Exploring the potential of Sentinel-2 broadband spectral features for discrimination of invasive *Lantana camara* L. among co-occurring species

Julius Maina Waititu[1], Charles Ndegwa Mundia[2], Arthur W Sichangi[2]
[1]Department of Spatial and Environmental Planning, Kenyatta University, Nairobi, Kenya
[2]Institute of Geomatics, GIS and Remote Sensing, Dedan Kimathi University of Technology, Nyeri, Kenya
waititujulius@gmail.com, waititu.julius@ku.ac.ke

DOI: https://dx.doi.org/10.4314/sajg.v14i2.2

Abstract

Lantana camara L. (LC), an alien invasive plant species, negatively impacts natural habitats globally. In East Africa, it is one of the habitat's transforming species that requires urgent mapping to support conservation actions. Its spatial distribution has not been adequately established, especially within forest habitats. In this study, we sought to identify Sentinel-2 broadband spectral features that could discriminate LC from co-occurring vegetation in a forest habitat. In-situ leaf-level hyperspectral measurements of LC and co-occurring species, namely, Neonotonia wightii (NW), Cucumis maderaspatanus (CM) and Ocimum gratissimum (OG), were collected in a one-hectare site in the Muringato forest area during the dry and wet seasons using two handheld spectroradiometers covering wavelength ranges of between 340–820 nm and 635–1100 nm. The leaf-level reflectance measurements were used to simulate Sentinel-2 wavebands, which were subsequently used to compute new band combination indices of type Normalized Difference (ND), Simple Ratio (SR), Difference (D) and Inverse Difference (ID) and published Sentinel-2 multispectral indices. The most spectrally significant features were selected using the Boruta and the Guided Regularized Random Forest (GRRF) methods. Jeffries–Matusita (JM) distance analysis was used to quantify the spectral separability of species class pairs using selected spectral features. The findings of this study showed that the selected Sentinel-2 spectral features that produced perfect separability accuracies of ≥97% for LC class pairs consisted of mostly the newly developed band combination indices and, to a lesser extent, the published Sentinel-2 indices. Notably, the separability analysis produced a unique set of spectral variables that accentuated the spectral properties of LC class pairs in both seasons and further pointed to the influence of the seasonal spectral variabilities of the species. Moreover, the Boruta method resulted in the selection of fewer spectral variables (2–12 variables) than the GRRF method, which resulted in the selection of two to19 variables for species class pair separability. Overall, the separability results demonstrate the potential of separating LC from other vegetation with freely available Sentinel-2 image data. The significant spectral variables identified in this study could be used for the seasonal mapping of LC and for further aid in the early detection and targeted management of the species in affected forest habitats.

Keywords: *Alien invasive plant species, Lantana camara L., Sentinel-2, field spectroscopy, multispectral vegetation indices*

1. Introduction

Alien invasive species pose greater risks to the optimal functioning of natural ecosystems and the growth of local native flora (Rajah *et al*., 2019). Despite history showing unintended negative outcomes after alien species have been introduced to ecosystems where they do not occur naturally, dangerous introductions continue. Some alien species, for example, *Lantana camara* L. and *Opuntia stricta,* have drawn global attention owing to their serious impacts on biological diversity and human activities and have been listed among the world's worst invasive alien species (Global Invasive Species Database, 2023). Positive strides have been made towards growing public awareness of the risks posed by alien invasive plant species. Witt *et al*. (2018) have provided a glimpse of the several species that are considered to have the greatest impacts in terms of transforming natural ecosystems within the East African region. With such information in the public domain, collective conservation actions could be devised through policy and community engagement. Key conservation areas in Kenya, particularly in the Mt. Kenya Forest and the Aberdare Forest reserves, have had to face the challenges of invading plant species such as *Lantana camara* L. (LC), *Caesalpinia decapelata*, *Datura dothistroma*, *Acacia melanoxylon*, *Solanum incanum*, *Acacia meansii, Resinus communis* and *Rubus stendineri* (Kenya Forest Service, 2010; Kenya Wildlife Service, 2010). Among these species, LC has been cited extensively in other jurisdictions as a problematic species whose adaptability to a wide range of natural environments has hindered the containment efforts in respect of its spread (Goncalves *et al*., 2014; Kimothi and Dasari, 2010; Negi *et al*., 2019). The risk of LC spreading into protected ecosystems such as Mt. Kenya and the Aberdare range may be greater under future changing climatic conditions (Waititu *et al*., 2022). Owing to its ability to dominate other plant species within infested areas, thereby leading to diminished agricultural production and livestock pasture areas, the negative impacts of LC on local livelihoods have been felt in some parts of Uganda (Shackleton *et al*., 2017).

Efforts to mitigate the spread of invasive plant species often require regular determinations of their spatial distribution to identify the trends and patterns of their spread for better decision-making and sustainable management of the environment (Royimani *et al*., 2019). In this regard, several studies have demonstrated the utility of leaf-level and canopy-level field hyperspectral reflectance datasets in discriminating plant species (Große-Stoltenberg *et al*., 2016; Mureriwa *et al*., 2016; Mudereri *et al.*, 2020). Hyperspectral datasets provide high spectral resolution information that enhances vegetation mapping at the community and species level (Hennessy *et al*., 2020). Therefore, the use of these datasets could be explored for the mapping of invasive plant species. However, owing to inadequate resources to acquire them, especially in low-income countries, there is a challenge in accessing hyperspectral imagery datasets for vegetation mapping. It is envisaged that the current situation might change soon as new hyperspectral earth observation sensors planned for deployment begin providing the much-needed hyperspectral imagery datasets for vegetation mapping (Transon *et al.*, 2018). In the meantime, some recent studies have exploited the benefits of the readily available multispectral images (e.g., those from

the new generation satellite imaging sensors) such as the Sentinel-2 Multispectral Instrument (MSI), which has improved performance accuracies in the detection of various vegetation species (Transon *et al*., 2018). Recently, Dube *et al*. (2020) established that mapping LC with Sentinel-2 image data produces higher classification accuracies than would be the case using Landsat 8 OLI data. The superior performance may be attributed to the spatial resolutions of 10 m, 20 m and 60 m of the new generation Sentinel-2 improved band, and its improved spectral wavelength ranges that enhance the detection of various geophysical variables (Transon *et al*., 2018).

It is well known that sound management decisions concerning invasive alien plant species require the development of a reliable information database in respect of the spatial distribution of the species in a given area. One of the strategies used for mapping the spatial distribution of species involves the use of spectral vegetation indices. Therefore, this study aims to derive useful Sentinel-2 waveband spectral indices from *in situ* leaf-level measurements through band combinations and to assess them with the published Sentinel-2 indices for discriminating LC from its co-occurring species. As pointed out in the literature, the use of spectral vegetation indices may identify changes in the bio-physical variables of plants (Mahlein *et al*., 2013) and allow for distinctions to be made among the different vegetation covers (Große-Stoltenberg *et al*., 2016). When extracting a given vegetation characteristic, vegetation indices help differentiate background reflectance from soils, atmospheric disturbances, the sensor angle of orientation, and the sun's azimuth (Fang and Liang, 2014). A review by Royimani *et al*. (2019) listed several vegetation indices, namely, Normalized Difference Vegetation Index (NDVI), Principal Component Analysis (PCA), Enhanced Vegetation Index (EVI), Tasseled Cap (TCap), Simple Ratio (SR), Soil Adjusted Vegetation Index (SAVI), Visible Atmospherically Resistant Index (VARI), and Normalized Difference Moisture Index (NDMI) as the most commonly used indices in the mapping of alien invasive plant species. This review points to the importance of assessing the applicability of the existing indices in vegetation studies as they may perform differently depending on the different vegetation covers. Furthermore, airborne and space-borne remote sensing sensors possess varying configurations, such as spectral sensing range, instrumentation, resolutions and platforms, thereby posing a challenge in developing a single vegetation index for mapping vegetation with the differing datasets that result (Xue and Su, 2017). By using field spectral reflectance data and resampling to the sensor-specific wavebands, a thorough analysis could be applied through feature selection and separability analysis to identify spectral indices or variables that could potentially single out LC from its other co-occurring species.

2. Materials and Methods

2.1. Area of Study

The area chosen for study is in Central Kenya, Nyeri County. It lies between Latitudes 0˚ 38’ 49” S and 0˚ 0’ 23” N and Longitudes 36˚ 36’ 17” E and 37˚ 18’ 30” E, as shown in Figure 1. Notable physical features in Nyeri County include Mt Kenya, to the east, which stands at an elevation of 5,199 m above mean sea level (a.s.l) and the Aberdare range, to the west, standing at 3,999 m a.s.l. The selected sampling site is a one-hectare area in the Muringato forest which is rich in native trees and floral diversity. The site contains a large patch of shrub vegetation composed of mainly LC and co-occurring climber species, namely, *Neonotonia wightii* (NW), and *Cucumis maderaspatanus* L. (CM) and the perennial herb species, *Ocimum gratissimum* L. (OG). The site slopes gradually towards the nearby Muringato River to the south and has well-drained red soil. The Nyeri County climate consists of two rainy seasons, generally occurring in March–May (rains of lengthy duration) and October–December (rains of short duration). The onset and duration of these rainy seasons may vary yearly (MoALF, 2016). On average, the annual rainfall ranges between 1200 and 1600 mm and 500 and 1500 mm during the long and short rain seasons, respectively, while the monthly mean temperatures range between 12.8˚C and 20.8˚C (Government of Kenya, 2018).

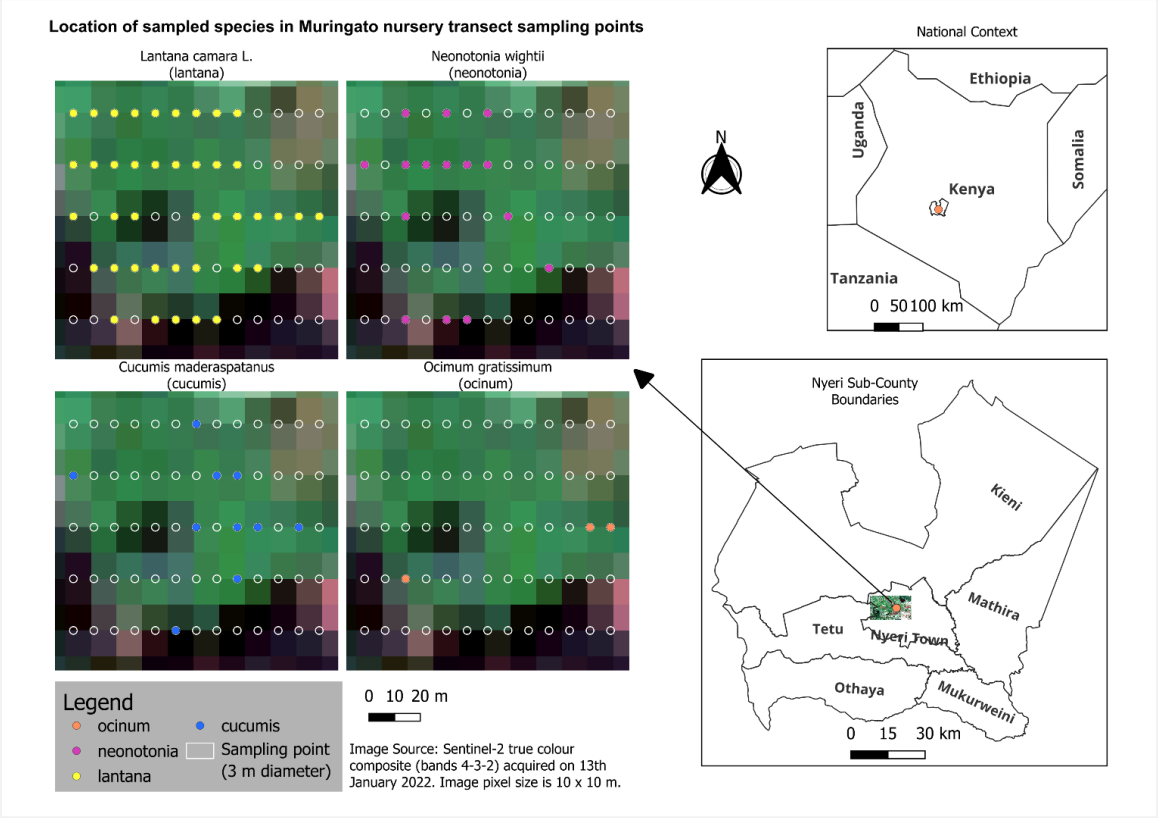

Figure 1. The geographical location of the study area and the hyperspectral data sampling layout at the one-hectare Muringato forest site. (Administrative boundary layer source: GADM database (www.gadm.org) under CC BY 4.0 license (https://gadm.org/license.html)).

2.2. Field Hyperspectral Data Collection

Systematic transect lines were designed over a Sentinel-2 image scene of the site covering 100m (length) and 100m (width). Five transect lines were laid out at a spacing of 20m and sampling locations were placed along the transect lines at eight metres apart (Figure 1). Spacing the sampling locations at eight metres ensured that the sampled species at these locations were individual plant species. LC grows to a height of two to three metres and its branches cover an area of approximately one square metre (Sharma *et al*., 1981). The field measurement campaign was planned to coincide with the Sentinel-2 image acquisition date (±2 days) during the dry season (August 2021) and wet season (January 2022) (see Figure 2). Sampling locations falling on bare ground were not sampled (Figure 1). Samples from 53 sampling locations were measured. Leaf-level hyperspectral reflectance measurements of the identified four dominant species, LC, NW, CM and OG, were taken using two portable field spectroradiometers manufactured by Apogee Instruments, Inc (https://www.apogeeinstruments.com) and having wavelengths ranging from 340 to 820nm and 635 to 1100nm. The wavelength resolution of the instrument is three nanometres (3nm) (full-width half maximum) and the spectrum recording interval is one nanometre (1nm). Three to five mature leaf samples were selected from branches forming part of the top canopy of the sunlit plants. The leaves were plucked from these branches and placed on a black cardboard sheet for immediate spectral reflectance measurement. The measurements were done with a nadir-looking 25° spectroradiometer reflectance head (Mureriwa *et al*., 2016). The internal averaging for the instrument was set to 3 before a final measurement was recorded for each leaf. Calibration of the spectroradiometers was done using a Zenith Polymer® white reflectance panel (~99% of reflectance) manufactured by https://sphereoptics.de/ and the instrument dark noise was corrected by taking the dark reflectance value by using the black cap supplied with the spectroradiometers. Spectral reflectance measurements were taken under sunny conditions − between 10:00 am and 02:30 pm (Mureriwa *et al*., 2016). The instruments were regularly calibrated during measurements to account for the sun's illumination changes in the field.

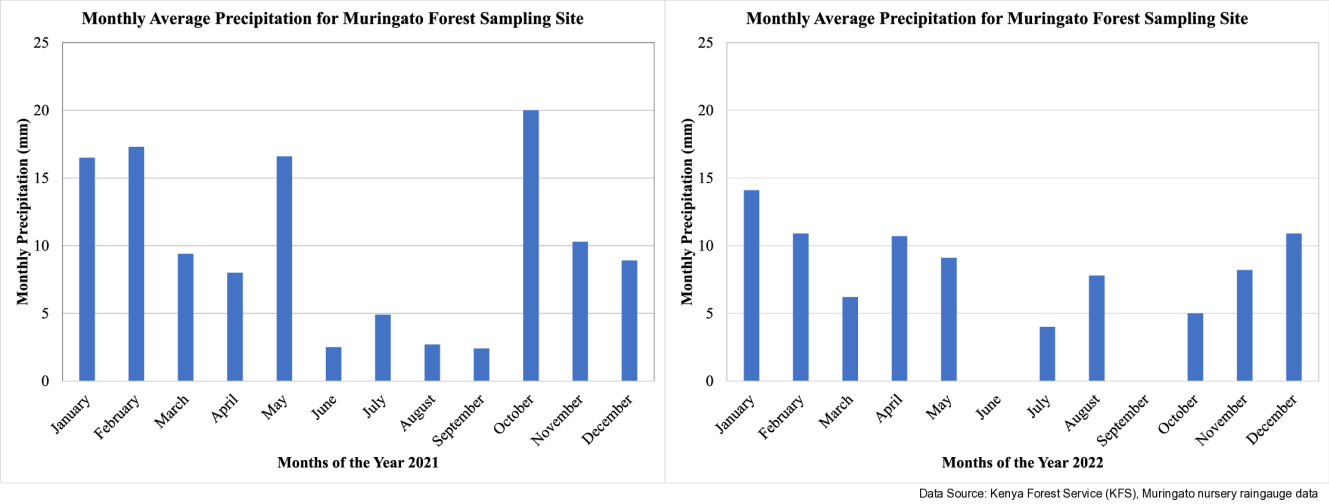

Figure 2. Monthly average precipitation values for the Muringato forest sampling site for 2021 and 2022.

2.3. Hyperspectral Data Pre-processing

Spectral reflectance data pre-processing and analysis were done in R Statistical software (R Core Team, 2021) and the R package "*hsdar*" (Lehnert *et al*., 2019). Raw spectral reflectance data were subjected to a filtering process to remove spectral reflectance noise. A Savitsky-Golay filter with a length of 25nm was used for this purpose and followed by a similarity test using spectral angle mapper (SAM) (Chauhan and Mohan, 2014) to detect and remove outliers. Pre-processed hyperspectral reflectance data (Table 1) were then resampled to Sentinel-2 sensor wavebands. Sentinel-2 wavebands 1 to 9, covering the field hyperspectral data wavelength range, were resampled, as detailed in Table 2.

Table 1. Pre-processed spectral reflectance data used in the species discrimination analysis.

Species code	Season	Wavelength range (340 – 820 nm)	Wavelength range (635 – 1110 nm)
LC	Dry	116	116
	Wet	210	212
NW	Dry	55	56
	Wet	75	75
CM	Dry	33	43
	Wet	38	39
OG	Dry	11	12
	Wet	12	13

Table 2. Spatial and spectral characteristics of resampled Sentinel-2 wavebands (B1 to B9). Source: (ESA, 2023)

Sentinel-2 band	Band name	Central wavelength (nm)	Bandwidth (nm)	Wavelength range (nm)	Spatial resolution (metres)
B1	Coastal aerosol	442.7	21	433–453	60
B2	Blue	492.4	66	458–523	10
B3	Green	559.8	36	543–578	10
B4	Red	664.6	31	650–680	10
B5	Vegetation Red edge 1	704.1	15	698–713	20
B6	Vegetation Red edge 2	740.5	15	733–748	20
B7	Vegetation Red edge 3	782.8	20	773–793	20
B8	NIR	832.8	106	785–900	10
B8A	Narrow NIR	864.7	21	855–875	20
B9	Water vapour	945.1	20	935–955	60
B10	SWIR - cirrus	1373.5	31	1360–1390	60
B11	SWIR1	1613.7	91	1565–1655	20
B12	SWIR2	2202.4	175	2100–2280	20

2.4. Computation of Multispectral Indices

Eighty-five (85) Sentinel-2 vegetation indices (VIs) published in the online index database (www.indexdatabase.de) were selected. They were based on the resampled Sentinel-2 wavebands and subsequently calculated for both seasons. In addition to these published indices, new multispectral indices were developed for species separability analysis. The newly developed indices were of four types, namely, simple ratio (SR), normalized difference ratio (ND), difference (D), and inverse difference (ID). These indices were calculated considering both the direct and inverse relationships (i.e. interchanging reflectance at a given wavelength, $R_{\lambda i,}$ with wavelength, $R_{\lambda j,}$ in the formulas). Formulas used to compute the new multispectral indices were adopted from Song and Wang (2022), as shown in equations 1–4 (R denotes the spectral reflectance at wavelengths, λ_i and λ_j).

$$SR_{i,j} = \frac{R_{\lambda i}}{R_{\lambda j}} \quad [1]$$

$$ND_{i,j} = \frac{R_{\lambda i} - R_{\lambda j}}{R_{\lambda i} + R_{\lambda j}} \quad [2]$$

$$D_{i,j} = R_{\lambda i} - R_{\lambda j} \quad [3]$$

$$ID_{i,j} = \frac{1}{R_{\lambda i}} - \frac{1}{R_{\lambda j}} \quad [4]$$

2.5. Statistical Analysis

Before analysing the resampled species spectral reflectance curves, we subjected the reflectance data to a normality test by using the Kolmogorov-Smirnov test (Berger and Zhou, 2014) in the R *"stats"* package (R Core Team, 2021). The test showed that the resampled spectral reflectance data did not follow a normal distribution ($p \leq 0.05$). Therefore, the Kruskal-Wallis H test (Kruskal and Wallis, 1952), a non-parametric test, was used to explore whether there were significant differences among the spectral reflectance curves of the species. We further performed a pairwise comparison of the reflectance curves of the paired species using the Wilcoxon rank sum test with continuity correction (*"BH"* method) to determine the pairs with significant differences.

2.6. Feature Selection

The feature selection spectral variables were grouped into six categories: resampled Sentinel-2 wavebands, published multispectral indices and new band combination indices (i.e. SR, ND, ID and D indices) for the dry and wet seasons. These grouped spectral variables were individually subjected to the feature selection process. The selected spectral variables per group were combined and used in the separability analysis of LC *vs* other species class pairs (i.e. LC vs NW, LC vs CM and LC vs OG). The feature selection procedure aimed to identify the most informative spectral variables that could discriminate LC among other species. The respective number of spectral variables used per group is presented in Table 3.

Table 3. The number of spectral variables subjected to the feature selection process.

Groups of spectral variables	Spectroradiometer wavelength range (340 – 820) nm	Spectroradiometer wavelength range (635 – 1100) nm
Resampled Sentinel-2 bands	n=8	n=7
Published Sentinel-2 multispectral indices	n=40	n=45
SR indices	n=56	n=42
ND indices	n=56	n=42
D indices	n=56	n=42
ID indices	n=56	n=42
Total	n = 272	n = 220

We used two popular feature selection methods, the Boruta algorithm (Kursa and Rudnicki, 2010), a wrapper method designed around the Random Forest (RF) algorithm, and the Guided Regularized Random Forest (GRRF) algorithm (Deng and Runger, 2013), an embedded approach. Both methods were implemented in R statistical software (R Core Team, 2021) and in the *"caret"* package (Kuhn, 2020). The GRRF algorithm was implemented in the R package, *"RRF",* (Deng, 2013) while the Boruta algorithm was implemented in the *"Boruta"* package (available at https://CRAN.R-project.org/package=Boruta). The advantage of the Random Forest method is that it can be used for

feature selection (Maxwell *et al*., 2018). Features selected through the Boruta method have been found to enhance land cover classification accuracies with multi-sourced, multi-sensor data (Duro *et al*., 2012). On the other hand, Izquierdo-Verdiguier and Zurita-Milla (2020 showed that substantial improvement in the accuracies of classification and regression models was achieved when GRRF-selected features were used instead of the ordinary RF-selected features. The GRRF method uses a coefficient of regularization parameter (*coefReg)* in the RRF algorithm to guide the feature selection process. The *coefReg* is obtained using two equations (equations 5 and 6).

$$imp = \frac{impRF}{\max(impRF)} \quad [5]$$

$$coefReg = (1 - gamma) + gamma \times imp \quad [6]$$

Equation 1 normalises the feature importance scores (*impRF*) obtained by means of the ordinary RF algorithm, while equation 2 gives a weighted average. The gamma values are user-defined and range between 0 and 1. Values close to 1 execute higher penalties, thereby leading to only a few feature selections, while values close to 0 lead to lower penalties and hence more features are selected. We chose a gamma value of eight (8) to obtain at least two (2) spectral features with the highest discriminatory power per spectral variable group, as indicated in Table 3.

2.7. Spectral Separability

A further step to quantify the potential of the Boruta and the GRRF-selected spectral features in separating species class pairs was taken by applying the Jeffries–Matusita (J-M) distance analysis. Several studies (Adam and Mutanga, 2009; Ouyang et al., 2013; Schmidt and Skidmore, 2003) have pointed to the importance of performing this step when determining species spectral separability. The J-M distance computations were done using the "*spectral.separability*" function in the "*spatialEco*" package (Evans and Murphy, 2021). The function gives J-M distance values ranging from 0 (lower bound) to ~1.4142 (√2) (upper bound) (Ouyang *et al*., 2013). This translates to 0% to 100% classification accuracies between the respective class pairs. Based on their contribution, the best-performing spectral features – as indicated by their J-M values − were identified and ordered from the highest to the lowest. Starting with the individual feature with the highest J-M value, additional features were added successively until a J-M distance value of ≥1.3718 (≥ 97% separability accuracy) was achieved. The ≥97% value was taken as the acceptable classification accuracy of our species class pairs (Adam and Mutanga, 2009).

2.8. Statistical Comparison of Spectral Variables selected by the GRRF and Boruta Methods

A statistical test was used to determine whether there were significant differences between the spectral features selected for species discrimination through the GRRF and Boruta methods. Before this analysis, these two sets of variables were subjected to a normality test using the Kolmogorov-Smirnov test (Berger and Zhou, 2014). They were found to be not normally distributed. Following this result, the Wilcoxon

rank sum test or the Mann-Whitney U test, a nonparametric test that compares values from two groups, was deemed appropriate since the datasets were not normally distributed (Taylor *et al*., 2012). Two sample Wilcoxon tests on rows were performed using the package *"matrixTests"* (Koncevicius, 2020) in R statistical software. The statistical differences between the two datasets i.e., the two sets of variables selected by the GRRF and Boruta methods, were assessed at a five percent (5%) significance level ($\alpha = 0.05$).

3. Results

3.1. Species Reflectance Curves

The mean spectral reflectance curves for the leaves of the study species are presented in Figure 2. The Kruskal-Wallis H test and the Wilcoxon rank sum test indicated that the OG reflectance curves significantly differed in bands 1, 2, 3, 6, 7, 8, 8A and 9 (Figure 2a). These channels may have the potential to discriminate OG from the rest of the species during the dry season. A comparison of the LC, CM and NW spectral curves for the dry season showed spectral similarity during the dry season. In contrast, the spectral reflectance curves for the species for the wet season indicated significant spectral differences between LC and the other species curves in bands 1 to 6. In particular, bands 4 and 5 showed significant differences between LC *vs* the CM, NW and OG class pairs. These results indicate the potential of the different spectral curve regions in discriminating LC from the other species. The level of separability of the species pairs was revealed in the computation of the J-M distance values.

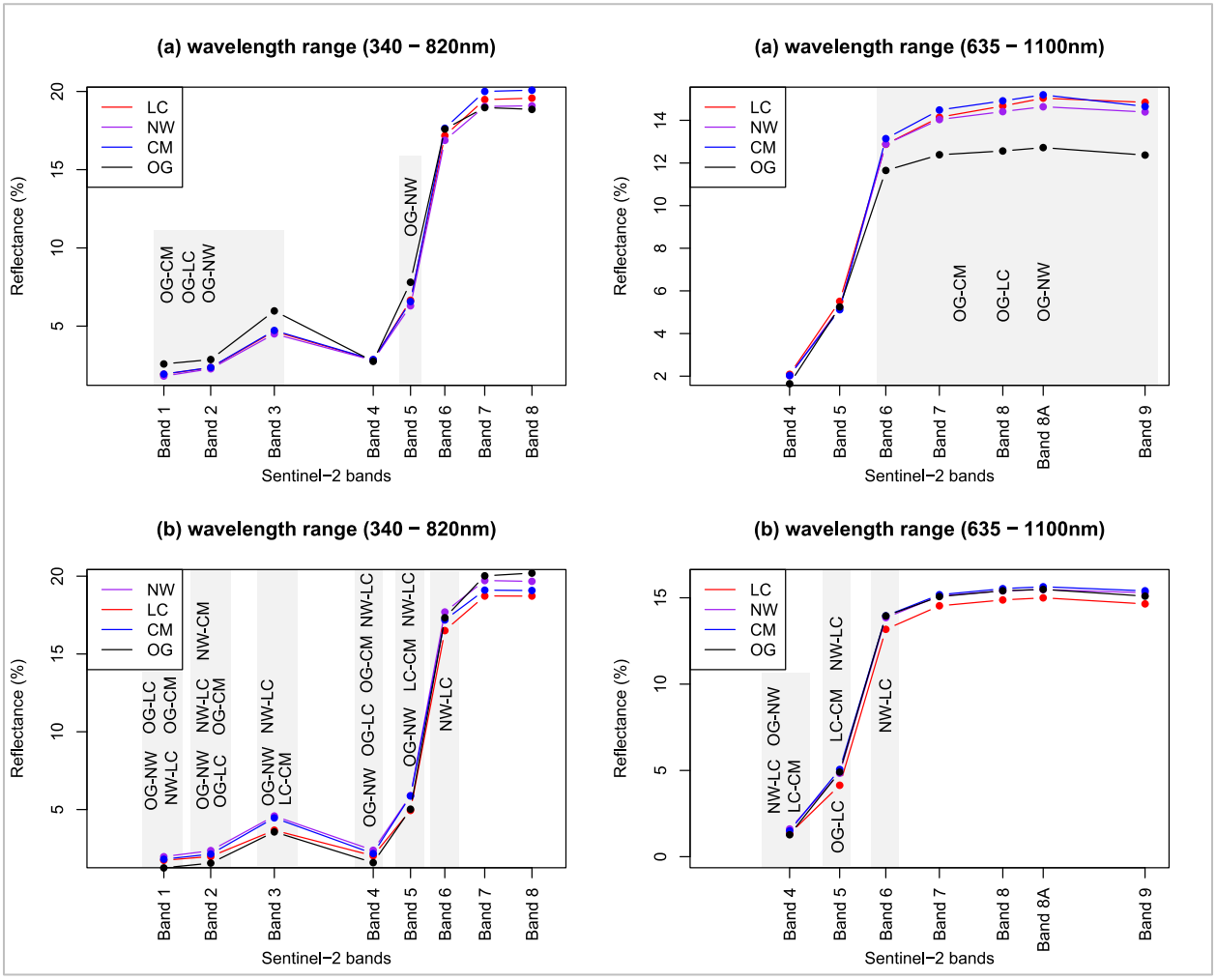

Figure 2. Resampled Sentinel-2 spectral reflectance curves for the dry season (a) and wet season (b). Bands highlighted in the grey background indicate significant spectral differences among the compared species pairs.

3.2. Feature Selection and Separability Analysis

The GRRF and Boruta methods yielded two sets of important spectral variables for species discrimination. Out of a total of *n=492* spectral features, the GRRF method selected considerably fewer spectral features per class pair than the Boruta method (Table 4).

Table 4. A summary of the number of selected spectral features for species class pair separation in the dry and wet seasons.

Species class pair		LC vs CM	LC vs NW	LC vs OG
GRRF method	Dry season	46	47	25
	Wet season	22	18	17
Boruta method	Dry season	158	86	252
	Wet season	186	213	134

Spectral separability analysis of the selected features using the J-M distance analysis produced fewer significant spectral features for species class pair separation. A set of spectral variables that produced acceptable class separability accuracies (≥97% of classification accuracy) is presented in Figure 3. These results reveal that, regardless of the feature selection method used, the selected features were unique for each of the species class pairs. Notably, relatively fewer spectral features that produced acceptable class separability accuracies were obtained from features selected by the Boruta method (LC *vs* CM (*n=12, n=12*), LC *vs* NW (*n=9, n=12*), and LC *vs* OG (*n=2, n=8*) for the dry and wet seasons, respectively) than those from the GRRF method (LC *vs* CM (*n=19, n=14*), LC *vs* NW (*n=15, n=14*), and LC *vs* OG (*n=2, n=4*) for the dry and wet seasons, respectively). In addition, the results showed that the SR, ND, ID and D indices constructed in this study dominated the list of spectral variables that give maximum LC class separability accuracy in both seasons. Several Sentinel-2 published indices, shown in Figure 3 and Table 5, featured among the selected spectral variables suitable for LC class separability in both seasons. The Sentinel-2 wavebands used in the construction of the new indices perfectly separating LC from the other species were distributed across the ten resampled Sentinel-2 wavebands (Bands 1 to 9) in both seasons, but apart from those separating the class pair, LC *vs* OG (Figure 4). The NIR waveband appeared to be most important in separating the LC *vs* CM and LC *vs* NW class pairs in the dry season, while the red, green and red-edge waveband regions were important for the separation of the LC *vs* OG class pair. During the wet season, most of the spectral indices chosen for the separation of the LC *vs* CM pair were constructed with Sentinel-2 bands 5 (Red edge 1) and 4 (Red), while indices giving perfect separation of the LC *vs* NW pair were constructed with bands 5 and 3 (Green). The separation of LC *vs* OG in the wet season consisted of indices constructed with bands 6 (Red edge 2) and 7 (Red edge 3).

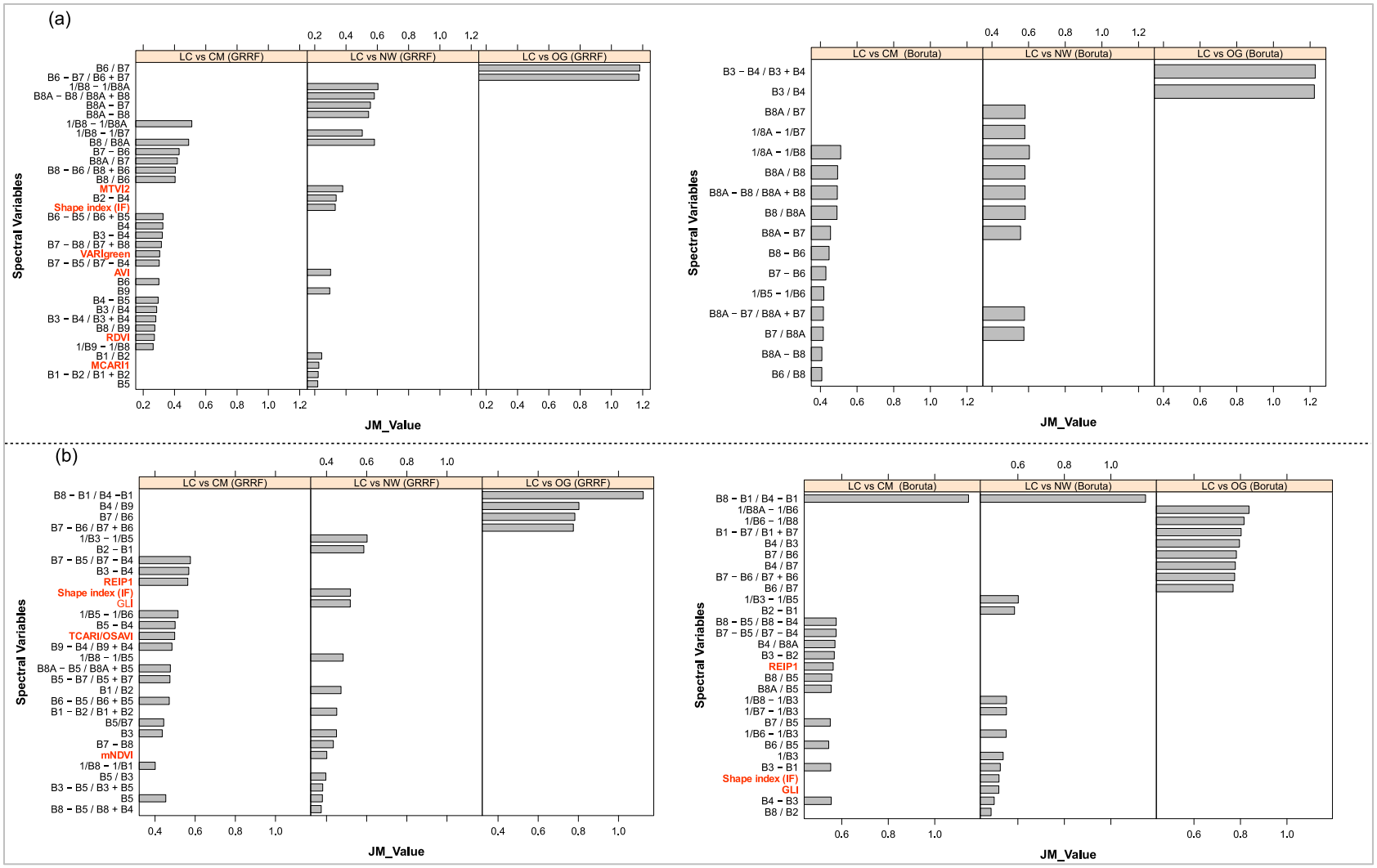

Figure 3. Sets of Boruta and GRRF selected spectral variables that produced ≥97% separability accuracy for the LC class pairs. The set of features for the dry season is shown in (a) and for the wet season in (b). Highlighted features in red font are the selected published Sentinel-2 indices. B_i in the spectral variables denotes the Sentinel-2 band number.

Table 5. The names and formulas of published Sentinel-2 multispectral indices selected by the GRRF and Boruta methods for species class pair separation per season. B_i denotes the Sentinel-2 band number.

S/N	Index name	Formula	Class pair	Season	Method
1	Anthocyanin reflectance index (ARI)	$1/B3 - 1/B5$	LC vs NW	Wet	Boruta
2	Ashburn Vegetation Index (AVI)	$2.0 \times B9 - B4$	LC vs NW	Dry	GRRF
3	CRI700 (Datt1)	$B8 - B5 \,/\, B8 - B4$	LC vs CM	Wet	Boruta
4	Green leaf index (GLI)	$(2 \times B3 - B5 - B1) \,/\, (2 \times B3 + B5 + B1)$	LC vs NW	Wet	GRRF/Boruta
5	Inverse reflectance 550 (IR550)	$1/B3$	LC vs NW	Wet	Boruta
6	Leaf Chlorophyll Index (LCI)	$B8 - B5 \,/\, B8 + B4$	LC vs NW	Wet	GRRF
7	Maccioni	$B7 - B5 \,/\, B7 - B4$	LC vs CM	Dry/Wet	GRRF/Boruta
8	Modified NDVI (mNDVI)	$(B8 - B4) \,/\, (B8 + B4 - (2 \times B1))$	LC vs NW	Wet	GRRF
9	Modified Chlorophyll Absorption in Reflectance Index 1 (mCARI1)	$1.2 \times (2.5 \times (B8 - B4) - 1.3 \times (B8 - B3))$	LC vs NW	Dry	GRRF
10	Modified Simple Ratio (mSR)	$B8 - B1 \,/\, B4 - B1$	LC vs OG	Wet	GRRF
			LC vs CM	Wet	Boruta
			LC vs NW	Wet	Boruta
11	Modified Triangular Vegetation Index 2 (mTVI2)	$1.5 \times ((1.2 \times (B8 - B3) - 2.5 \times (B4 - B3))/ ((2 \times B8 + 1)\ \hat{}2 - 6 \times B8 - 5 \times (B4)\ \hat{}0.5)\ \hat{}0.5)$	LC vs NW	Dry	GRRF
12	Normalized Difference 550/650 Photosynthetic vigour ratio (PVR)	$B3 - B4 \,/\, B3 + B4$	LC vs OG	Dry	Boruta
13	Pigment-specific simple ratio C2 (PSSRc2)	$B8 \,/\, B2$	LC vs NW	Wet	Boruta
14	Renormalized Difference Vegetation Index (RDVI)	$(B8 - B4) \,/\, (B8 + B4)\ \hat{}0.5$	LC vs CM	Dry	GRRF
15	Red-Edge Inflection Point 1 (REIP1)	$700 + 40 \times ((((B4 + B7)/2) - B5) \,/\, (B6 - B5))$	LC vs CM	Wet	GRRF/Boruta
16	Shape index (IF)	$(2 \times B5 - B3 - B1) \,/\, (B3 - B1)$	LC vs NW	Dry/Wet	GRRF/Boruta
17	Simple Ratio 550/670	$B3 \,/\, B4$	LC vs OG	Dry	Boruta
18	Simple Ratio 735/710	$B6 \,/\, B5$	LC vs CM	Wet	Boruta
19	Simple Ratio 850/710 (Datt2)	$B8 \,/\, B5$	LC vs CM	Wet	Boruta
20	Simple Ratio 860/708	$B8A \,/\, B5$	LC vs CM	Wet	Boruta
21	TCARI/OSAVI	$3 \times (B5 - B4) - (0.2 \times (B5 - B3) \times (B5/B4)) \,/\, (1 + 0.16) \times ((B8 - B4)/ (B8 + B4 + 0.16))$	LC vs CM	Wet	GRRF
22	Visible Atmospherically Resistant Index Green (VARIgreen)	$(B3 - B4) \,/\, (B3 + B4 - B2)$	LC vs CM	Dry	GRRF

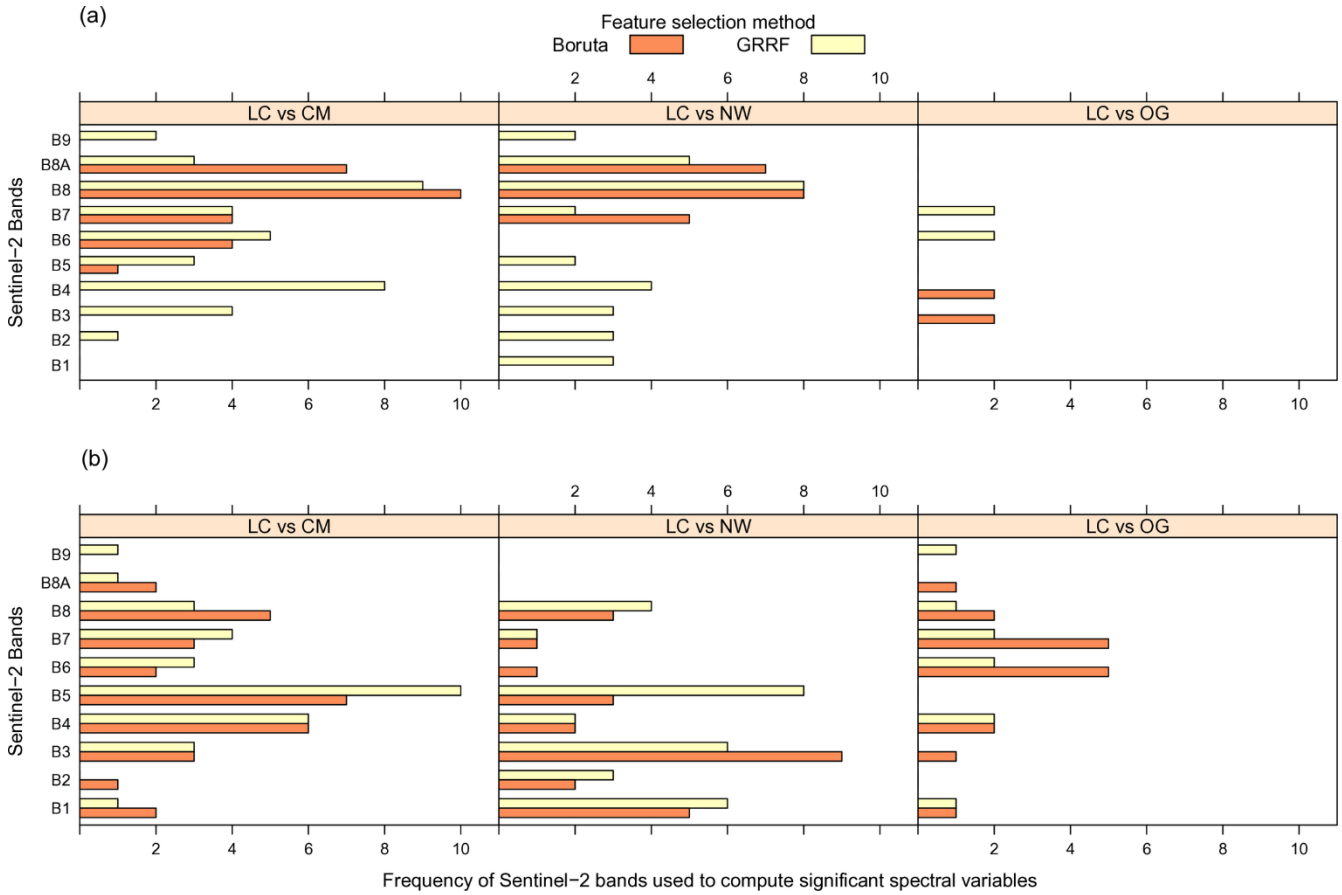

Figure 4. The frequency of the selected Sentinel-2 wavebands that were used to construct new spectral indices for the separation of species class pairs for the dry season (a) and wet season (b).

3.3. Significance Test between the GRRF and Boruta-selected Spectral Variables

The results of the Wilcoxon rank sum test presented in Figure 5 show that the p-values were greater than 0.05 ($p>0.05$), thereby indicating that the two sets of spectral variables did not present with significant differences in their performance. This suggests that despite the differences in the sets of individually selected spectral variables through the two methods, each set contained spectrally significant variables that perfectly separated the LC class pairs.

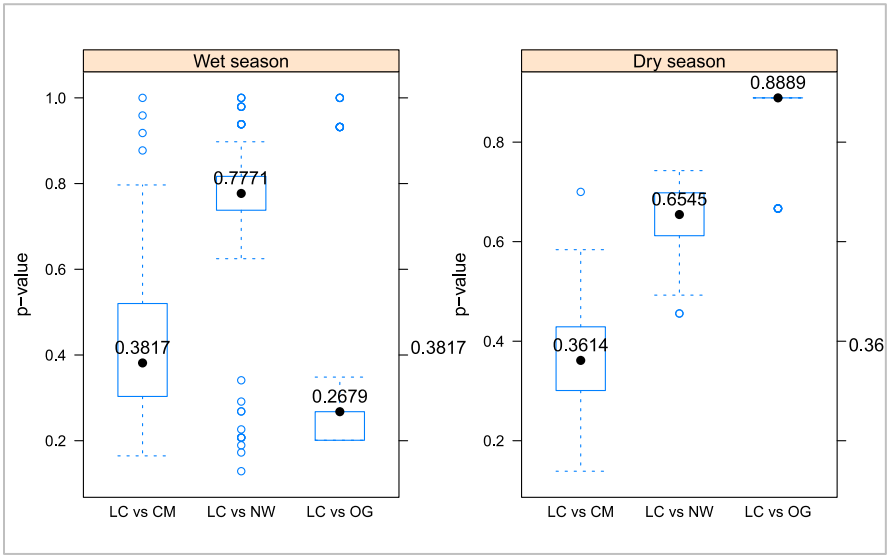

Figure 5. Boxplot of p-values obtained from a comparison of the selected GRRF and Boruta method spectral variables for LC class pair separation.

4. Discussion

Mapping alien invasive plant species such as the LC in natural environments is crucial to the management of the species. The successful mapping of LC using remote sensing image data may pose a challenge when useful spectral features that would enhance its discrimination are unknown. This study sought to identify Sentinel-2 multispectral variables that can discriminate LC from co-occurring vegetation. The results from this study have demonstrated that LC could be discriminated from co-occurring vegetation by using a set of significant Sentinel-2 spectral variables selected through feature selection and a separability analysis strategy.

The initial exploratory analysis of the LC, CM, NW and OG spectral reflectance curves indicated visual similarities in some Sentinel-2 wavelength regions. Further statistical analysis performed in this study revealed potential spectral distinctions among the species curves. This study has demonstrated that using feature selection methods such as GRRF and Boruta and performing class pair spectral separability analysis provides an optimal set of spectrally significant variables for species class pair discrimination. This concurs with other works, such as Ouyang *et al*. (2013); Mureriwa *et al*. (2016); Mudereri *et al*. (2020), which point out that successful discrimination among vegetation covers is achievable through feature selection and separability analysis. Results of this study highlight the importance of identifying spectrally unique features to differentiate among vegetation covers in satellite images which could be particularly useful in the management of invasive species like LC.

One way to reduce data dimensionality, redundancy and the extraction of unique spectral information for species discrimination is through vegetation indices (Thenkabail *et al*., 2013). Vegetation indices, especially the band combination indices presented in this study, were largely selected as significant

spectral variables for LC discrimination. This suggests that vegetation indices provide useful spectral information for LC class pair discrimination as opposed to the individual Sentinel-2 spectral bands and published indices. This concurs with other studies such as one by Rajah *et al*. (2019), where high classification accuracies (~80%) of *Rubus cuneifolius* were obtained. In this case, Sentinel-2 vegetation indices were used. Notably, Sentinel-2 bands in the red, in the red-edge, and leading up into the near-infrared regions, were the main ones the that were selected in our study as significant in terms of their spectral features. These regions have also been termed crucial in other studies (e.g.Mureriwa *et al*., 2016) in that they discriminate among vegetation covers. In particular, the red-edge region is sensitive to leaf spectral variations brought about by the structural characteristics, pigmentation, water content and size of the leavges, and are, therefore, valuable in species discrimination (Odindi *et al.*, 2016).

Although indices of type normalized difference (ND) (e.g., the NDVI) take advantage of the reflectance contrast between bands in the NIR range and other bands (Ouyang *et al*., 2013), the construction of such indices may involve bands other than the NIR for enhancing spectral information in a given species. For instance, the selected ND indices (B6 – B5 / B6 + B5) and (B3 – B4 / B3 + B4) among those enhancing separability of the LC *vs* CM class pair in the dry season make use of the red-edge bands (B5 and B6) and the red (B4) and green (B3) bands, respectively. The importance of blue, green, red-edge 2, red-edge 3 and NIR Sentinel-2 wavebands in discriminating LC have also been reported by Dube *et al.* (2020). In addition, the results of this study indicate that the SR, D and ID indices significantly contribute to the enhancement of the spectral separability of LC from other species.

Identifying the unique spectral characteristics of a given species facilitates its successful discrimination from other vegetation cover types. However, getting such information from multispectral sensors may prove difficult on account of the sensor's inability to record pure signals of individual plants in a given location. These sensors often give mixed pixels (a mixture of signals from different plants), thereby throwing the whole exercise of image classification to some level of uncertainty (Huang and Asner, 2009; Royimani et al., 2019). Nevertheless, several researchers have explored the capabilities of such datasets, especially those from new-generation sensors (e.g., Sentinel-2 and Landsat 8) in mapping LC in various habitats. For instance, (Dube *et al*. (2020) reported that Sentinel-2 imagery datasets detected LC from other landcovers in a semiarid rangeland ecosystem with an overall accuracy of ~78% as opposed to ~65% obtained with Landsat 8. In that study, the Sentinel-2 derived indices, especially the red-edge-derived Normalized Difference Vegetation Index (NDVI), were found to improve the LC classification compared to the Landsat 8 indices. This concurs with the capabilities of the ND Sentinel-2 indices derived in our study to detect LC in natural habitats.

The use of hyperspectral datasets, on the other hand, easily allows for the identification of the unique spectral characteristics of species owing to the high spectral resolution inherent in the datasets. However, these datasets are usually associated with several challenges, including "the curse of dimensionality"

inherent in them, thereby requiring large-scale training samples for reliable image classification (Thenkabail *et al*., 2013) and high costs of accessing the imagery (Royimani *et al*., 2019). Feature selection procedures, such as those employed in this research, are useful in addressing the challenge of dimensionality. However, the high cost of hyperspectral imagery data acquisition may render long-term monitoring of invasive species with such datasets impossible for developing nations (Royimani et al., 2019). Recently, field spectroscopy has gained interest among the scientific community on account of the availability of relatively "low-cost" handheld spectroradiometers such as the ones used in this study. This has allowed for studies of the spectral characteristics of vegetation and species-level spectral separability analysis. The ultimate goal has been to enhance the monitoring of the vegetation cover on a landscape scale (e.g., the LC invasion in forest habitats). In addition, the optimal period for species discrimination can be determined through analyses of seasonal hyperspectral datasets, as demonstrated in studies such as Ouyang *et al*. (2013). This present study obtained seasonal datasets (i.e., during the dry and wet seasons). Although plant phenology (variability) is influenced by seasonality, the results of this study show no preference for any given season over the others in the context of the separability of LC from its co-occurring species. A set of spectral variables has been selected for LC discrimination in both seasons. These variables may be used in conjunction with Sentinel-2 data on a landscape scale in the fractional cover mapping of the Muringato forest area. Future work could also investigate the applicability of these indices in mapping LC in other areas outside the Muringato forest.

5. Conclusions

This study sought to identify spectrally significant Sentinel-2 spectral variables that could be used in LC discrimination. The findings of this study have shown that feature selection using the GRRF and Boruta methods, and separability analysis using the J-M distance method, can provide for spectrally significant variables for LC class pair discrimination in wet and dry seasons. It has been found that although some of the available published Sentinel-2 multispectral indices may enhance LC discrimination, the use of band combination indices of types ND, SR, D, and ID could significantly reveal distinct spectral information important in discriminating LC from other species. This may suggest the need for an exploratory analysis of band combination indices of the types mentioned above in conjunction with the published indices to extract those that would enhance LC discrimination. By doing this, processing times for image classification would be lowered and the detection of LC from Sentinel-2 images would be improved. This study also confirms the capability of both the GRRF and the Boruta method as equally important in selecting a set of the most informative spectral features for LC discrimination in both seasons. The spectral features identified in this study could be used for a landscape-level classification of LC in the entire Muringato forest with Sentinel-2 imagery datasets. The processing chains presented in this study could also be used for other invasive species. It is envisaged

that the creation of regular and reliable maps of LC cover using satellite imagery data and the appropriate spectral variables would enhance monitoring programmes focused on invasive species. This will further aid conservation managers in making informed decisions on conservation actions to deal with invasive species, especially in forested areas.

6. Acknowledgements

The authors would like to express their gratitude to the Kenya Forest Service and the Muringato Forest station manager for their support during this work. This work was part of the work funded by the International Foundation for Science (IFS), Grant/ Award Number: 1-1-D-6489-1, 2020.

7. References

Adam, E., Mutanga, O., 2009. Spectral discrimination of papyrus vegetation (*Cyperus papyrus* L.) in swamp wetlands using field spectrometry. ISPRS Journal of Photogrammetry and Remote Sensing 64, 612–620. https://doi.org/10.1016/j.isprsjprs.2009.04.004

Berger, V.W., Zhou, Y., 2014. Kolmogorov–Smirnov Test: Overview, in: Wiley StatsRef: Statistics Reference Online. Wiley, pp. 1–5. https://doi.org/10.1002/9781118445112.stat06558

Chauhan, H., Mohan, B.K., 2014. Effectiveness of Spectral Similarity Measures to develop Precise Crop Spectra for Hyperspectral Data Analysis. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences II–8, 83–90. https://doi.org/10.5194/isprsannals-II-8-83-2014

Deng, H., 2013. Guided Random Forest in the RRF Package 1–2.

Deng, H., Runger, G., 2013. Gene selection with guided regularized random forest. Pattern Recognition 46, 3483–3489. https://doi.org/10.1016/j.patcog.2013.05.018

Dube, T., Shoko, C., Sibanda, M., Madileng, P., Maluleke, X.G., Mokwatedi, V.R., Tibane, L., Tshebesebe, T., 2020. Remote Sensing of Invasive *Lantana camara* (*Verbenaceae*) in Semiarid Savanna Rangeland Ecosystems of South Africa. Rangeland Ecology & Management 73, 411–419. https://doi.org/10.1016/j.rama.2020.01.003

Duro, D.C., Franklin, S.E., Dubé, M.G., 2012. Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests. International Journal of Remote Sensing 33, 4502–4526. https://doi.org/10.1080/01431161.2011.649864

ESA, 2023. Sentinel-2 - Missions - Resolution and Swath - Sentinel Handbook - Sentinel Online (esa.int) [WWW Document]. URL https://sentinel.esa.int/web/Sentinel/missions/Sentinel-2/instrument-payload/resolution-and-swath (Date accessed: 27 September 2022).

Evans, J.S., Murphy, M.A., 2021. spatialEco.

Fang, H., Liang, S., 2014. Leaf Area Index Models, Reference Module in Earth Systems and Environmental Sciences. Elsevier Inc. https://doi.org/10.1016/B978-0-12-409548-9.09076-X

Global Invasive Species Database, 2023. 100 of the World's Worst Invasive Alien Species [WWW Document]. URL http://www.iucngisd.org/gisd/100_worst.php (Date accessed: 23 June 2023).

Goncalves, E., Herrera, I., Duarte, M., Bustamante, R.O., Lampo, M., Velásquez, G., Sharma, G.P., García-Rangel, S., 2014. Global invasion of *Lantana camara*: Has the climatic niche been conserved across continents? PLoS ONE 9. https://doi.org/10.1371/journal.pone.0111468

Government of Kenya, 2018. Nyeri County Integrated Development Plan 2018-2022.

Große-Stoltenberg, A., Hellmann, C., Werner, C., Oldeland, J., Thiele, J., 2016. Evaluation of continuous VNIR-SWIR spectra *versus* narrowband hyperspectral indices to discriminate the invasive *Acacia longifolia* within a Mediterranean dune ecosystem. Remote Sensing (Basel) 8, 334. https://doi.org/10.3390/rs8040334

Hennessy, A., Clarke, K., Lewis, M., 2020. Hyperspectral Classification of Plants: a Review of Waveband Selection Generalisability. Remote Sensing 12, 113. https://doi.org/10.3390/rs12010113

Huang, C., Asner, G.P., 2009. Applications of remote sensing to alien invasive plant studies. Sensors (Switzerland) 9, 4869–4889. https://doi.org/10.3390/s90604869

Izquierdo-Verdiguier, E., Zurita-Milla, R., 2020. An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. International Journal of Applied Earth Observation and Geoinformation 88, 102051. https://doi.org/10.1016/j.jag.2020.102051

Kenya Forest Service, 2010. Aberdare forest reserve management plan [WWW Document]. URL http://www.kenyaforestservice.org/documents/Aberdare.pdf (Date accessed: 3 April 2020).

Kenya Wildlife Service, 2010. Mt Kenya Ecosystem Management Plan, 2010-2020 [WWW Document]. URL www.kws.go.ke/sites/default/files/parksresorces:/Mt. Kenya Ecosystem Management Plan (2010-2020).pdf (Date accessed: 3 April 2020).

Kimothi, M.M., Dasari, A., 2010. Methodology to map the spread of an invasive plant (*Lantana camara* L.) in forest ecosystems using Indian remote sensing satellite data. International Journal of Remote Sensing 31, 3273–3289. https://doi.org/10.1080/01431160903121126

Koncevicius, K., 2020. matrixTests: Fast Statistical Hypothesis Tests on Rows and Columns of Matrices.

Kruskal, W.H., Wallis, W.A., 1952. Use of Ranks in One-criterion Variance Analysis. Journal of the American Statistical Association 47, 583. https://doi.org/10.2307/2280779

Kuhn, M., 2020. Caret: Classification and Regression Training.

Kursa, M.B., Rudnicki, W.R., 2010. Feature Selection with the Boruta Package. Journal of Statistical Software 36, 1–13. https://doi.org/10.18637/jss.v036.i11

Lehnert, L.W., Meyer, H., Obermeier, W.A., Silva, B., Regeling, B., Bendix, J., 2019. Hyperspectral Data Analysis in R: The Hsdar Package. Journal of Statistical Software 89. https://doi.org/10.18637/jss.v089.i12

Mahlein, A.K., Rumpf, T., Welke, P., Dehne, H.W., Plümer, L., Steiner, U., Oerke, E.C., 2013. Development of spectral indices for detecting and identifying plant diseases. Remote Sensing of Environment 128, 21–30. https://doi.org/10.1016/j.rse.2012.09.019

Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: an applied review. International Journal of Remote Sensing 39, 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

MoALF, 2016. Climate Risk Profile for Nyeri. Kenya County Climate Risk Profile Series. Nairobi, Kenya.

Mudereri, B.T., Dube, T., Niassy, S., Kimathi, E., Landmann, T., Khan, Z., Abdel-Rahman, E.M., 2020. Is it possible to discern Striga weed (*Striga hermonthica)* infestation levels in maize agro-ecological systems using *in-situ* spectroscopy? International Journal of Applied Earth Observation and Geoinformation 85, 102008. https://doi.org/10.1016/j.jag.2019.102008

Mureriwa, N., Adam, E., Sahu, A., Tesfamichael, S., 2016. Examining the Spectral Separability of *Prosopis glandulosa* from Co-existent Species using Field Spectral Measurement and Guided Regularized Random Forest. Remote Sensing (Basel) 8, 144. https://doi.org/10.3390/rs8020144

Negi, G.C.S., Sharma, S., Vishvakarma, S.C.R., Samant, S.S., Maikhuri, R.K., Prasad, R.C., Palni, L.M.S., 2019. Ecology and Use of *Lantana camara* in India. Botanical Review 85, 109–130. https://doi.org/10.1007/s12229-019-09209-8

Odindi, J., Mutanga, O., Rouget, M., Hlanguza, N., 2016. Mapping alien and indigenous vegetation in the KwaZulu-Natal Sandstone Sourveld using remotely sensed data. Bothalia - African Biodiversity & Conservation 1–9. http://dx.doi.org/10.4102/abc.v46i2.2103

Ouyang, Z.T., Gao, Y., Xie, X., Guo, H.Q., Zhang, T.T., Zhao, B., 2013. Spectral Discrimination of the Invasive Plant *Spartina alterniflora* at Multiple Phenological Stages in a Saltmarsh Wetland. PLoS One 8, 1–12. https://doi.org/10.1371/journal.pone.0067315

R Core Team, 2021. R: A Language and Environment for Statistical Computing.

Rajah, P., Odindi, J., Mutanga, O., Kiala, Z., 2019. The utility of Sentinel-2 Vegetation Indices (VIs) and Sentinel-1 Synthetic Aperture Radar (SAR) for invasive alien species detection and mapping. Nature Conservation 35, 41–61. https://doi.org/10.3897/natureconservation.35.29588

Royimani, L., Mutanga, O., Odindi, J., Dube, T., Matongera, T.N., 2019. Advancements in satellite remote sensing for mapping and monitoring of alien invasive plant species (AIPs). Physics and Chemistry of the Earth, Parts A/B/C 112, 237–245. https://doi.org/10.1016/j.pce.2018.12.004

Schmidt, K.S., Skidmore, A.K., 2003. Spectral discrimination of vegetation types in a coastal wetland. Remote Sensing of Environment 85, 92–108. https://doi.org/10.1016/S0034-4257(02)00196-7

Shackleton, R.T., Witt, A.B., Aool, W., Pratt, C.F., 2017. Distribution of the invasive alien weed, *Lantana camara*, and its ecological and livelihood impacts in eastern Africa. African Journal of Range & Forage Science 34, 1–11. https://doi.org/10.2989/10220119.2017.1301551

Sharma, O., Makkar, H.P.S., Dawra, R.K., Negi, S.S., 1981. A Review of the Toxicity of *Lantana camara* (Linn) in Animals. Clinical Toxicology 18, 1077–1094. https://doi.org/10.3109/15563658108990337

Song, G., Wang, Q., 2022. Developing Hyperspectral Indices for assessing Seasonal Variations in the Ratio of Chlorophyll to Carotenoid in Deciduous Forests. Remote Sensing 14. https://doi.org/10.3390/rs14061324

Taylor, S., Kumar, L., Reid, N., Lewis, C.R.G., 2012. Optimal band selection from hyperspectral data for *Lantana camara* discrimination. International Journal of Remote Sensing 33, 5418–5437. https://doi.org/10.1080/01431161.2012.661093

Thenkabail, P.S., Mariotto, I., Gumma, M.K., Middleton, E.M., Landis, D.R., Huemmrich, K.F., 2013. Selection of Hyperspectral Narrowbands (HNBs) and Composition of Hyperspectral Two-band Vegetation Indices (HVIs) for Biophysical Characterization and Discrimination of Crop Types using Field Reflectance and Hyperion/EO-1 Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 427–439. https://doi.org/10.1109/JSTARS.2013.2252601

Transon, J., d'Andrimont, R., Maugnard, A., Defourny, P., 2018. Survey of hyperspectral Earth observation applications from space in the Sentinel-2 context. Remote Sensing 10, 1–32. https://doi.org/10.3390/rs10020157

Waititu, J.M., Mundia, C.N., Sichangi, A.W., 2022. Assessing distribution changes of selected native and alien invasive plant species under changing climatic conditions in Nyeri County, Kenya. PLoS ONE 17, e0275360. https://doi.org/10.1371/journal.pone.0275360

Witt, A., Beale, T., van Wilgen, B.W., 2018. An assessment of the distribution and potential ecological impacts of invasive alien plant species in eastern Africa. Transactions of the Royal Society of South Africa 73, 217–236. https://doi.org/10.1080/0035919X.2018.1529003

Xue, J., Su, B., 2017. Significant remote sensing vegetation indices: A review of developments and applications. Journal of Sensors 2017. https://doi.org/10.1155/2017/1353691