Machine learning algorithms for mapping *Prosopis glandulosa* and land cover change using multi-temporal Landsat products: a case study of Prieska in the Northern Cape Province, South Africa

Colette de Villiers[1], Cilence Munghemezulu[1,2], George Chirima[1,2], Philemon Tsele[1] and Zinhle Mashaba-Munghemezulu[1,2]

[1]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa email: colettedev007@gmail.com
[2]Agricultural Research Council-Institute for Soil, Climate and Water (ARC-ISCW), Division of Geoinformation Science, Pretoria, South Africa

DOI: http://dx.doi.org/10.4314/sajg.v9i2.13

Abstract

Invasive alien plants (IAPs) are responsible for loss in biodiversity and the depletion of water resources in natural ecosystems. Prosopis species are IAPs previously introduced by farmers to provide shade and fodder for livestock. In the Northern Cape, Prosopis spp. invasions are associated with the loss of native species resulting in overgrazing and degrading rangelands. Mapping Prosopis glandulosa is essential for management initiatives to assist the government in minimising the spread and impact of IAPs. This study aims to evaluate the performance of two machine learning algorithms i.e., Support Vector Machine (SVM) and Random Forest (RF) to map the spatial dynamics of P. glandulosa in Prieska. The spatial invasion extent of P. glandulosa was mapped using multitemporal Landsat data spanning the period from 1990 to 2018. Validation of the results was done through an estimated error matrix with the use of the proportion of area and the estimates of overall accuracy, user's accuracy and producer's accuracy with a 95% confidence interval. The performance of the SVM and RF classifiers showed similar results in the overall accuracy and Kappa statistics throughout the years. These methods showed an overall increase of at least 3.3% of the area invaded by P. glandulosa from 1990 to 2018. The study indicates the importance of Landsat imagery for mapping historical and current land cover change of IAPs. The spread of P. glandulosa was confirmed by an increase in the total area of invasion, which enables decision-makers to improve monitoring and eradication initiatives.

Keywords: *Invasive alien plants, Prosopis glandulosa, Machine learning, Landsat data, Random Forest, Support Vector Machine*

1. Introduction

Invasive alien plants (IAPs) are non-native taxa introduced by human-related activities (accidentally or intentionally) that spread throughout large geographic regions while invading natural ecosystems (Richardson et al., 2000). The IAPs cause a decline in biodiversity and indigenous ecosystems, they are a continuous threat to environments because they outcompete native vegetation (Van Wilgen and Richardson, 2014). Additionally, the impacts of IAPs result in the loss of plant species richness and species composition of indigenous vegetation (Gaggini et al., 2019, Hejda et al., 2009). Ultimately, invasive species are responsible for the economic and environmental impacts that negatively affect human welfare (Beck et al., 2008). The economy is at risk when the pressures of invasive species on the agricultural industry threaten human well-being and livelihood (Liu and Piper, 2016).

The *Prosopis* spp. (Fabaceae) is native to North and South America and was intentionally introduced to South Africa by farmers to provide shading and fodder for livestock (Zimmermann, 1991). The rapid dispersal of *Prosopis* spp. seeds was caused primarily by animals and flooding events and this resulted in widespread formation of dense stands (Zachariades et al., 2011). Large-scale invasions were responsible for the loss of native vegetation, which caused overgrazing and degrading rangelands where very poor grazing capacity was observed (Ndhlovu et al., 2011). Shackleton et al. (2015a) observed that *Prosopis* spp. thickets seem to flourish in areas with high water availability throughout South Africa. Widespread invasion of ephemeral river systems by the IAP is responsible for large-scale depletion of groundwater (Van Wilgen et al., 2007). According to the Department of Water Affairs (2013), IAPs are responsible for the 695 million m^3 of annual reduction in the water supply of dams and river systems in South Africa. In South Africa, Van Wilgen et al. (2008) identified *Prosopis* spp. as a high risk to water catchment areas in the Nama and Succulent Karoo. Water supplies are continuously at risk in semi-arid environments such as in the Northern Cape, where there is an increased abundance of *Prosopis* spp. invasions. The study by Dzikiti et al. (2013), found that clearing of *Prosopis* spp. would result in significant conservation of groundwater supplies throughout the province.

Selecting a suitable classification approach is important to produce reliable maps for *P. glandulosa* and other land cover types. Various classification methods have been used to map *Prosopis* spp., including pixel-based (Mirik and Ansley, 2012), object-based, rule-based (Laliberte et al., 2012) and machine learning classifications (Ku and Popescu, 2019). Random Forest (RF) and Support Vector Machine (SVM) classifiers are the two major machine learning algorithms used in the land cover mapping of intricate landscapes (Adam et al., 2014, Zhang et al., 2017, Li et al., 2016). For example, in Somaliland the mapping of *Prosopis* spp. in the same study area produced more favourable results with the RF classifier (Meroni et al., 2017) compared to the initial study using the supervised Maximum Likelihood (ML) classifier (Rembold et al., 2015). Machine learning algorithms reduce the occurrence of class misclassification when identifying *Prosopis* spp. from other land covers during classification. For example, Landsat 8 Operational Land Imager (OLI) satellite imagery was classified with the object-oriented approach which produced considerably more misclassifications between land cover classes than the RF pixel-based approach in Somaliland (Ng et al., 2016).

In South Africa, machine learning algorithms have seldom been studied in the mapping of the spatial extent of the IAP *P. glandulosa*. In the study by Adam et al. (2017), they mapped *P. glandulosa* with the RF and SVM classifiers using WorldView-2 datasets. Both classifications achieved high accuracy classification results to distinguish *P. glandulosa* from other native vegetation such as *A. mellifera* and *A. karoo.* Evaluating the effectiveness of mapping *P. glandulosa* with different machine learning algorithms needs statistical analysis to determine the best performing classifier. Several statistical methods have been used to assess the performance of different classifiers in remote sensing studies, for example, the Kappa statistics and McNemar's test (Li et al., 2013, Rodriguez-Galiano and Chica-Rivas, 2014).

Mapping the extent of IAPs is essential to improving the understanding of management initiatives in terms of planning and implementation (Shackleton et al., 2014). Mapping techniques help improve planning procedures to prioritise the areas identified for the clearing of IAPs (Van Wilgen et al., 2007, Rouget et al., 2004). Van Wilgen et al. (2012) indicated that only 4% of *Prosopis* spp. in the estimated total invaded area throughout South Africa was cleared in all the arid biomes despite 435 million Rands spent. For management initiatives, it is crucial to understand the spatial dynamics of *Prosopis* spp. over time. Amboka and Ngigi (2015) mapped *Prosopis* spp. over a time-series using multi-temporal Landsat Thematic Mapper (TM) and Landsat Enhanced Thematic Mapper Plus (ETM+) imagery. This study mapped the change of *Prosopis* spp. invasion from 1985 to 2010 in Kenya. Van den Berg et al. (2013) is one of the few long-term studies of *Prosopis* spp. invasion monitoring in the Northern Cape, South Africa. The study used terrain analysis and remote sensing techniques to classify Landsat datasets that produced time series maps of the distribution of *Prosopis* spp. for 30 years from 1970 to 2007.

The continuous spread of *P. glandulosa* is a serious problem for management initiatives in South Africa. Mapping techniques are improving control programs that inform managers on current distribution patterns. This study aimed to accurately map the extent of the *P. glandulosa* with multi-temporal Landsat imagery for the period 1990 to 2018 in Prieska, Northern Cape. The main objective of this study was to map the IAP *P. glandulosa* and other land cover types with two machine learning algorithms using Landsat imagery. Two machine learning algorithms were evaluated by comparing the performance of the SVM and RF classifiers. The results were used to assess land cover change detection using multi-temporal Landsat imagery for 28 years at 5-year intervals (1990, 1997, 2005, 2013 and 2018). Mapping of the invasion and change of land cover in the study area can inform decision-makers on the risk of future invasions and where to focus current control initiatives.

2. Materials and methods

2.1. Study area

The study area (Figure 1) is located around the town of Prieska within the Siyathemba local municipality of the Northern Cape Province of South Africa. Prieska is located at 29°41'25.7"S latitude and 22°44'26.8"E longitude in the eastern part of the province. This region is distinguished by a semi-arid climate with low annual rainfall. The Blaaukrans weather station in Prieska recorded an average annual rainfall of 270 mm from 1996 to 2018. The study area is part of the Nama-Karoo the largest biome in South Africa with vast dwarf shrubland vegetation. Indigenous species (i.e.

Ziziphus mucronata, *Acacia erioloba,* and *Tamarix usneoides)* in the region form dense thickets found next to dry river banks (Van den Berg, 2010). The town of Prieska was built along the Orange River, known as the longest-running river system in South Africa. Farmers in the area depend on the river as a source of water for crop irrigation. The river ecosystem is very vulnerable to the loss of native species with the increasing spread of the IAP *P. glandulosa*.

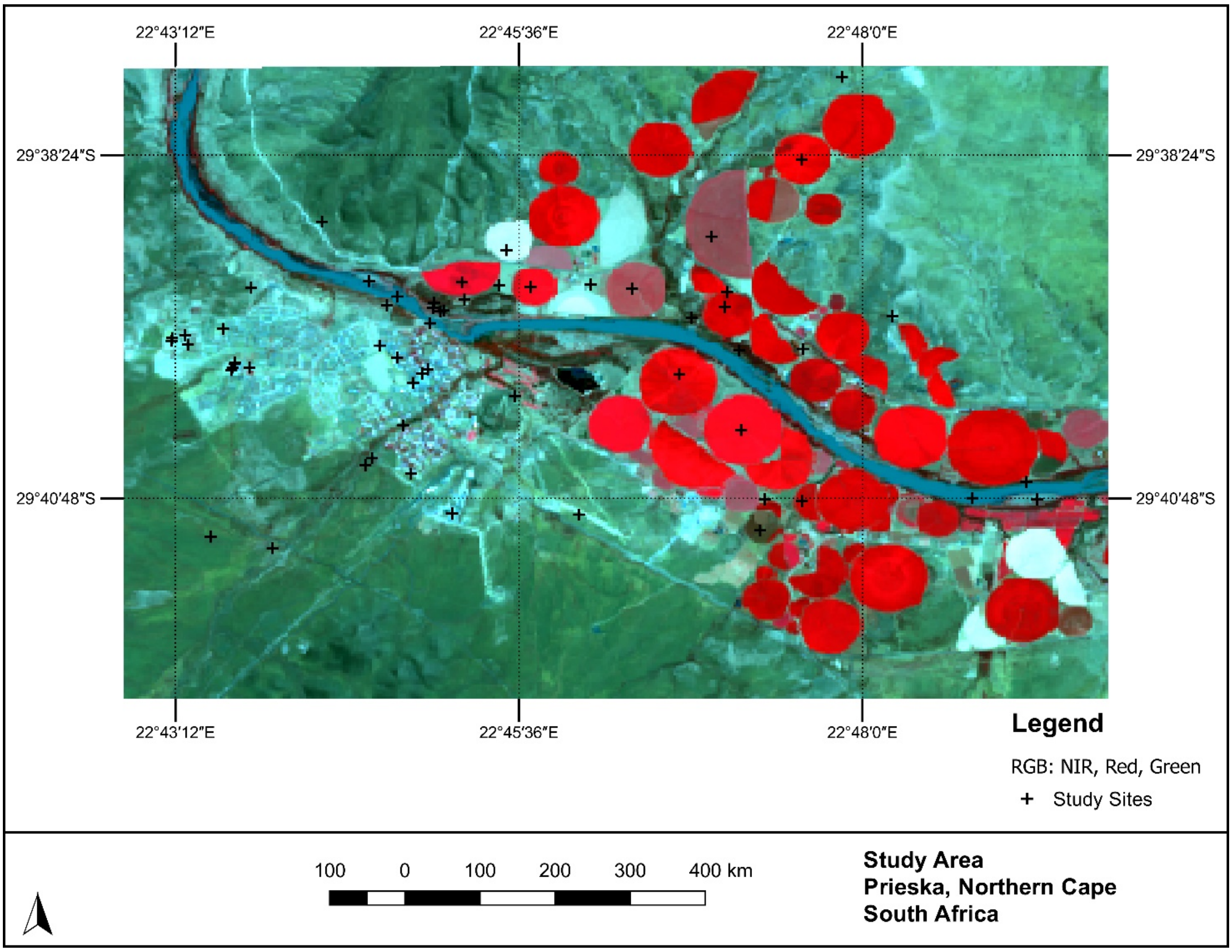

Figure 1. The false colour composite Landsat 8 OLI image of the study area around Prieska in the Northern Cape. The cross symbols indicate in situ data collection sites used for training and validation.

2.2. Data sets used

2.2.1. Satellite data products

The Landsat 5 TM and Landsat 8 OLI satellite imagery (Path171/83) Level-2 surface reflectance products were acquired from the U.S. Geological Survey (USGS) website (https://earthexplorer.usgs.gov/). The USGS employs the Land surface Reflectance Code (LaSRC) algorithm on Landsat 8 products to generate surface reflectance (Vermote et al., 2018). Data from TM sensor are corrected to surface reflectance using Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm (Schmidt et al., 2013). Nazeer et al. (2014) evaluated different surface reflectance models (FLASH, LEDAPS, atmospheric correction (ATCOR), dark object subtraction (DOS), and the empirical line method (ELM)) on Landsat TM/ETM sensors and found that all the models performed differently on different land cover types. Therefore, we adopted the standard USGS surface reflectance models for this study. The images were obtained for 28 years from 1990 to 2018 at roughly 5-year intervals (1990, 1997, 2005, 2013, 2018) based on the availability of Landsat 5 TM and Landsat 8 OLI data. This study used the six available 30-meter spatial resolution spectral bands of the Landsat products for classification, including blue, green, red, infrared, shortwave infrared (SWIR) 1 and 2. The visible and near-infrared bands were chosen as they encourage spectral separability ideal for the identification of vegetation and land cover (Adam et al., 2017). The SWIR bands were included to ensure vegetation with higher water content would be separated from low water content vegetation.

2.2.2. Field measurements

Training and validation data were acquired in the spring of 2018 from 15th to 21st of October. Random sampling was used to generate points of interest to identify initial land cover classes in the study area. These classes were used to guide the collection of ground reference data in Prieska. Farmers in the area were contacted to get permission to access roads, properties and the Orange River system. The Carlson Archer 2 Differential GPS allowed for high precision positioning with sub-centimetre accuracy used for the training data collection and verification of land cover types. At each site, at least four GPS points were collected with site descriptions and type of land cover observed. Ground reference sites (polygons) were created with the GPS point data collected at each site to classify the satellite imagery. Time and cost constraints limited the amount of in situ reference data collection and additional polygons were identified through Google Earth and satellite imagery. A total of seven classes were identified by grouping together several land cover types, including Agriculture, Bare Soil, Built-up, Indigenous, *P. glandulosa*, Shrublands, and Water. The classes were identified based on the dominant classes in our study area and their natural stability over time. For example, we assumed that the Orange River is there at present, then there is high probability that the river was there in our time frame of study which is 25 years in this case. The reference data were split into two categories i.e., the training and the validation data used to train the classifier and validate the results post-classification.

2.3. Description of machine learning algorithms

2.3.1. Support vector machine classifier

The SVM is a supervised machine learning algorithm developed by Cortes and Vapnik (1995). The algorithm is applied by obtaining a separating hyperplane or decision boundary by maximising the distance between the two classes. The distance between the two classes is called the margin. The data points nearest to the maximum and minimum margins define the hyperplane and are known as support vectors (Vapnik, 2000). Originally, the SVM is only known as a binary (linear) classifier in a nonseparable two-dimensional space. Once the kernel method is used in classification, for example, the polynomial kernel, data become separated in a high-dimensional space (Noble, 2006). The advantages of the SVM algorithm is the ability to classify a high-dimensional space and to eliminate the need for feature selection (Joachims, 1998). The accuracy of the polynomial kernel is dependent on the input values and can significantly influence the performance of the SVM classifier (Huang et al., 2002). The SVM algorithm is beneficial for minimising errors such as misclassification (Tso and Mather, 2009). The SVM algorithm was executed in the ENVI 5.1 software by using the Support Vector Machine Classification tool and the degree of kernel polynomial was set to 6.

2.3.2. Random forest classifier

The RF method proposed by Breiman (2001) is a machine learning algorithm that consists of a collection of traditional decision trees used to classify the sample. The original dataset is used to create a bootstrap sample training dataset. The bootstrap samples were used to grow an unpruned classification tree through bagging by randomly selecting features. For each node of the tree, variables are randomly chosen to choose the best split among all the trees. The new dataset is classified by selecting the majority vote from each predictor in the random forest (Liaw and Wiener, 2002, Breiman, 2001). The RF classification can perform better than other methods as it can mitigate problems of missing values and overfitting (Ali et al., 2012, Pal, 2005). The RF classification only requires two user-defined parameters: the number of decision trees and the number of random variables to be split (Rodriguez-Galiano et al., 2012). The RF classification was executed in Python using the Scikit-learn machine learning library and the number of decision trees in the forest was set to 500 (Pedregosa et al., 2011).

2.4. Classification accuracy assessment

Accuracy assessment was based on the overall accuracy, user's accuracy, producer's accuracy and Kappa (KHAT) statistic for each classification image. Validation data collected during the field assessment was used to produce a confusion matrix, which is used to calculate the overall, producer and user accuracies. The overall accuracy is calculated from the sum of the correctly classified pixels divided by the total of samples in the image. The producer's accuracy represents the probability that the sample data is classified correctly, whereas the user's accuracy is the probability of the sample pixels classified as a specific class belongs to that specific class (Al-Fares, 2013). The KHAT statistic is a measure of the level of agreement between the remotely sensed data and the reference data. In this study, polygons were used to validate each class during the classification of Landsat image data. Each validation polygon was selected away from the training polygons to eliminate the chance of overlapping pixels. The performance of each classifier (SVM and RF) were assessed based on overall accuracy and Kappa statistics for the individual error matrices. Pairwise analysis of the error matrices for both classifiers can be tested with the Z statistic, expressed by:

$$Z = \frac{|K_1 - K_2|}{\sqrt{var(K_1) - var(K_2)}}. \quad \textbf{[1]}$$

Where K1 and K2 are the Kappa values and var (K1) and var (K2) is the variance of the Kappa for classifier 1 and classifier 2. The Z statistic in Equation (1) is normally distributed given the null hypothesis HO: (K1 - K2) = 0, and the alternative H1: (K1 - K2) ≠ 0. HO is rejected if Z ≥ Zα/2, thus there is no significant difference between the two classifiers. However, if the null hypothesis is not rejected then there is a significant difference between the classifications. Therefore, the classifiers were tested to determine if the results were statistically different from one another at the 95% confidence interval (α = 0.05). The method described by Olofsson et al. (2013) was used to adjust the estimated proportion of area and the estimated accuracies as well as to produce the 95% confidence interval (CI) for the estimated adjusted areas . This procedure is recommended since the land cover change map may differ greatly from the true area of change due to classification errors. It is also an important procedure since the level of uncertainty can be quantified using CI. We refer the reader to Olofsson et al. (2013) for detail procedure on the computation of the adjusted error matrix and CI. The adjusted areas were used to obtain area changes for five-year intervals starting from 1990 to 2018.

3. Results

3.1. Classification of land cover classes with RF and SVM classifiers

The SVM and RF algorithms were used to classify land cover types in Prieska using Landsat imagery to produce maps of *P. glandulosa* spatial extent (Figure 2). Both classification algorithms identified the Shrubland class as the most common land cover class. The least dominant class were different for both classifications where the areal percentage for the Indigenous class in the SVM classification is 1.61% and in the RF classification the water class is 1.90%. The SVM classifier detected less (868.23 ha) of the IAP *P. glandulosa* than the RF classifier (1055.52 ha).

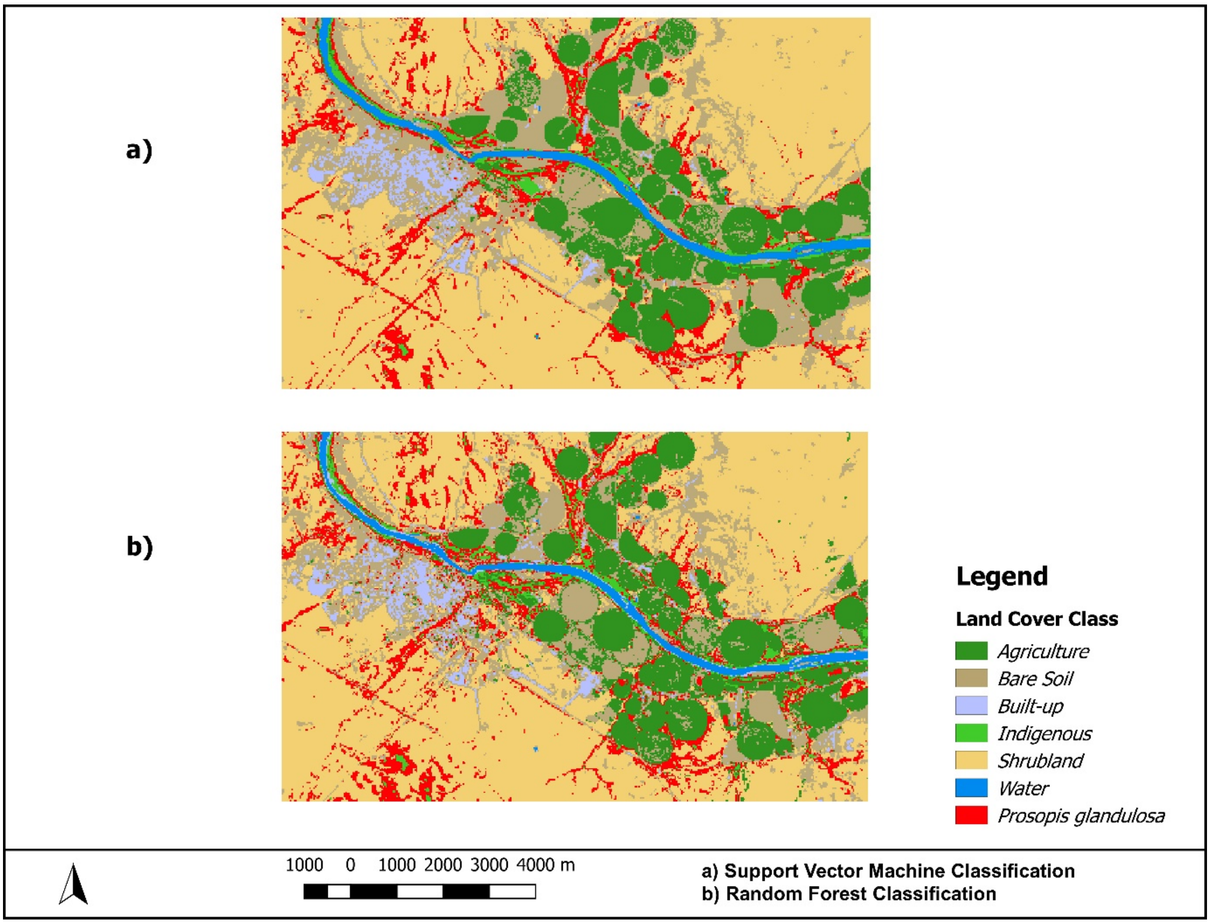

Figure 2: The maps illustrate SVM (a) and RF (b) land cover classifications for the study area in 2018. The maps represent each land cover class in a different colour where P. glandulosa is shown in red.

3.2. Performance of the two machine learning algorithms

The SVM and RF classifiers were evaluated using confusion matrix and Z statistical test to compare the performance of each classifier for mapping land cover in Prieska. Table 1 summarises the overall accuracy assessment results and the Kappa analysis for the 10 classification maps using the two machine learning algorithms, SVM and RF, from 1990 to 2018. The SVM overall accuracy ranged from 61% to 89% and the RF from 57% to 83% for 1990 to 2018. Generally, each year showed an increase in the area-adjusted overall accuracy compared to the unadjusted overall accuracy. The area-adjusted overall accuracy for the reference year 2018 is 88.58% for the SVM classifier. The RF classifier produced an 86.06% area-adjusted overall accuracy for the reference year. The area-adjusted overall accuracies for 1990 to 2013 ranged between 71% and 77% for the SVM and between 69% and 77% for the RF classifiers. The differences in accuracies of different classes can be attributed to several factors: (1) Climate variability may cause variations in spectral signature of land cover during dry/wet season; it is possible that classes such as agriculture have become another class such as bare soil. Possibly because the farmer did not plant during that season. (2) Higher mixel effect throughout the years compared to another i.e., a pixel that contain more than one type of land cover type, which can reduce the accuracies. (3) Landsat 5 and Landsat 8 surface reflectance data were processed by USGS using two different algorithms i.e., LEDAPS and LaSRC respectively. According to Vermote et al., (2016), the LaSRC algorithm produces higher atmospheric correction accuracy than the LEDAPS (Landsat 5 ETM+) by utilising the enhanced radiometric and spectral resolution of Landsat 8 OLI. These two algorithms are designed for the respective Landsat products and USGS does not use them interchangeably.

The overall accuracy and Kappa analysis indicated that throughout the year's different classification results were obtained. In 2018, the SVM classifier was more accurate than the RF classifier. The unadjusted overall accuracy and KHAT values improved from 83% and 0.78 respectively to 89% and 0.85. Similar results were observed for the SVM classification maps in 2005 and 1990. In 2013, the RF classification statistically outperformed the SVM classifier. The overall accuracy and KHAT values were observed at 61% and 0.46 for the RF classifier compared to 66% and 0.52 for the SVM classification. In 1997, there was no statistical difference between the overall accuracy for the SVM and RF algorithms and the KHAT values were 0.52 for both classifiers. The Z statistic indicates a statistical difference at the 95% confidence interval if the Z value is above 1.96. The statistic indicates Kappa analysis results for the pairwise comparison of the error matrices between the SVM and RF for each year. The Z values suggested that the SVM were statistically different from the RF for the years 2018, 2013, 2005 and 1990 at the 95% CI. In 1997, the results show a small difference between the overall accuracy and a Z value of 0.043 observed for the RF and SVM classifications. This indicates there was no statistical difference between both the SVM and RF classification.

Based on the confusion matrices for the images from 1990 to 2018, the accuracy assessment produced the adjusted user's accuracy (UA) and producer's accuracy (PA). Table 2 shows for both classification methods the UA and PA of five years for each land cover class identified in the study. The 2018 reference year achieved the highest UA and PA when classifying all the land cover classes for both SVM and RF classification compared to previous years. The agriculture land cover class showed above 60% UA for the SVM and RF classification from 1990 to 2018, however, the PA was below 65% for all the years except 2018. The range of commission error (CE) for the *P. glandulosa* land cover class was calculated to be approximately between 41.10% and 71.23% for the SVM and between 28.77% and 52.05% for the RF classification. The CE% for *P. glandulosa* indicates SVM at each year had more false classifications compared to RF classification. The omission error (OE) for *P. glandulosa* throughout the years ranges between 14.89% and 61.99% using the SVM, while the RF classifier ranges between 23.56% and 56.08%. In 1990, the OE% for *P. glandulosa* was high for both the SVM (61.99%) and RF (48.63%) classifications. The high OE% suggests that there was misclassification of the *P. glandulosa* land cover class.

Table 1: Overall accuracy assessment and Kappa statistics for SVM and RF classifiers for different images used in the study.

		Overall Accuracy			
Year	**Method**	**Unadjusted**	**Adjusted**	**KHAT**	**Z**
2018	SVM	89.09	88.58	0.85	3.672
	RF	83.44	86.06	0.78	
2013	SVM	61.44	73.33	0.46	2.112
	RF	66.03	76.90	0.52	
2005	SVM	67.27	76.72	0.54	2.863
	RF	60.10	72.48	0.45	
1997	SVM	63.54	71.33	0.52	0.043
	RF	62.87	68.54	0.52	
1990	SVM	62.30	71.33	0.48	2.269
	RF	56.84	67.61	0.41	

Table 2: The summary of the adjusted user's and producer's accuracies for the SVM and RF classification of each year from 1990 to 2018.

Support Vector Machine

Land Cover Class	19900916		19971005		20050909		20131118		20180929	
	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy
Agriculture	69.46	63.28	59.61	36.16	83.50	48.41	79.80	49.65	91.13	95.51
Bare Soil	33.46	39.64	44.12	65.28	19.12	37.09	6.25	11.14	97.43	80.43
Built-up	53.33	33.79	80.00	68.74	73.33	94.21	80.00	93.46	66.67	95.74
Indigenous	46.67	77.06	30.00	18.54	86.67	73.53	73.33	59.36	56.67	37.07
Prosopis glandulosa	28.77	38.01	58.90	64.11	47.95	85.11	45.21	53.97	47.95	67.63
Shrubland	92.93	94.42	94.44	94.09	95.45	96.46	92.42	99.41	92.42	94.95
Water	100.00	92.31	100.00	89.73	100.00	100.00	100.00	95.32	100.00	94.41

Random Forest

Land Cover Class	19900916		19971005		20050909		20131118		20180929	
	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy	User's Accuracy	Producer's Accuracy
Agriculture	59.11	55.95	60.59	39.28	64.04	37.61	86.21	55.69	79.80	93.26
Bare Soil	33.82	39.68	39.71	64.71	20.96	36.58	9.93	18.74	90.07	76.63
Built-up	66.67	49.66	60.00	34.13	80.00	88.47	66.67	91.44	53.33	55.55
Indigenous	43.33	41.15	20.00	21.70	83.33	83.11	43.33	38.75	53.33	32.78
Prosopis glandulosa	56.16	51.37	69.86	43.92	47.95	59.37	71.23	76.44	53.42	72.78
Shrubland	93.94	91.24	94.44	94.93	94.95	95.23	94.44	99.05	95.96	94.96
Water	23.53	63.43	98.04	71.39	100.00	96.20	100.00	97.52	98.04	96.49

3.3. Land cover change: from 1990 to 2018

Land cover change detection was studied using correlation matrices for each year to calculate the area-adjusted estimates based on Olofsson et al. (2013). The land cover changes over the period (from 1990 to 2018) for the study area were assessed by considering the adjusted areas and the percentage of change. The Shrubland land cover class shows small changes from 1990 to 1997, although a steep increase occurred from 1997 to 2005. The agriculture class changed constantly over time and is one of the most dynamic land cover classes since farmers constantly plant and rotate crops. The land cover class *P. glandulosa* shows little land change throughout the years, especially where the range of the error bars overlap which can be attributed to the lower UA and PA in this land cover class compared to other classes.

Table 3 shows the percentage area change for the seven land cover classes from year to year and the percentage change from 1990 to 2018. The two classification methods indicate a similar change over time throughout all the land cover classes. The agriculture class changed from 1990 to 2018 with a decrease of 17.5% and 14.8% for the SVM and RF classifiers, respectively. The percentage area of the Shrubland land cover class increased with 7.5% and 6.5% during the SVM and RF classifications between 1990 and 2018. The results show *P. glandulosa* variations over time as the percentage area fluctuated. There was a 1.9% increase in area covered by *P. glandulosa* from 1990 to 1997 compared to the decrease between 1.03% and 1.5% observed from 1997 to 2005. The overall change was high for the *P. glandulosa* class with above a 3.3% and 3.7% increase in the percentage change from 1990 to 2018 with the SVM and RF classifiers, respectively.

Table 3: The difference percentage of land cover change over time is represented in the tables below for the SVM and RF classifiers.

	Support Vector Machine %Δ				
Class	**1990-1997**	**1997-2005**	**2005-2013**	**2013-2018**	**1990-2018**
Agriculture	-10.95	6.04	3.29	-15.87	-17.49
Bare Soil	5.35	-11.56	0.11	14.55	8.45
Built-up	-1.56	-1.08	0.53	-1.11	-3.21
Indigenous	1.91	0.52	-1.28	0.42	1.58
Prosopis glandulosa	1.94	-1.56	0.39	2.55	3.32
Shrubland	3.48	7.74	-3.42	-0.32	7.49
Water	-0.18	-0.11	0.38	-0.23	-0.13

Table 3: Continues.

	Random Forest				
	%Δ				
Class	**1990-1997**	**1997-2005**	**2005-2013**	**2013-2018**	**1990-2018**
Agriculture	-9.11	4.75	5.89	-16.36	-14.83
Bare Soil	0.85	-7.67	-3.39	13.66	3.45
Built-up	1.71	-3.92	-0.43	0.50	-2.14
Indigenous	2.16	0.79	-2.07	1.28	2.16
Prosopis glandulosa	1.89	-1.03	1.75	1.04	3.65
Shrubland	0.33	8.06	-2.07	0.19	6.51
Water	2.16	-0.97	0.32	-0.31	1.20

4. Discussion

The study evaluated two machine learning algorithms for mapping land cover types that include the IAP *P. glandulosa* in the semi-arid environment of Prieska. The two classifications methods i.e. SVM and RF classifiers were investigated and produced high accuracy results using 30-meter Landsat imagery from 1990 to 2018 period. The performance of the two machine learning algorithms varied throughout the years. The SVM classifier outperformed the RF classifier for three of the five years that were classified (2018, 2005 and 1997). In 2013, the RF classifier outperformed the SVM classifier with a 4.6% difference in overall accuracy. Both machine learning algorithms produced high accuracy for the 1997 classification. The results suggest that both the SVM and RF classifiers have the potential to produce high accuracy classification for mapping land cover types. Our results are in agreement with Li et al. (2013) that compared three machine learning algorithms. In this study a forest ecosystem was classified with decision trees, RF and SVMs classifiers using multi-temporal Landsat imagery. The performance of the RF and SVM classifiers were comparable, although they both outperformed the decision tree classifier in terms of the overall accuracy with high-dimensional data.

Higher UA and PA were observed with the RF classifier than the SVM when mapping of *P. glandulosa* in the study area. Previous studies have shown the success of RF to discriminate between IAPs and other vegetation with high spatial resolution multispectral data (Odindi et al., 2016, Ng et al., 2016) and hyperspectral data (Große-Stoltenberg et al., 2016). The most confusion between classes occurred between the indigenous and *P. glandulosa* land cover classes where the lowest UA and PA were observed using 30-meter medium spatial resolution Landsat imagery. The use of high spatial resolution has been reported to perform better for species-level mapping where spectral characteristics successfully distinguish between physiological differences of IAPs. For example, the mapping of *Prosopis* spp. with WV-2 high-resolution imagery (Adam et al., 2017, Robinson et al., 2016). Unfortunately, the use of WV-2 imagery often can become costly and imagery is not available for long-term land cover change studies.

The land cover change detected in this study indicates that *P. glandulosa* invasion from 1990 to 2018 is responsible for a considerable transformation in the study area. Both the SVM and RF classifiers indicated that the area of *P. glandulosa* increased between 3.3% and 3.7% over 28 years. Figure 3 shows the change in *P. glandulosa* distribution in the study area as the IAP gradually increased from low density stands to the highest density between 1990 and 2018. Declines of indigenous vegetation are often reported in semi-arid to arid regions of South Africa, for example, the Northern Cape, where large scale changes of *Prosopis* spp. invasions were observed (Richardson and Van Wilgen, 2004, Ndhlovu et al., 2016).

Water catchments in arid environments are especially threatened by the continuous invasion of IAP as uncontrolled spread throughout the catchments decreases river flow and water yields (Preston et al., 2018). The prioritisation of catchment areas throughout South Africa indicated that the Northern Cape needed the most management strategies to combat *Prosopis* spp. invasion (Shackleton et al., 2017). In this study, the high density of *P. glandulosa* invasion was seen along the Orange River and other water catchments (Figure 3) found throughout the study area. In the study by Van den Berg (2010) high-density *Prosopis* spp. was found to have invaded most of the Northern Cape catchments. Management strategies for removal of *Prosopis* spp. invasions are essential to ensure a decrease in groundwater and surface water loss by implementing more clearing programmes (Dzikiti et al., 2017).

The spread of *Prosopis* spp. is recognized throughout the Northern Cape by locals as a problematic plant even though the plant provides some positive incentive to local livelihoods (Shackleton and Shackleton, 2018). The plant is an essential source of fodder for livestock during times where other food sources are not available. The distribution of *P. glandulosa* (Figure 5) indicates higher densities in some areas where livestock are present. Livestock dispersal is responsible for the invasion of *P. glandulosa* as seeds are consumed and spread through dung that provides favourable conditions for germination or deposited into the seed bank (Ansley et al., 2017). Livestock also uncovers seeds when the pods fall to the ground which allows the distribution away from the parent plant and further germination to occur (Alvarez et al., 2017). Farmers have raised concerns about the spread of *Prosopis* spp. through their pods and the impact of its invasion on agriculture. Many farmers do clearing to decrease the presence of *Prosopis* spp. on their farms and they often sell the wood to local communities to cover the cost of clearing (Shackleton et al., 2015b).

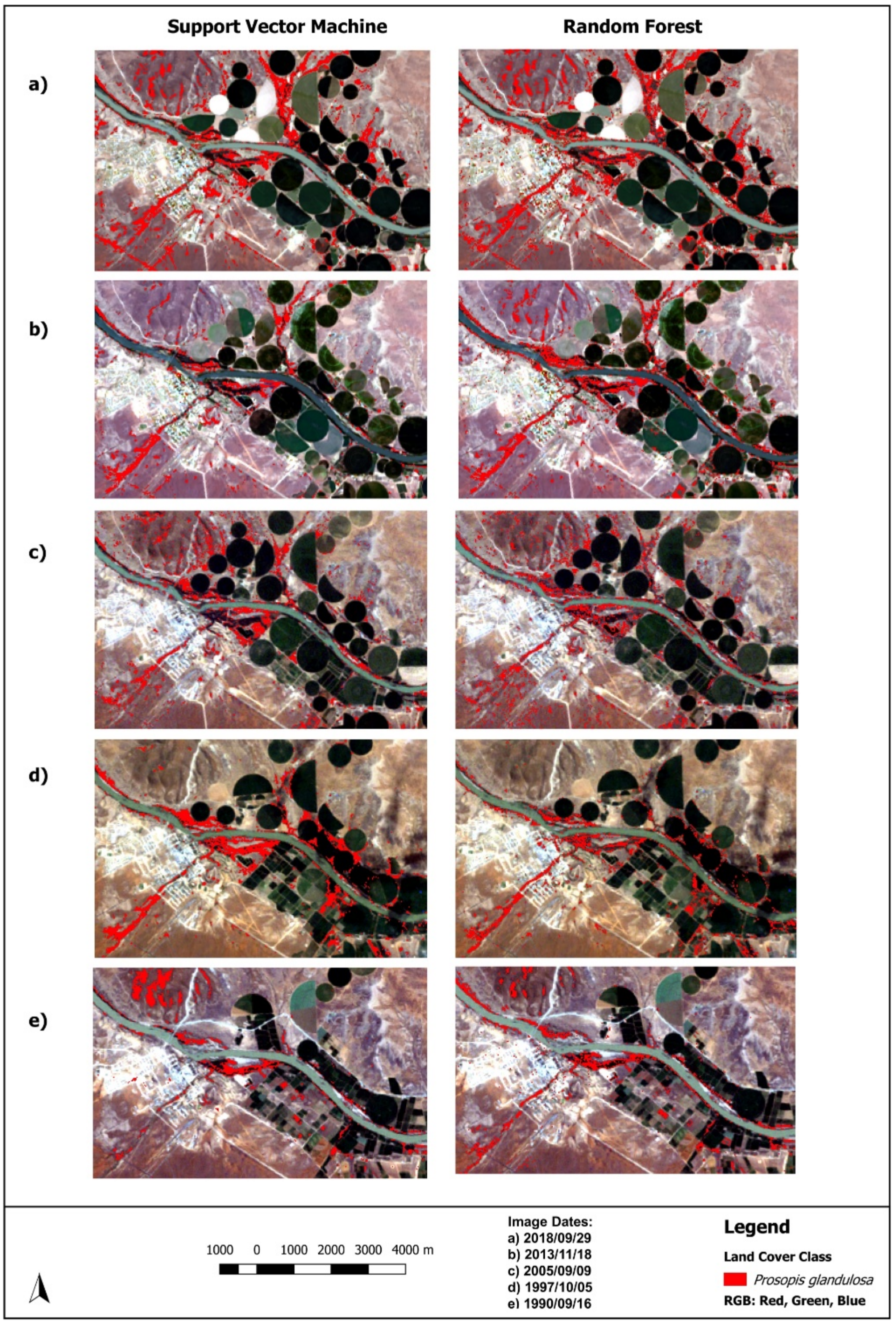

Figure 3: The distribution maps for *P. glandulosa* were produced by the SVM (on the left) and the RF (on the right) classification. The maps show a subset of the study area to illustrate the densities of the IAP for the years: 2018 (a), 2013 (b), 2005 (c), 1997 (d) and 1990 (e).

5. Conclusion

The objective of the study was to use machine learning algorithms to map the distribution and spread of *P. glandulosa* over 28 years. The research investigated the ability of two machine learning algorithms to map the *P. glandulosa* using multitemporal Landsat imagery. The results suggest that both algorithms produced statistically similar classifications. Overall, the results indicate that *P. glandulosa* invasion increased from 1990 to 2018, especially next to the water catchment areas (i.e. Orange River) where high-density canopies threaten water yields of the Northern Cape Province. The findings can be used to inform local government management initiatives. Future research should focus on other municipalities using high spatial resolution datasets to improve classification products.

6. Acknowledgements

The research was funded by the Geoinformatics Division at the Agricultural Research Council-Institute for Soil, Climate and Water (ARC-ISCW). The researchers are also grateful to the Centre for Geoinformation Sciences at the University of Pretoria for providing the differential GPS instrumentation that was used in this study. Furthermore, the authors would like to thank Mr. Eric Economon for assisting with data acquisition during the field campaign.

7. References

Adam, E., Mureriwa, N. & Newete, S. 2017, 'Mapping *Prosopis glandulosa* (mesquite) in the semi-arid environment of South Africa using high-resolution WorldView-2 imagery and machine learning classifiers', *Journal of Arid Environments*.

Adam, E., Mutanga, O., Odindi, J. & Abdel-Rahman, E. M. 2014, 'Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers', *International Journal of Remote Sensing,* vol. 35**,** no. 10, pp. 3440-3458.

Al-Fares, W. 2013, *Historical Land Use/land Cover Classification Using Remote Sensing: A Case Study of the Euphrates River Basin in Syria,* Springer Science & Business Media.

Ali, J., Khan, R., Ahmad, N. & Maqsood, I. 2012, 'Random forests and decision trees', *International Journal of Computer Science Issues (IJCSI),* vol. 9**,** no. 5, pp. 272.

Alvarez, M., Leparmarai, P., Heller, G. & Becker, M. 2017, 'Recovery and germination of *Prosopis juliflora* (Sw.) DC seeds after ingestion by goats and cattle', *Arid land research and management,* vol. 31**,** no. 1, pp. 71-80.

Amboka, A. & Ngigi, T. 2015, 'Mapping and monitoring spatial-temporal cover change of *Prosopis* species colonization in Baringo Central, Kenya', *International Journal of Engineering Science Invention,* vol. 4**,** pp. 2319-6734.

Ansley, R., Pinchak, W. & Owens, M. 2017, 'Mesquite pod removal by cattle, feral hogs, and native herbivores', *Rangeland Ecology & Management,* vol. 70**,** no. 4, pp. 469-476.

Beck, K. G., Zimmerman, K., Schardt, J. D., Stone, J., Lukens, R. R., Reichard, S., Randall, J., Cangelosi, A. A., Cooper, D. & Thompson, J. P. 2008, 'Invasive species defined in a policy context: Recommendations from the Federal Invasive Species Advisory Committee', *Invasive Plant Science and Management,* vol. 1**,** no. 4, pp. 414-421.

Breiman, L. 2001, 'Random forests', *Machine learning,* vol. 45**,** no. 1, pp. 5-32.

Cortes, C. & Vapnik, V. 1995, 'Support-vector networks', *Machine learning,* vol. 20**,** no. 3, pp. 273-297.

Department of Water Affairs, D. 2013, *National Water Resource Strategy: Water for an Equitable and Sustainable Future,* Department of Water Affairs.

Dzikiti, S., Ntshidi, Z., Le Maitre, D., Bugan, R., Mazvimavi, D., Schachtschneider, K., Jovanovic, N. & Pienaar, H. 2017, 'Assessing water use by *Prosopis* invasions and Vachellia karroo trees: Implications for groundwater recovery following alien plant removal in an arid catchment in South Africa', *Forest Ecology and Management,* vol. 398**,** pp. 153-163.

Dzikiti, S., Schachtschneider, K., Naiken, V., Gush, M., Moses, G. & Le Maitre, D. C. 2013, 'Water relations and the effects of clearing invasive *Prosopis* trees on groundwater in an arid environment in the Northern Cape, South Africa', *Journal of Arid Environments,* vol. 90, pp. 103-113.

Gaggini, L., Rusterholz, H.-P. & Baur, B. 2019, 'The invasion of an annual exotic plant species affects the above- and belowground plant diversity in deciduous forests to a different extent', *Perspectives in Plant Ecology, Evolution and Systematics,* vol. 38, pp. 74-83.

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,* vol. 8, no. 4, pp. 334.

Hejda, M., Pyšek, P. & Jarošík, V. 2009, 'Impact of invasive plants on the species richness, diversity and composition of invaded communities', *Journal of ecology,* vol. 97, no. 3, pp. 393-403.

Huang, C., Davis, L. & Townshend, J. 2002, 'An assessment of support vector machines for land cover classification', *International Journal of remote sensing,* vol. 23, no. 4, pp. 725-749.

Joachims, T. 1998. 'Text categorization with support vector machines: Learning with many relevant features' *European conference on machine learning*. pp. 137-142.

Ku, N.-W. & Popescu, S. C. 2019, 'A comparison of multiple methods for mapping local-scale mesquite tree aboveground biomass with remotely sensed data', *Biomass and bioenergy,* vol. 122, pp. 270-279.

Laliberte, A. S., Browning, D. & Rango, A. 2012, 'A comparison of three feature selection methods for object-based classification of sub-decimeter resolution UltraCam-L imagery', *International Journal of Applied Earth Observation and Geoinformation,* vol. 15, pp. 70-78.

Li, M., Im, J. & Beier, C. 2013, 'Machine learning approaches for forest classification and change analysis using multi-temporal Landsat TM images over Huntington Wildlife Forest', *GIScience & Remote Sensing,* vol. 50, no. 4, pp. 361-384.

Li, X., Chen, W., Cheng, X. & Wang, L. 2016, 'A comparison of machine learning algorithms for mapping of complex surface-mined and agricultural landscapes using ZiYuan-3 stereo satellite imagery', *Remote sensing,* vol. 8, no. 6, pp. 514.

Liaw, A. & Wiener, M. 2002, 'Classification and regression by randomForest', *R news,* vol. 2, no. 3, pp. 18-22.

Liu, L. & Piper, B. 2016, 'Predicting the total economic impacts of invasive species: The case of B. rubostriata (red streaked leafhopper)', *Ecological Economics,* vol. 128, pp. 139-146.

Meroni, M., Ng, W. T., Rembold, F., Leonardi, U., Atzberger, C., Gadain, H. & Shaiye, M. 2017, 'Mapping *Prosopis juliflora* in west Somaliland with Landsat 8 satellite imagery and ground information', *Land Degradation & Development,* vol. 28, no. 2, pp. 494-506.

Mirik, M. & Ansley, R. J. 2012, 'Comparison of ground-measured and image-classified mesquite (*Prosopis glandulosa*) canopy cover', *Rangeland Ecology & Management,* vol. 65, no. 1, pp. 85-95.

Nazeer, M., Nichol, J. E. & Yung, Y.-K. 2014, 'Evaluation of atmospheric correction models and Landsat surface reflectance product in an urban coastal environment', *International journal of remote sensing,* vol. 35, no. 16, pp. 6271-6291.

Ndhlovu, T., Milton-Dean, S. & Esler, K. 2011, 'Impact of *Prosopis* (mesquite) invasion and clearing on the grazing capacity of semiarid Nama Karoo rangeland, South Africa', *African Journal of Range & Forage Science,* vol. 28, no. 3, pp. 129-137.

Ndhlovu, T., Milton, S. J. & Esler, K. J. 2016, 'Effect of *Prosopis* (mesquite) invasion and clearing on vegetation cover in semi-arid Nama Karoo rangeland, South Africa', *African journal of range & forage science,* vol. 33, no. 1, pp. 11-19.

Ng, W.-T., Meroni, M., Immitzer, M., Böck, S., Leonardi, U., Rembold, F., Gadain, H. & Atzberger, C. 2016, 'Mapping *Prosopis* spp. with Landsat 8 data in arid environments: Evaluating effectiveness of different methods and temporal imagery selection for Hargeisa, Somaliland', *International Journal of Applied Earth Observation and Geoinformation,* vol. 53, pp. 76-89.

Noble, W. S. 2006, 'What is a support vector machine?', *Nature biotechnology,* vol. 24, no. 12, pp. 1565.

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,* vol. 46, no. 2, pp. 1-9.

Olofsson, P., Foody, G. M., Stehman, S. V. & Woodcock, C. E. 2013, 'Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation', *Remote Sensing of Environment,* vol. 129, pp. 122-131.

Pal, M. 2005, 'Random forest classifier for remote sensing classification', *International Journal of Remote Sensing,* vol. 26, no. 1, pp. 217-222.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011, 'Scikit-learn: Machine learning in Python', *Journal of machine learning research,* vol. 12, no. Oct, pp. 2825-2830.

Preston, I., Le Maitre, D. C., Blignaut, J., Louw, L. & Palmer, C. 2018, 'Impact of invasive alien plants on water provision in selected catchments', *Water SA,* vol. 44, no. 4, pp. 719-729.

Rembold, F., Leonardi, U., Ng, W.-T., Gadain, H., Meroni, M. & Atzberger, C. 2015. 'Mapping areas invaded by Prosopis juliflora in Somaliland on Landsat 8 imagery' *Remote Sensing for Agriculture, Ecosystems, and Hydrology XVII*. pp. 963723.

Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Panetta, F. D. & West, C. J. 2000, 'Naturalization and invasion of alien plants: concepts and definitions', *Diversity and distributions,* vol. 6, no. 2, pp. 93-107.

Richardson, D. M. & Van Wilgen, B. W. 2004, 'Invasive alien plants in South Africa: how well do we understand the ecological impacts?: working for water', *South African Journal of Science,* vol. 100, no. 1-2, pp. 45-52.

Robinson, T., Wardell-Johnson, G., Pracilio, G., Brown, C., Corner, R. & Van Klinken, R. 2016, 'Testing the discrimination and detection limits of WorldView-2 imagery on a challenging invasive plant target', *International Journal of Applied Earth Observation and Geoinformation,* vol. 44, pp. 23-30.

Rodriguez-Galiano, V. F. & Chica-Rivas, M. 2014, 'Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models', *International Journal of Digital Earth,* vol. 7, no. 6, pp. 492-509.

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

Rouget, M., Richardson, D. M., Nel, J. L., Le Maitre, D. C., Egoh, B. & Mgidi, T. 2004, 'Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability', *Diversity and Distributions,* vol. 10, no. 5-6, pp. 475-484.

Schmidt, G., Jenkerson, C. B., Masek, J., Vermote, E. & Gao, F. 2013. Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description. US Geological Survey.

Shackleton, R. T., Le Maitre, D. C., Pasiecznik, N. M. & Richardson, D. M. 2014, '*Prosopis*: a global assessment of the biogeography, benefits, impacts and management of one of the world's worst woody invasive plant taxa', *AoB PLANTS,* vol. 6.

Shackleton, R. T., Le Maitre, D. C. & Richardson, D. M. 2015a, '*Prosopis* invasions in South Africa: Population structures and impacts on native tree population stability', *Journal of Arid Environments,* vol. 114, pp. 70-78.

Shackleton, R. T., Le Maitre, D. C. & Richardson, D. M. 2015b, 'Stakeholder perceptions and practices regarding *Prosopis* (mesquite) invasions and management in South Africa', *Ambio,* vol. 44, no. 6, pp. 569-581.

Shackleton, R. T., Le Maitre, D. C., Van Wilgen, B. W. & Richardson, D. M. 2017, 'Towards a national strategy to optimise the management of a widespread invasive tree (*Prosopis* species; mesquite) in South Africa', *Ecosystem Services,* vol. 27, pp. 242-252.

Shackleton, S. E. & Shackleton, R. T. 2018, 'Local knowledge regarding ecosystem services and disservices from invasive alien plants in the arid Kalahari, South Africa', *Journal of arid environments,* vol. 159, pp. 22-33.

Singh, A. 1989, 'Review article digital change detection techniques using remotely-sensed data', *International journal of remote sensing,* vol. 10, no. 6, pp. 989-1003.

Tso, B. & Mather, P. M. 2009, *Classification methods for remotely sensed data,* 2nd ed., CRC Press,Boca Raton.

Van Den Berg, E. C. 2010. *Detection, quantification and monitoring Prosopis spp. in the Northern Cape Province of South Africa using remote sensing and GIS.* North-West University.

Van Den Berg, E. C., Kotze, I. & Beukes, H. 2013, 'Detection, quantification and monitoring of *Prosopis* in the Northern Cape Province of South Africa using Remote Sensing and GIS', *South African Journal of Geomatics,* vol. 2, no. 2, pp. 68-81.

Van Wilgen, B., Nel, J. & Rouget, M. 2007, 'Invasive alien plants and South African rivers: a proposed approach to the prioritization of control operations', *Freshwater Biology,* vol. 52, no. 4, pp. 711-723.

Van Wilgen, B. W., Forsyth, G. G. & Le Maitre, D. C. 2008, 'The prioritization of species and primary catchments for the purposes of guiding invasive alien plant control operations in the terrestrial biomes of South Africa', *Unpublished report. CSIR Natural Resources and the Environment, Stellenbosch*.

Van Wilgen, B. W., Forsyth, G. G., Le Maitre, D. C., Wannenburgh, A., Kotzé, J. D., Van Den Berg, E. & Henderson, L. 2012, 'An assessment of the effectiveness of a large, national-scale invasive alien plant control strategy in South Africa', *Biological Conservation,* vol. 148, no. 1, pp. 28-38.

Van Wilgen, B. W. & Richardson, D. M. 2014, 'Challenges and trade-offs in the management of invasive alien trees', *Biological invasions,* vol. 16, no. 3, pp. 721-734.

Vapnik, V. N. 2000, *The nature of statistical learning theory,* 2nd ed., Springer,New York.

Vermote, E., Justice, C., Claverie, M. & Franch, B. 2016, 'Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product', *Remote Sensing of Environment,* vol. 185, pp. 46-56.

Vermote, E., Roger, J.-C., Franch, B. & Skakun, S. 2018. 'LaSRC (Land Surface Reflectance Code): Overview, application and validation using MODIS, VIIRS, LANDSAT and Sentinel 2 data's' *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. pp. 8173-8176.

Zachariades, C., Hoffmann, J. & Roberts, A. 2011, 'Biological control of mesquite (Prosopis species)(Fabaceae) in South Africa', *African Entomology,* vol. 19, no. 2, pp. 402-415.

Zhang, X., He, G., Zhang, Z., Peng, Y. & Long, T. 2017, 'Spectral-spatial multi-feature classification of remote sensing big data based on a random forest classifier for land cover mapping', *Cluster Computing,* vol. 20, no. 3, pp. 2311-2321.

Zimmermann, H. 1991, 'Biological control of mesquite, *Prosopis* spp.(Fabaceae), in South Africa', *Agriculture, ecosystems & environment,* vol. 37, no. 1-3, pp. 175-186.