VALUE OF FEATURE REDUCTION FOR CROP DIFFERENTIATION USING MULTI-TEMPORAL IMAGERY, MACHINE LEARNING, AND OBJECT-BASED IMAGE ANALYSIS

J. K. Gilbertson [a], A. Van Niekerk [a,b]*

[a] Department of Geography and Environmental Studies, University of Stellenbosch, Stellenbosch, South Africa – 16255135@sun.ac.za
[b] School of Plant Biology, University of Western Australia, 35 Stirling Hwy, Crawley WA 6009, Perth, Australia – avn@sun.ac.za

KEY WORDS: Feature Selection, Machine Learning, Object-Based Image Analysis, Multi-Temporal, Supervised Classification

ABSTRACT

This study examined the value of automated and manual feature selection, when applied to machine learning and object-based image analysis (OBIA), for the differentiation of crops in a Mediterranean climate. Five Landsat8 images covering the phenological stages of seven major crops types in the study area (Cape Winelands, South Africa) were acquired and processed. A statistical image fusion technique was used to enhance the spatial resolution of the imagery. The pan-sharpened imagery was used to produce a range of spectral features, textural measures, indices and colour transformations, after which it was segmented using the multi-resolution (MRS) algorithm. The entire set of 205 features (41 per image capture date) was then subjected to different feature selection and reduction methods. The feature selection and reduction methods included manual feature removal (i.e. grouping into semantic themes), filter methods (such as classification and regression trees (CART) and random forest (RF)), and statistical principal components analysis (PCA). The experiments were carried out in two scenarios, namely 1) on all input images in combination; and 2) on each individual image date. The feature subsets were used as input to decision trees (DTs), k-nearest neighbour (k-NN), support vector machine (SVM), and random forest (RF) machine learning classifiers. In order to assess the value of each feature reduction method (comprising feature reduction and selection techniques), overall accuracy, kappa coefficient and McNemar's test were employed to assess classification accuracy and compare the results. The results show that feature selection was able to improve the overall crop identification accuracy for the DT, k-NN, and RF classifiers, but was unable to do so for SVM. SVM scored the highest overall accuracy and kappa coefficient, even without applying feature reduction or selection. Based on these results it was concluded that, although feature selection can aid the crop differentiation process, it is not a necessity.

1. INTRODUCTION

Accurate crop maps are required for the health of an economy's agricultural sector, as they can be used for yield forecasting and keeping agricultural database statistics up to date (Monfreda et al. 2008). The production of crop maps has traditionally been done through field visits, which are costly and biased. Remote sensing has been proposed as a cost-effective solution (Castillejo-Gonzalez & Lopez-Granados 2009), as it can be linked to climate, soil properties, terrain characteristics, and light-use efficiency and therefore provide farmers with crucial information on crop health and moisture content, resulting in better decisions about irrigation and fertilization (Monfreda et al. 2008; Xin et al. 2015). In recent years, multi-temporal optical data has been the preferred data source for crop type mapping. However, multi-temporal approaches often lead to large datasets and features, which often result in the so-called 'curse of dimensionality'. This phenomenon occurs when an increase in the number of input features leads to a decrease in classification accuracy due to feature space sparseness (Gislason et al. 2006; Rodriguez-Galiano et al. 2012). One way of mitigating this effect is through feature selection (Rodriguez-Galiano et al. 2012) and/or reduction (Zhang et al. 2009).

2. MAIN BODY

2.1 Study Site

The study was carried out in the Cape Winelands region of South Africa (Figure 1). The 1040 km^2 study site, which extends from 33°34'39" to 33°52'17" S and 18°32'24" to 18°54'43" E, was selected based on the availability of multitemporal cloud free Landsat8 imagery and the variety of winter and summer crops produced in the region. The study site has a Mediterranean climate with cool wet winters and warm dry summers, an average annual rainfall of 550 mm, a mean annual temperature minima of 11°C and a maxima of 22°C (Tererai et al. 2015).

* Corresponding author

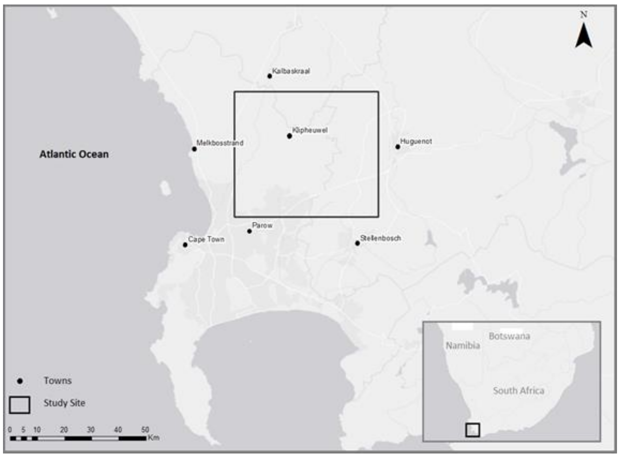

Figure 1. Location of the study area in the Western Cape, South Africa.

The area produces a wide range of crops, the most common of which are canola, grapes (mainly for wine production), lucerne (alfalfa), lupine, olives, managed pasture, and wheat. The phenological and agricultural production stages of these crops are shown in Figure 2. The annuals canola, lupine, pasture grasses and wheat are grown during southern hemisphere winter (April to August), while grapes are harvested during summer months (December to February). Lucerne is cut throughout the year, while olives are harvested during the early winter months.

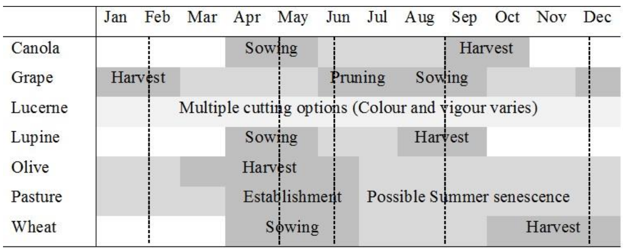

Figure 2. Phenological and agricultural production phases of targeted crops. Dark grey represents important agricultural phases, light grey the growth of the crop, and white represents bare field. The selected imagery dates are indicated with dotted lines.

2.2 Segmentation

Multiresolution segmentation (MRS) was chosen as the segmentation algorithm. This function uses scale, shape, compactness, and layer weighting as input parameters. The ESP (Estimation Scale Parameter) tool (developed by Dragut et al. (2010)), which attempts to quantitatively estimate an optimal scale parameter, was applied but recommended an unrealistically large scale factor. A qualitative assessment was consequently carried out by comparing systematic segmentation outputs (with increasing scale parameters) to shapefiles of digitised and verified crop fields. The literature on similar segmentation attempts was also consulted. Based on visual assessment, a favourable delineation of objects was achieved using 0.45, 0.55, and 14 for shape, compactness and scale factor respectively, with a slightly heavier weighting for the green, red, near-infrared, and short wave infrared spectral bands.

2.3 Image Feature Set Generation

Several other features (in addition to the Landsat8 spectral bands) were used as input to the classifications. Table 2 outlines the 205 features considered (41 per image capture date). For the spectral features, mean and standard deviation values were calculated for each of the objects. Coastal aerosol (band 1), cirrus (band 8), thermal infrared 1 (band 10), and thermal infrared 2 (band 11) were excluded as they were deemed unsuitable for crop type differentiation. Nine indices commonly used in vegetation studies were generated from the remaining bands. The textural features considered included contrast, correlation, and entropy, as recommended by Clausi (2002). Several image transforms relating to principal component analysis (PCA), colour transformations and tasselled cap were carried out. Coefficients for the latter were obtained from Baig et al. (2014).

Table 1. Features used as input for the DT, NN, SVM, and RT classifiers.

Type	*Subtype*	*Features*
Spectral Features	Mean	Blue, Green, Red, NIR, SWIR 1, SWIR 2, Panchromatic
	SD	Green, Red, NIR, SWIR 1, SWIR 2
Indices		ARVI, EVI, GCI, GNDVI, Green Index, NDVI, RGRI, SAVI, SRI
Textural features	GLCM	Contrast, Correlation, Entropy, Homogeneity
	GLDV	Correlation, Entropy, Mean
Image transforms	PCA	PC1, PC2, PC3, PC4
	Tasselled cap	Brightness, Greenness, Wetness, Transformation 4, Transformation 5, Transformation 6
	HSI	Hue, Saturation, Intensity

2.4 Feature Selection

A number of feature selection and reduction methods were used in this study. This included the two filter-based feature selection methods CART and RF, which were carried out using Salford Systems Data Mining and Predictive Software. The feature reduction technique PCA was performed in ArcMap. Manual feature subset selection was also done.

For RF the number of trees to build was set to 1000 for better performance, as computation time and power was not as issue. The number of variables to be considered at each node was based on the Breiman (2001) recommendation of the square root of K (where K is the number of predictors). Boot strap sample size was left at AUTO and a balanced class weighting was applied, as recommended by the Salford Systems user's manual.

For CART the search intensity was set to the maximum (400), the splitting method was set to GINI (as done by Yu et al. (2006) and Lewis (2000)), while the V-fold cross-validation was set to 10, as done by Lewis (2000) and recommended by the Salford Systems user's manual. The maximum number of nodes and depth was set to AUTO, the "apply no penalties to any variables" setting was enabled and a threshold level of 15 was set for enabling intelligent categorical split search.

Variable importance, as produced by CART and RF, was used to select a relevant subsets of features. Feature reduction and selection were done iteratively with user defined intervals of 75, 50, 25, 20, 15, 10, 5, 4, 3, 2, and 1 variable(s). The CART variable importance scores tapered off as soon 75 features were used, so no selection iterations larger than 75 were implemented.

PCA was generated using the original image bands, and the number of components used was based on eigenvalue percentages so as to remove components that accounted for insignificant covariance.

2.5 Classification Software

The Supervised Learning and Image Classification Environment (SLICE) software developed by the Centre for Geographical Analysis was used for classification and accuracy assessment. SLICE was developed using the C++ programming language and libraries from OpenCV 2.2 (Bradski & Pisarevsky 2000) and Libsvm (Chang and Lin 2011). SLICE includes a range of classification algorithms, including decision trees (DTs), k-nearest neighbour (k-NN), random forests (RF) (called random trees in OpenCV), and support vector machine (SVM). For SVM the radial basis function kernel was chosen for this study, as recommended by Hsu et al. (2010), and for the k-NN classification k was set to 1 as used by Qian et al. (2015). The geospatial data abstraction library is used to manipulate shapefiles and raster files. A 3:2 sample split ratio was employed for classification and accuracy assessment (the same set of training and testing samples were used for all experiments).

3. RESULTS AND DISCUSSION

The results (Table 2) show that feature selection and reduction improved classification accuracies for three out of the four classifiers. In Figures 3 and 4 it is clear that DT, NN, and RF initially benefits from feature selection/reduction, but then accuracy is reduced as the information content diminishes. The two tree classifiers showed smaller improvements with feature selection/reduction compared to NN. This is expected, given that NN has been shown to be sensitive under conditions of high dimensionality (Myburgh & Van Niekerk 2014).

Table 2. Overall accuracies and kappa coefficients for all classifications scenarios. All scenarios are also assigned a run number.

Run No.	Feature Reduction	Feature Count	DT OA	DT K	NN OA	NN K	RF OA	RF K	SVM OA	SVM K
1	None	205	81,8	0,78	78,9	0,75	87,5	0,82	95,9	0,95
2	CART	75	86,3	0,83	88,1	0,85	89,3	0,87	93,9	0,92
3	CART	60	83,0	0,79	87,1	0,84	88,7	0,86	92,3	0,90
4	CART	50	80,2	0,76	78,4	0,74	84,9	0,82	91,9	0,90
5	CART	40	75,2	0,70	77,7	0,73	84,5	0,81	90,2	0,88
6	CART	30	73,8	0,68	76,9	0,72	82,3	0,79	87,7	0,85
7	CART	20	73,8	0,68	64,6	0,58	79,4	0,75	82,4	0,79
8	CART	10	68,2	0,62	63,4	0,57	72,5	0,67	74,0	0,69
9	CART	5	55,8	0,47	60,0	0,52	65,2	0,58	70,4	0,64
10	RF	150	78,9	0,75	82,0	0,78	85,0	0,82	90,5	0,88
11	RF	100	83,7	0,80	91,6	0,90	89,5	0,87	93,7	0,92
12	RF	75	84,8	0,82	94,9	0,93	89,5	0,87	94,9	0,93
13	RF	60	81,8	0,78	89,9	0,88	88,5	0,86	92,4	0,91
14	RF	50	81,0	0,77	88,2	0,86	86,6	0,84	92,1	0,90
15	RF	40	80,2	0,76	87,5	0,85	86,3	0,84	89,8	0,87
16	RF	30	76,7	0,72	84,4	0,81	79,2	0,75	90,8	0,89
17	RF	20	76,4	0,72	78,0	0,74	79,1	0,75	89,0	0,87
18	RF	10	73,4	0,69	73,5	0,68	78,9	0,75	82,1	0,78
19	RF	5	61,3	0,54	58,8	0,51	74,1	0,69	68,8	0,63
20	Band Means	35	71,3	0,66	88,7	0,86	86,2	0,83	90,5	0,88
21	Band Means & SD	60	72,6	0,67	89,8	0,87	86,0	0,83	92,9	0,91
22	PCA All Dates	20	67,5	0,61	83,1	0,79	84,5	0,81	89,1	0,87
23	PCA Per Image	20	82,1	0,78	93,0	0,91	90,1	0,88	96,2	0,95
24	TCT	30	78,0	0,73	94,5	0,93	84,4	0,81	92,7	0,91
25	Indices	45	67,2	0,61	64,6	0,58	78,6	0,74	87,8	0,85
26	Transforms	110	85,7	0,83	84,1	0,81	92,1	0,90	95,2	0,94
27	Texture Values	35	49,8	0,40	56,2	0,48	55,4	0,47	58,1	0,50

Rodrigues-Galiano et al. (2012) tested RF classification and feature selection on 972 potential variables under similar conditions to this study (multi-seasonal data, Mediterranean climate, classifiers, feature ranking method). Some of the best results achieved in their study were achieved when using reduced feature sets (of around 117). This finding corresponds to the results obtained in this study which indicate that the best classifications were obtained when 75–205 features were used as input. It seems that, for multi-seasonal differentiation of crop types, a feature set of approximately 100 is ideal.

4. CONCLUSION

This paper aimed to determine whether or not feature selection/reduction is of any value when discriminating crops using multi-temporal Landsat8 imagery. The results show that feature selection can, under certain scenarios, prove useful, but is ultimately not beneficial when the SVM classifier is used. The DT, NN, and RF classifiers did, however, improve with variable selection based on CART and RF. Given the effort of feature selection and/or reduction, we recommend using SVM (with the full set of features) for mapping crop types using multi-temporal Landsat8 imagery.

ACKNOWLEDGEMENTS

This work forms part of a larger project titled "Wide-scale modelling of water use and water availability with earth observation/satellite imagery" which was initiated and funded by the Water Research Commission (WRC) of South Africa (contract number K5/2401//4). More information about this project is available in the 2014/2015 WRC Knowledge Review available at www.wrc.org.za. The authors thank the USGS for providing the Landsat8 data, the Western Cape Department of Agriculture for supplying the crop data census, and the Centre for Geographic Analysis for use of the classification and accuracy assessment software.

REFERENCES

Bradski G.R., & Pisarevsky V., 2000. Intel's computer vision library: applications in calibration, stereo, segmentation, tracking, gesture, face and object recognition, 1(1), pp. 2796.

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

Castillejo-Gonzalez, I., & Lopez-Granados, F., 2009. Object and pixel-based analysis for mapping crops and their agro-environmental associated measures using QuickBird imagery. *Computers and Electronics in Agriculture*, 68(2), pp.207–215.

Chang C.C., & Lin C.J., 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp. 27.

Clausi D.A. 2002. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, 28(1), pp 45-62.

Dragut L., Tiede D., & Levick S., 2010. ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. International Journal of Geographic Information Science, 24(6), pp. 859-871.

Gislason P.O., Benediktsson J.A., & Sveinsson J.R., 2006. Random forests for land cover classification. *Pattern Recognition Letters* 27(4), pp. 294-300.

Hsu C.W., Chang C.C., & Lin C.J., 2003. A practical guide to support vector classification. Available at: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Lewis R.J., 2000. An Introduction to Classification and Regression Tree (CART) Analysis. In Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California. 1-14.

Monfreda C., Ramankutty N., & Foley J. A., 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global biogeochemical cycles* 22(1).

Myburgh G., & Van Niekerk A., 2014. Effect of feature dimensionality on object-based land cover classification: A comparison of three classifiers. *South African Journal of Geomatics* 2(1), pp. 13-27.

Qian Y., Zhou W., Yan J., Li W., & Han L., 2015. Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery. *Remote Sensing*, 7(1), pp. 153-168.

Rodriguez-Galiano V.F., M. Chica-Olmo, F. Abarca-Hernandez P.M., Atkinson, and Jeganathan C., 2012. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment* 121(1), pp. 93-107.

Tererai F., Gaertner M., & Jacobs S.M., 2015. Resilience of invaded riparian landscapes: The potential role of soil-stored seed banks. *Environmental Management* 55(1), pp.86-99

Yu Q.P., Gong N., Clinton G., Biging M., Kelly M., & Schirokauer D., 2006. Object-based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery. *Photogrammetric Engineering and Remote Sensing* 72(7), pp. 799-811.

Xin Q., Broich M., Suyker A.E., Yu L., & Gong P. 2015. Multi-scale evaluation of light use efficiency in MODIS gross primary productivity for croplands in the Midwestern United States. *Agricultural and Forest Meteorology* 201(1), pp. 111-119.

Zhang, Y., Wu, L., Neggaz, N., Wang, S. and Wei, G., 2009. Remote-sensing image classification based on an improved probabilistic neural network. *Sensors*, 9(9), pp.7516-7539.