ajayi, S., Reddy, S., Gowda, P., Xue, Q., Rudd, J., Pradhan, G., Liu, S., Stewart, B., Biradar, C., Jessup, K., 2016. Spectral reflectance models for characterizing winter wheat genotypes. J. Crop Improve. 30, 176–195.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. 7, 16398–16421.

Anyasi, T.A., Jideani, A.I., Mchau, G.R., Safety, F., 2013. Functional properties and

postharvest utilization of commercial and noncommercial banana cultivars. Comprehens. Rev. Food Sci. 12, 509–522.
Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sens. 5, 949–981.
Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal. Methods 5, 3790–3798.
Barnes, M.L., Breshears, D.D., Law, D.J., van Leeuwen, W.J., Monson, R.K., Fojtik, A.C., Barron-Gafford, G.A., Moore, D.J., 2017. Beyond greenness: Detecting temporal changes in photosynthetic capacity with hyperspectral reflectance data. PLoS ONE 12, e0189539.
Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31.
Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
Carfagna, E., Gallego, F.J., 2005. Using remote sensing for agricultural statistics. Int. Stat. Rev. 73, 389–404.
Carletto, G., Jolliffe, D., Banerjee, R., 2015. From tragedy to renaissance: improving agricultural data for better policies. J. Dev. Study 51, 133–148.
Chivasa, W., Mutanga, O., Biradar, C., 2019. Phenology-based discrimination of maize (Zea mays L.) varieties using multitemporal hyperspectral data. J. Appl. Remote Sens. 13, 017504.
Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69.
Christiaensen, L., Demery, L., 2017. Agriculture in Africa: Telling Myths from Facts. The World Bank, Washington DC.
Crabbe, R.A., Lamb, D., Edwards, C., 2020. Discrimination of species composition types of a grazed pasture landscape using Sentinel-1 and Sentinel-2 data. Int. J. Appl. Earth Observ. Geoinf. 84, 101978.
Daniells, J., Jenny, C., Karamura, D., Tomekpe, K., 2001. Musalogue: a catalogue of Musa germplasm. Diversity in the genus Musa (E. Arnaud and S. Sharrock, compil.). International Network for the Improvement of Banana and Plantain, Montpellier, France.
Digital Globe, 2014. Worldview 3 data sheet. https://www.spaceimagingme.com/downloads/sensors/datasheets/DG_WorldView3_DS_2014.pdf. (date accessed 15 Jan 2020).
Duan, S.B., Li, Z.L., Wu, H., Tang, B.H., Ma, L., Zhao, E., Li, C., 2014. Inversion of the PROSAIL model to estimate leaf area index of maize, potato, and sunflower fields from unmanned aerial vehicle hyperspectral data. Int. J. Appl. Earth Observ. Geoinf. 26, 12–20.
Esbensen, K.H., Guyot, D., Westad, F., Houmoller, L.P., 2002. Multivariate Data Analysis: In Practice: An Introduction to Multivariate Data Analysis and Experimental Design. Aalborg University, Esbjerg, Oslo, Norway.
FAO, 2011. Food and Agriculture Organization of the United Nations. FAOSTAT.
FAO, 2017a. Data: Production Quanitity. FAO, Rome Italy.
FAO, 2017b. Review of the Available Remote Sensing Tools, Products, Methodologies and Data to Improve Crop Production Forecasts. FAO, Rome, Italy.
Féret, J.B., Gitelson, A., Noble, S., Jacquemoud, S., 2017. PROSPECT-D: towards modeling leaf optical properties through a complete lifecycle. Remote Sens. Environ. 193, 204–215.
Fletcher, R.S., Reddy, K.N., 2016. Random forest and leaf multispectral reflectance data to differentiate three soybean varieties from two pigweeds. Comput. Electron. Agric. 128, 199–206.
Fletcher, R.S., 2016. Using vegetation indices as input into Random forest for Soybean and Weed classification. Am. J. Plant Sci. 7, 2186–2198.
Fu, P., Meacham-Hensold, K., Guan, K., Bernacchi, C., 2019. Hyperspectral leaf reflectance as proxy for photosynthetic capacities: an ensemble approach based on multiple machine learning algorithms. Front. Plant Sci. 10.
Gara, T.W., Skidmore, A.K., Darvishzadeh, R., Wang, T., 2019. Leaf to canopy upscaling approach affects the estimation of canopy traits. GIScience Remote Sens. 56 (4), 554–575.
Garriga, M., Romero-Bravo, S., Estrada, F., Escobar, A., Matus, I.A., del Pozo, A., Astudillo, C.A., Lobos, G.A., 2017. Assessing wheat traits by spectral reflectance: do we really need to focus on predicted trait-values or directly identify the elite genotypes group? Front. Plant Sci. 8, 280.
Hansen, P., Schjoerring, J., 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sens. Environ. 86, 542–553.
Heckmann, D., Schlüter, U., Weber, A.P., 2017. Machine learning techniques for predicting crop photosynthetic capacity from leaf reflectance spectra. Mol. Plant 10, 878–890.
Hennessy, A., Clrake, K., Lewis, M., 2020. Hyperspectral classification of Plants: A review of waveband selection generalisability. Remote Sens. 12, 113. https://doi.org/10.3390/rs12010113.
Indahl, U.G., Liland, K.H., Næs, T., 2009. Canonical partial least squares—a unified PLS approach to classification and regression problems. J. Chemomet.: J. Chemomet. Soc. 23, 495–504.
Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P.J., Asner, G.P., François, C., Ustin, S.L., 2009. PROSPECT+ SAIL models: A review of use for vegetation characterization. Remote Sens. Environ. 113, S56–S66.
Jay, S., Gorretta, N., Morel, J., Maupas, F., Bendoula, R., Rabatel, G., Dutartre, D., Comar, A., Baret, F., 2017. Estimating leaf chlorophyll content in sugar beet canopies using millimeter-to centimeter-scale reflectance imagery. Remote Sens. Environ. 198, 173–186.
Jiménez, M., Díaz-Delgado, R., 2015. Towards a standard plant species spectral library protocol for vegetation mapping: a case study in the Shrubland of Doñana National Park. ISPRS Int. J. Geo-Inf. 4, 2472–2495.
Johansen, K., Phinn, S., Witte, C., Philip, S., Newton, L., 2009. Mapping banana plantations from object-oriented classification of SPOT-5 imagery. Photogramm. Eng. Remote Sens. 75, 1069–1081.
Johansen, K., Sohlbach, M., Sullivan, B., Stringer, S., Peasley, D., Phinn, S., 2014. Mapping banana plants from high spatial resolution orthophotos to facilitate plant health assessment. Remote Sens. 6, 8261–8286.
Kaufman, Y.J., 1993. Aerosol optical thickness and atmospheric path radiance. J. Geophys. Res. 98, 2677–2692.
Kiiza, B., Abele, S., Kalyebara, R., 2004. Market opportunities for Ugandan banana products: National, regional and global perspectives. Uganda J. Agric. Sci. 9, 743–749.
Knyazikhin, Y., Schull, M.A., Stenberg, P., Mõttus, M., Rautiainen, M., Yang, Y., Marshak, A., et al., 2013. Hyperspectral remote sensing of foliar nitrogen content. Proc. Natl. Acad. Sci. 110 (3), E185–E192. https://doi.org/10.1073/pnas.1210196109.
Kosmowski, F., Aragaw, A., Kilian, A., Ambel, A., Ilukor, J., Yigezu, B., Stevenson, J.J.E.A., 2019. Varietal identification in household surveys: results from three household-based methods against the benchmark of DNA fingerprinting in southern Ethiopia. J. Exp. Agric. 55, 371–385.
Lee, L.C., Liong, C.-Y., Jemain, A.A., 2018. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst 143, 3526–3539.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.
Liland, K.H., Indahl, U.G., 2009. Powered partial least squares discriminant analysis. J. Chemomet.: J. Chemomet. Soc. 23, 7–18.
Lobell, D.B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., Murray, S., 2018. Eyes in the Sky, Boots on the Ground: Assessing Satellite-and Ground-based Approaches to Crop Yield Measurement and Analysis in Uganda. Policy Research working paper;no. WPS 8374;LSMS Washington, D.C. The World Bank.
Mahmood, T., Martens, H., Saebo, S., Warringer, J., Snipen, L., 2011. A partial least squares based algorithm for parsimonious variable selection. Algorithms Mol. Biol. 6, 27. https://doi.org/10.1186/1748-7188-6-27.
Martínez-Martínez, V., Gomez-Gil, J., Machado, M.L., Pinto, F.A., 2018. Leaf and canopy reflectance spectrometry applied to the estimation of angular leaf spot disease severity of common bean crops. PLoS ONE 13, e0196072.
Meacham-Hensold, K., Montes, C.M., Wu, J., Guan, K., et al., 2019. High-throughput field phenotyping using hyperspectral reflectance and partial least squares regression (PLSR) reveals genetic modifications to photosynthetic capacity. Remote Sens. Environ. 231, 111176.
Mevik, B.-H., Wehrens, R., 2007. The Pls Packgae: principal component and partial least squares and regression in R. J. Stat. Softw. 18, 1–24.
Mishra, P., Asaari, M.S.M., Herrero-Langreo, A., Lohumi, S., Diezma, B., Scheunders, P., 2017. Close range hyperspectral imaging of plants: a review. J. Biosyst. Eng. 164, 49–67.
Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 66, 247–259.
Mutanga, O., Skidmore, A.K., 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. Int. J. Remote Sens. 25, 3999–4014.
Mutanga, O., Skidmore, A.K., 2007. Red edge shift and biochemical content in grass canopies. ISPRS J. Photogramm. Remote Sens. 62, 34–42.
Nyombi, K., 2013. Towards sustainable highland banana production in Uganda: opportunities and challenges. Afr. J. Food Agric. Nutr. Dev. 13, 7544–7561.
Øvergaard, S.I., Isaksson, T., Korsaeth, A., 2013a. Prediction of wheat yield and protein using remote sensors on plots—Part I: Assessing near infrared model robustness for year and site variations. J. Near Infrared Spectrosc. 21, 117–131.
Øvergaard, S.I., Isaksson, T., Korsaeth, A., 2013b. Prediction of wheat yield and protein using remote sensors on plots—Part II: Improving prediction ability using data fusion. J. Near Infrared Spectrosc. 21, 133–140.
PARAM, 2015. Platform for Agricultural Risk Management: Agricultural Risk Assessment Study Uganda, 2015.
Peñuelas, J., Filella, I., 1998. Visible and near-infrared reflectance techniques for diagnosing plant physiological status. Trends Plant Sci. 3, 151–156.
Peerbhay, K.Y., Mutanga, O., Ismail, R., 2013. Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu–Natal, South Africa. ISPRS J. Photogramm. Remote Sens. 79, 19–28.
Perry, E.M., Goodwin, I., Cornwall, D., 2018. Remote sensing using canopy and leaf reflectance for estimating natrogen status in red-blush pears. HortSc 53 (1), 78–83.
Rajkumar, P., Wang, N., EImasry, G., Raghavan, G., Gariepy, Y., 2012. Studies on banana fruit quality and maturity stages using hyperspectral imaging. J. Food Eng. 108, 194–200.
Rao, N.R., Garg, P., Ghosh, S.K., 2007. Development of an agricultural crops spectral library and classification of crops at cultivar level using hyperspectral data. Precis. Agric. 8, 173–185.
Rapaport, T., Hochberg, U., Shoshany, M., Karnieli, A., Rachmilevitch, S., 2015. Combining leaf physiology, hyperspectral imaging and partial least squares-regression (PLS-R) for grapevine water status assessment. ISPRS J. Photogramm. Remote Sens. 109, 88–97.
Robson, A., Rahman, M.M., Muir, J., 2017. Using Worldview satellite imagery to map yield in Avocado (Persea americana): a case study in Bundaberg, Australia. Remote Sens. 12 (9), 1223.
Rodriguez, D., Fitzgerald, G.J., Belford, R., Christensen, L.K., 2006. Detection of nitrogent difficiency in wheat from spectral reflectance indices and basic crop eco-physiological concepts. Austr. J. Agric. Res. 57, 781–789.
Sahoo, R.N., Ray, S., Manjunath, K., 2015. Hyperspectral remote sensing of agriculture. Curr. Sci. 848–859.

Schmidt, K., Skidmore, A., 2004. Smoothing vegetation spectra with wavelets. Int. J. Remote Sens. 25, 1167–1184.
Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. 406, 109–120.
Schwieder, M., Leitão, P., Suess, S., Senf, C., Hostert, P., 2014. Estimating fractional shrub cover using simulated EnMAP data: A comparison of three machine learning regression techniques. Remote Sens. 6, 3427–3445.
Serbin, S.P., Dillaway, D.N., Kruger, E.L., Townsend, P.A., 2011. Leaf optical properties reflect variation in photosynthetic metabolism and its sensitivity to temperature. J. Exp. Bot. 63, 489–502.
Shapira, U., Herrmann, I., Karnieli, A., Bonfil, D.J., 2013. Field spectroscopy for weed detection in wheat and chickpea fields. Int. J. Remote Sens. 34, 6094–6108.
Shi, T., Wang, J., Liu, H., Wu, G., 2015. Estimating leaf nitrogen concentration in heterogeneous crop plants from hyperspectral reflectance. Int. J. Remote Sens. 36, 4652–4667.
Silva-Perez, V., Molero, G., Serbin, S.P., Condon, A.G., Reynolds, M.P., Furbank, R.T., Evans, J.R., 2017. Hyperspectral reflectance as a tool to measure biochemical and physiological traits in wheat. J. Exp. Bot. 69, 483–496.
Stewart Jr., C.N., Via, L.E., 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques 14 (5), 748–750.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9, 307.
Suarez, L., Apan, A., Werth, J., Sensing, R., 2016. Hyperspectral sensing to detect the impact of herbicide drift on cotton growth and yield. ISPRS J. Photogramm. Remote Sens. 120, 65–76.
Suarez, L., Apan, A., Werth, J., Sensing, R., 2017. Detection of phenoxy herbicide dosage in cotton crops through the analysis of hyperspectral data. Int. J. Remote Sens. 38 (23), 6528–6553. https://doi.org/10.1080/01431161.2017.1362128.
Sun, L., Latifovic, R., Poulliot, D., 2017. Haze remoeval based on fully automated and improved haze optimized transformation of Landsat imagery over land. Remote Sens. 9, 972.
Tinzaara, W., Ocimati, W., Kikulwe, E., Otieno, G., Stoian, D., Blomme, G., 2018. Challenges and opportunities for smallholders in banana value chains. In Kema, A.G. D. (Ed.), chieving sustainable cultivation of bananas. Vol. 1 Cultivation techniques. Burleigh Dodds, pp. 1–26.
Tripathi, L., Tripathi, J.N., Vroh-Bi, I., 2007. Bananas and plantains (Musa spp.): Transgenics and biotechnology. Transgenic Plant J. 1, 185–201.
Ubos, 2010. Uganda Bureau of Statistics- Uganda Census of Agriculture 2008/2009, Volume III, Agriculture House Hold and Holding Characteristics Report (2010).
Vaiphasa, C., 2006. Consideration of smoothing techniques for hyperspectral remote sensing. ISPRS J. Photogramm. Remote Sens. 60, 91–99.
Van Asten, P., Gold, C.S., Wendt, J., De Waele, D., Okech, S., Ssali, H., Tushmereirwe, W.K., 2003. The contribution of soil quality to yield and its relationship with other factors in Uganda. In: Blomme, G., Gold, C.S., Karamura, E. (Eds.), Farmer Participatory Testing of IPM Options for Sustainable Banana Production in Eastern Africa, Seeta, Uganda. International Plant Genetic Resources Institute, Montpellier, pp. 100–115.
Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J.P., Veroustraete, F., Clevers, J.G., Moreno, J., 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–a review. ISPRS J. Photogramm. Remote Sens. 108, 273–290.
Vijaya Kumar, P., Ramakrishna, Y., Bhaskara Rao, D., Sridhar, G., Srinivasa Rao, G., Rao, G., 2005. Use of remote sensing for drought stress monitoring, yield prediction and varietal evaluation in castor beans (Ricinus communis L.). Int. J. Remote Sens. 26, 5525–5534.
Wang, H., Chen, J., Lin, H., Yuan, D., 2010. Research on effectiveness of hyperspectral data on identifying rice of different genotypes. Remote Sens. Lett. 1, 223–229.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58, 109–130.
Wong, C.Y., Gamon, J.A., 2015. Three causes of variation in the photochemical reflectance index (PRI) in evergreen conifers. New Phytol. 206, 187–195.
Zha, Y., Gao, J., Jiang, J., Lu, H., Huand, J., 2012. Normalized diffrence haze index: a new spectral index for monitoring urban air poluution. Int. J. Remote Sens. 33 (1), 309–321.
Zhao, D., Reddy, K.R., Kakani, V.G., Read, J.J., Koti, S., 2007. Canopy reflectance in cotton for growth assessment and lint yield prediction. Eur. J. Agron. 26, 335–344.
Zhao, J., Bodner, G., Rewald, B., 2016. Phenotyping: using machine learning for improved pairwise genotype classification based on root traits. Front. Plant Sci. 7, 1864.