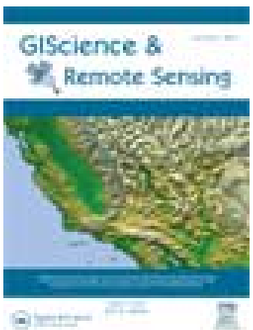


GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: http://www.tandfonline.com/loi/tgrs20

Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application

Stefanos Georganos, Tais Grippa, Sabine Vanhuysse, Moritz Lennert, Michal Shimoni, Stamatis Kalogirou & Eleonore Wolff

To cite this article: Stefanos Georganos, Tais Grippa, Sabine Vanhuysse, Moritz Lennert, Michal Shimoni, Stamatis Kalogirou & Eleonore Wolff (2017): Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application, GIScience & Remote Sensing

To link to this article: https://doi.org/10.1080/15481603.2017.1408892

Accepted author version posted online: 21 Nov 2017.

Publisher: Taylor & Francis

Journal: *GIScience & Remote Sensing*

DOI: 10.1080/15481603.2017.1408892

Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application

Stefanos Georganos[a*], Tais Grippa[a], Sabine Vanhuysse[a], Moritz Lennert [a], Michal Shimoni[b], Stamatis Kalogirou[c], Eleonore Wolff[a]

[a]*Universite Libre de Bruxelles, Institute of Environmental Management and Spatial Planning (IGEAT), Department of Geosciences, Environment and Society (DGES), Avenue Franklin Roosevelt 50, Bruxelles, Belgium, 1050;* [b]*Royal Military Academy, Signal and Image Center, Rue Hobbema 8, Bruxelles, Belgium, 1000;* [c]*Harokopio University of Athens, Department of Geography, Athens, Greece*

******Corresponding Author***, E-mail:** *sgeorgan@ulb.ac.be*

Funding

This work was supported by BELSPO (Belgian Science Policy Office) in the frame of the STEREO III programm – project REACT (SR/00/337).

27

28

29

30

Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application

Abstract This study evaluates the impact of four Feature Selection (FS) algorithms in an Object-Based Image Analysis (OBIA) framework for Very-High-Resolution (VHR) Land Use-Land Cover (LULC) classification. The selected FS algorithms, Correlation Based Feature Selection (CFS), Mean Decrease in Accuracy (MDA), Random Forest (RF) based Recursive Feature Elimination (RFE) and Variable Selection Using Random Forest (VSURF), were tested on the Extreme Gradient Boosting (Xgboost), Support Vector Machine (SVM), K-Nearest Neighbor, Random Forest (RF) and Recursive Partitioning (RPART) classifiers, respectively. The results demonstrate that the selection of an appropriate FS method can be crucial to the performance of a machine learning classifier in terms of accuracy but also parsimony. In this scope, we propose a new metric to perform model selection named Classification Optimization Score (COS) that rewards model simplicity and indirectly penalizes for increased computational time and processing requirements by using the number of features for a given classification model as a surrogate. Our findings suggest that applying rigorous feature selection along with utilizing the COS metric may significantly reduce the processing time and the storage space while at the same time producing higher classification accuracy than by using the initial data set.

Keywords: OBIA, land cover classification, extreme gradient boosting, feature selection, machine learning

1. Introduction

In recent years, with a significant increase in the acquisition of very-high-resolution (VHR) satellite data from Earth Observation (EO) missions such as Pleiades, Quickbird and Worldview, Remote Sensing (RS) information is being collected at various temporal and spatial resolutions. The adequate interpretation of RS data is of vital importance for the extraction of rigorous and robust results that may drive policy

making in various fields as health and epidemiology (Giardina, Franke, and Vounatsou 2015), climate monitoring (Georganos et al. 2017), environmental management (Pradhan and Nampak 2017) or spatial planning (Grippa et al. 2017).

One of the cornerstones of information that can be produced from VHR images is classified maps that most commonly depict land cover (LC) and land use (LU) categories. The classification of these images (the process in which each image pixel acquires a categorical label) is usually attained through Object-Based Image Analysis (OBIA) methods rather than traditional per-pixel approaches (Blaschke et al. 2014). In the OBIA framework, the segmentation step aims to group neighbouring image pixels into objects (segments) that share common characteristics and ideally aim to represent real world objects that are larger than the original pixel resolution (e.g. roofs, streets) (Conchedda, Durieux, and Mayaux 2008). The superiority of OBIA in comparison to traditional per-pixel approaches for VHR imagery has been well established in the remote sensing community (Blaschke 2010). The main advantage is the consideration of not only the spectral but also spatial representation of data, allowing for the utilization of a plethora of morphological, contextual, textural and proximity characteristics of the input data to be effectively used (Cheng et al. 2014).

Although the improvement OBIA offers over the pixel based approach is evident, new challenges have emerged. The number of features which are used as input to the classification has exponentially increased due to each layer of information that would normally be represented by one feature in the pixel based approach, now being depicted through a large set of descriptive statistics (Puissant, Rougier, and Stumpf 2014; Georganos et al. 2017). In addition, OBIA-specific features which are particularly useful for VHR classifications (e.g. textural and spatial measures) can add up to several hundreds, sometimes irrelevant and redundant, features (Laliberte and Rango 2009).

This might have explicit negative effects on the performance of a classifier, an artefact described as the curse of dimensionality, also called 'Hughes phenomenon' as first mentioned by Hughes (1968). In other words, the expansion of feature space over a finite sample size may reduce the accuracy of the prediction. This issue often arises in RS OBIA supervised classifications, as the collection of reference data is an expensive, time-consuming and application-dependent process, thus limiting the sample size. The negative effects of high dimensionality in comparison to the training data size can be present in Machine Learning (ML) classifiers, which are most commonly used for RS classifications such as Support Vector Machine (SVM) and Random Forest (RF) (Pal and Foody 2010). Additional consequences of using datasets with a very large number of predictors may relate to (i) overfitting the classification algorithm due to random variation from irrelevant features captured as meaningful information, (ii) constructing intrinsically complex models making model interpretation a challenging task and finally (iii) requiring additional computational effort, data storage and processing than a simpler, more parsimonious dataset would (Genuer, Poggi, and Tuleau-Malot 2015). It remains unclear to which extend the aforementioned factors may affect the classifiers' predictive capabilities but there is evidence that large RS datasets can interact negatively with respect to the classification accuracy regardless of the training data (Löw et al. 2013; Pal and Foody 2010).

To tackle these issues, feature selection (FS) may be applied, where the classification model is fit on a reduced subset of the most discriminant and informative predictors (Che et al. 2017). Most FS methods operate by calculating a score for each feature or feature subset which allows for ranking predictors based on their discriminative power (Guyon and Elisseeff 2003; Janecek et al. 2008). Pal and Foody (2010), demonstrated that in hyperspectral datasets, SVM classification accuracy tended to decrease as

111 additional features were added in the classification. By applying various FS methods,
112 they showed that the accuracy based on any FS method was always higher than when
113 not performing FS. Demarchi et al. (2014) suggested that reducing dimensionality in
114 machine learning classifiers can be of significant modelling and computational merit in
115 hyperspectral data for urban mapping, without compromising on classification accuracy.
116 Waske et al. (2010) highlighted the importance of selecting relevant features with the
117 SVM classifier for crop mapping applications. In a similar fashion, Löw et al. (2013)
118 demonstrated that, in time-series of multispectral RS data, FS was more beneficial in
119 terms of prediction accuracy than using all features for an SVM classifier. Nonetheless,
120 the degree of success of each feature selection method in terms of number of selected
121 features and classification accuracy, varied between each application and classifier.
122 The concept of applying features selection methods has also been studied in OBIA
123 (Chubey, Franklin, and Wulder 2006; Marpu et al. 2008; Carleer and Wolff 2006).
124 Laliberte, Browning, and Rango (2012) showed that different FS algorithms provided
125 significantly different results in terms of accuracy and feature size for an OBIA
126 classification. Ma et al. (2017) pointed out that appropriate FS could significantly
127 increase accuracy, especially for smaller sample sizes for the SVM and RF classifiers.
128 Albeit, with limited literature, the consensus regarding feature selection in OBIA is still
129 unclear and no standardized procedures exist (Ma et al. 2017). Many studies abstain
130 from applying FS before classification, although the use of large sets of predictors, by
131 assuming that ML classifiers are robust to high dimensional datasets even though may
132 include noisy and correlated features. Studies that have tackled the issue in OBIA, have
133 concluded that RF feature importance, is a FS technique suitable for multiple classifiers
134 such as SVM or RF (Cánovas-Garcia and Alonso-Sarria 2015). Nonetheless, numerous
135 studies have demonstrated RF based importance to be biased in favour of redundant

features, along with problematic permutation schemes in high dimensionality (Strobl et al. 2007; Nicodemus et al. 2010; Gregorutti, Michel, and Saint-Pierre 2016; Nguyen, Huang, and Nguyen 2015). Critically, the impact of FS on different state of the art classifiers other than SVM and RF has not been systematically examined in OBIA literature. This issue is crucial, particularly in studies where different classifiers are evaluated without examining appropriate feature subsets (Kavzoglu, Colkesen, and Yomralioglu 2015; Chan and Paelinckx 2008; Grippa et al. 2017). From the existing literature we can infer that efficient FS can have as much of an impact in increasing classification accuracy as selecting between different classifiers. In this study, we employ a framework for examining the effects of FS algorithms of various complexities and designs, on several machine learning classifiers. In detail, we used Correlation Based Feature Selection (CFS), Mean Decrease in Accuracy (MDA) as derived from the RF classifier, and two wrapper algorithms, Variable Selection Using Random Forests (VSURF) and Recursive Feature Elimination (RFE). The classifiers we tested FS with are the SVM, RF, Extreme Gradient Boosting (Xgboost), Recursive Partitioning (RPART) and K-Nearest Neighbour (KNN), in an OBIA environment. The rationale behind the selection of these algorithms is given in section 3.3.

Finally, most of the research towards the benefits of feature selection is focused towards its impact on classification accuracy. So far, no studies have tried to quantify additional benefits such as data storage, processing time and model simplicity which can be of equal importance, especially in large scale applications. Thus, we propose a new metric to select the most appropriate model which considers not only the accuracy of the classification but also the computational burden, data storage and model interpretation by using as a surrogate the number of features included. The objective is therefore to investigate if more sophisticated FS algorithms can improve classification accuracy

while maintaining a small set of predictors in comparison to the widely used RF feature rankings and to promote parsimonious model selection when deciding between different classifiers and feature subsets. To our knowledge, this is the first study that tackles advanced feature selection techniques in multispectral VHR data in urban environments and quantifying issues related to data storage and processing time in evaluating classification performance.

2. Materials and Methods

2.1 Case Study and Software

The classification scheme is applied in Ouagadougou (615 km^2), the capital of Burkina Faso and a major Sahelian city. Ouagadougou has been facing rapid urban growth for the last two decades, with a constantly increasing rate of urbanization, which doubled between the years of 1980 and 2000 (United Nations Urbanization Prospects, 2014). Consequently, the urban environment has been rapidly shaped, combining heterogeneous features, from planned neighbourhoods in the core areas to an unregulated expansion in the peri-urban zones (Figure 1).

2.2 Data, Image Segmentation and Training Samples

Our initial dataset consists of stereo Worldview-3 imagery (0.5 m) with three bands in the visible (RGB) and one in the near infrared (NIR) wavelengths, collected in October 2015. Additionally, a photogrammetric normalized digital surface model (nDSM) layer was derived from these stereo Worldview-3 pairs at the same spatial resolution. For segmentation, we utilized an open source processing chain developed by Grippa et al. (2017), based on an integration of GRASS GIS (Neteler et al. 2012) and Python programming environment. The four Worldview bands were used as input to a region

185 growing (RG) segmentation algorithm and an Unsupervised Segmentation Parameter
186 Optimization (USPO) approach was implemented (Lennert et al. 2016) to evaluate the
187 quality of the segmentation and to calibrate its parameters as suggested by Johnson et al.
188 (2015). In the RG implementation of GRASS GIS, the most critical parameter that
189 affects the size and shape of the segments is the *threshold* which ranges between *0* to *1*,
190 with *0* leading to the creation of no segments, while *1* describes the situation where the
191 whole image is segmented in a single object. The USPO suggested a value of 0.012 for
192 the image of Ouagadougou.
193 We considered 169 features as initial input to the classification. These include object
194 descriptive statistics (mean, median, min, max, range, standard deviation, sum,
195 coefficient of variation, first quantile, third quantile) for each raster layer (VNIR,
196 nDSM, Brightness, NDWI, NDVI, angular second moment (ASM) of the Gray Level
197 Co-occurrence Matrix (GLCM) for each spectral band (5 and 11 windows), as well as
198 object compactness, perimeter, area and fractal dimension.
199 The classifications were trained with 10 LULC categories (Table 1). A category
200 describing unclassified land due to shadow was also included. The collection of training
201 data labels was performed using random sampling. For built-up areas and roads
202 stratified random sampling, based on Open Street Map (OSM) information, was utilized
203 as a complementary measure to make the sampling more efficient. Through detailed
204 visual interpretation we labelled in total 1880 object samples which were randomly split
205 into a training set (50%) and a validation set (50%).
206
207

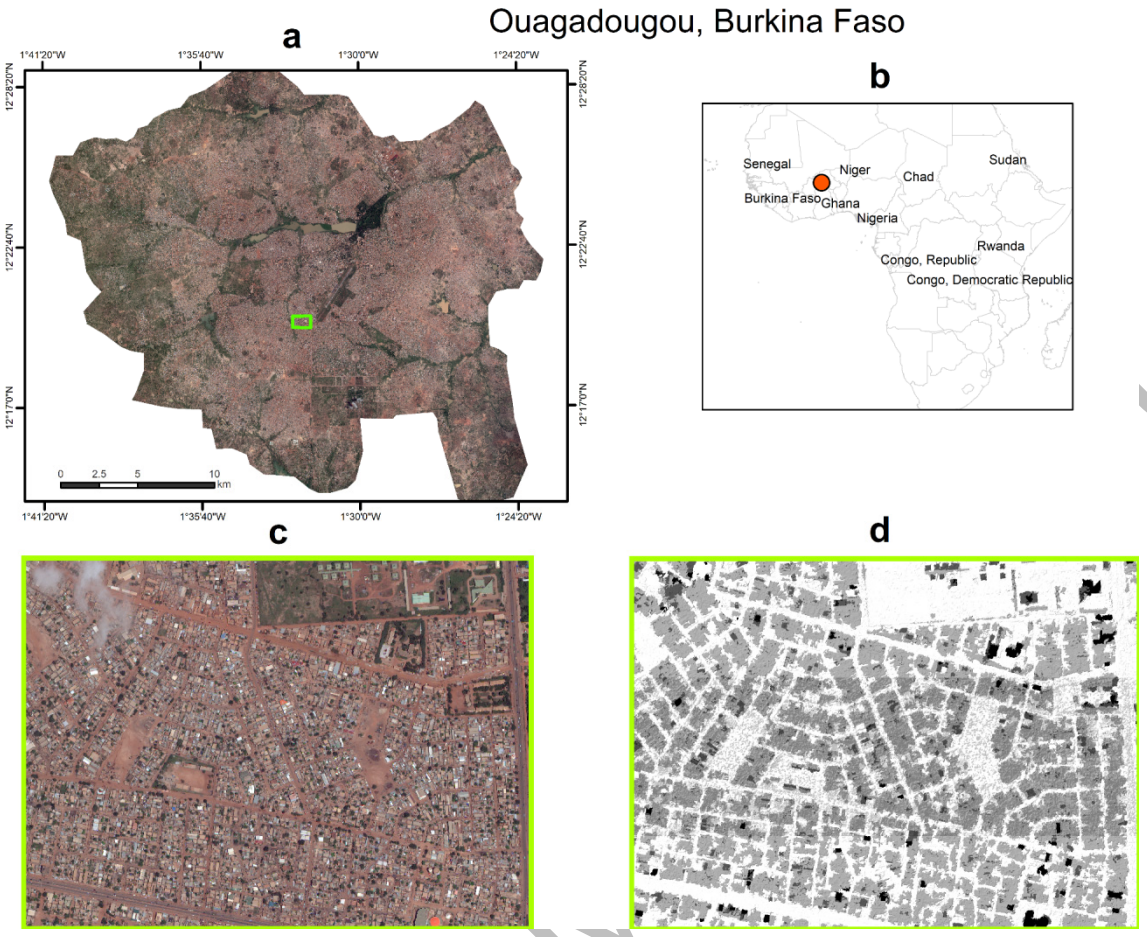


Figure 1 Ouagadougou, Burkina Faso. (a) RGB composite of the whole study area, (b) reference map of Ouagadougou within the African continent, (c) high density built up area and (d) nDSM for the same area of interest.

Table 1 Classification scheme, train and test data for Ouagadougou.

LULC	Training Set Size	Test Set Size	Segments Median Area (m^2)
Buildings (BU)	157	157	273
Swimming Pools (SP)	68	69	119
Asphalt (AS)	56	56	30240
Brown/Red Bare Soil (BS)	90	91	55450
White/Grey Bare Soil (WS)	72	72	11300
Trees (TV)	77	77	380
Mixed Bare Soil/Vegetation (MV)	76	75	24390

Dry Vegetation (DV)	70	70	30420
Other Vegetation (OV)	139	141	7464
Inland Waters (IW)	75	75	2073
Shadow (SH)	58	59	72

214

215 ***2.3 Classification Algorithms***

216 To evaluate the functionality of FS, five classifiers were selected: RF, SVM, Xgboost,
217 RPART and KNN. RF and SVM are well established in the RS community and
218 excessive reviews of their applications can be found in (Mountrakis, Im, and Ogole
219 2011; Belgiu and Drăgu 2016). KNN is among the simplest ML algorithms and its
220 implementations have been used in OBIA (Kavzoglu, Colkesen, and Yomralioglu 2015)
221 along with RPART, a traditional Classification, and Regression Tree (CART) algorithm
222 (Grippa et al. 2016). Xgboost is a new implementation of boosting CART algorithms
223 (Chen and Guestrin 2016) that has been very popular in machine learning community,
224 usually outperforming benchmark classifiers such as SVM or RF (Song et al. 2016;
225 Mustapha and Saeed 2016). The main difference between Xgboost and existing
226 boosting techniques is that it provides a stronger regularization framework which helps
227 to avoid overfitting (Xia et al. 2017). The aforementioned classifiers were applied
228 through their most well-known implementations in R software, namely the '*e1071*'
229 (Meyer et al. 2014), '*xgboost*' (Chen and He 2015), '*randomForest*' (Liaw et al. 2002),
230 '*kknn*' (Schliep and Hechenbichler 2014), and '*rpart*' (Therneau, Atkinson, and Ripley
231 2015) packages. The parameters of the classifiers were optimized by performing grid
232 search (GS) through cross-validation as depicted in Table 2. In GS, the system loops
233 through a predefined subset of potential parameter values and the model with the
234 smallest error is selected as optimal. The most important parameters for Xgboost is the

number of trees (*nrounds),* the learning rate (*eta)* and the depth of each tree *(depth)*. They control the complexity and the fitness of the model. The rest of the parameters are complementary and help to avoid situations of over or under-fitting. The other classifiers require just less parameters to be optimized. For RF, an appropriate number of trees (*ntrees)* and number of randomly selected predictors at each tree node (*mtry)* are specified. For SVM, a regularization or complexity parameter (C) and the radial kernel search parameter (*gamma)* that minimize cross validation error are selected. For KNN, the number of nearest neighbors *(k)* that are considered for labeling a data point in feature space is defined.

Table 2 Optimized parameter values for the Xgboost, RF and SVM, and KNN models.

Classifier	**Parameters**	**Value**
Xgboost	nrounds	600
	eta	0.015
	gamma	0
	depth	4
	child_weight	2
	subsample	0.4
	column_sample	0.3
RF	ntree	2000
	mtry	1/3 of input features
SVM	gamma	0.003
	C	15
KNN	k	10

2.4 Feature Selection

FS techniques can be divided into three general categories: i) filters, ii) embedded techniques and iii) wrappers. Filter methods represent the simplest, fastest and most generic approaches for selecting relevant attributes. They do not require any learning algorithm but rather rank and select features and feature subsets based on statistical measures such as correlation. In that manner, they attempt to allow only the most important variables to manifest. Their main disadvantage is that they do not take model prediction and feature interaction under consideration. On the other hand, wrappers use model prediction of a machine learning algorithm to decide on the set of most suitable features. By creating models with different numbers of features as input and evaluating them through an objective function and cross validation, the most discriminant variables can be identified and ranked. However, they require higher computational resources than filter techniques and may inevitably resort to heuristic searching methods (i.e. stepwise search) in high dimensional data while Strobl et al. (2007) criticized them for a tendency to overfit the training data. Finally, the embedded selection can be described as a feature ranking method that is built-in the learning process of the selected classifier. The algorithms CART, Lasso and Ridge are examples of such functionalities. Embedded techniques are more computationally efficient than wrappers as they only need one model to produce results, but because of that, they are limited to the biases and weaknesses of a specific classifier. Moreover, they take feature interaction under consideration only partially compared to wrappers, as the results are derived from a model including all variables. In this study, we considered four FS techniques covering the main categories described previously.

The application of all FS methods was performed in R software by their respective package implementations.

2.4.1. Correlation-based feature selection (CFS)

The CFS method is a popular filter approach in which feature subsets are evaluated by maximizing the dependency of a feature subset with the target class, while minimizing the intercorrelation within the subset (Yu and Liu 2003):

$$g(F) = \frac{FCi}{IFCi}, \quad (1)$$

where *F is* a candidate feature subset, *g(F)* is the score of the evaluation, *FCi* is the average correlation between subset F and the dependent variable and *IFCi* is the average inter-correlation within the subset *F*. Given an initial set of input, CFS will identify a set of features that best optimizes *g(F)*. To derive a relative ranking from this method, we used CFS sequentially by removing the CFS proposed feature set from the initial input each time and re-applying the algorithm on the reduced variable set until no features were left. In our study, conditional Entropy is used to evaluate the correlations and best first search strategy is employed to explore the potential feature subsets.

2.4.2. Random forest mean decrease in accuracy (MDA)

RF's Mean Decrease in Accuracy is a feature ranking approach used in several RS studies (Pal and Foody 2010; Rodriguez-Galiano et al. 2012; Duro, Franklin, and Dubé 2012). It utilizes the capability of the RF to evaluate features during the training stage. Each tree in the RF is trained using a random subset of the input features while one third of the training data, the out-of-bag (OOB) sample, is kept for validation. This leads to the calculation of the internal measure of error in RF, the OOB error, which is

consequently used to assess the importance of each predictor and perform selection as the classifier is being trained (Breiman 2001). MDA is the most frequently used measure to do so and it is produced in the following way: the relationship between a feature and a dependent variable is being decoupled by randomly permuting the features values (Belgiu and Drăgu 2016). This should result in changes in model accuracy if the feature is important, or no changes if the feature is insignificant and thus feature rankings can be derived based on their impact on model accuracy.

2.4.3. Random forest recursive feature elimination (RFE)

Recursive Feature Elimination (RFE) is a well-established wrapper method to derive discriminant features rankings in RS datasets (Pal and Foody 2010; Ma et al. 2017). The approach followed in this study is an iterative backwards elimination process in which a ML classifier is fit with the initial input and afterwards, the variable which is the least useful for changing an objective function such as the reduction of a cross-validation metric is removed (Kuhn et al. 2014). Then, a model is refit on the reduced feature subset and again, the least significant predictor is eliminated. The algorithm continues until all variables have been removed and consequently, ranked. To reduce the potential model combinations, as this is a NPP-complete algorithm, heuristics are often used. In our case, we used a RF model to fit the data and we removed the least important feature based on the RF MDA importance rankings at each fitted model.

2.4.3. Variable Selection Using Random Forest (VSURF)

VSURF is a wrapper based algorithm which uses RF as the base classifier (Genuer, Poggi, and Tuleau-Malot 2015) and operates in the following way: First, the features are ranked based on MDA rankings over typically 50 permutations using raw instead of

316 normalized values. Afterwards, features that have zero or almost zero contribution to
317 the classification are removed. The remaining variables are tested in a descending
318 manner of nested RF models and the smallest most accurate model is retained. From
319 there, an ascending stepwise function is optimized for removing redundancy. In detail,
320 if the introduction of a feature provides only a small or no decrease in OOB error, then
321 the said feature is rejected. The rejection threshold is defined based on an objective
322 function that minimizes OOB. At the final step of the procedure, a list of the most
323 discriminant features is obtained. By ranking these variables subsets backwards, a
324 ranked list of features can be derived for the whole dataset.

325 ***2.3.3. Classification optimization score (COS)***

326 In many large-scale applications, an additional factor which can be influential in
327 selecting an appropriate classification model is the number of features selected. The
328 amount of input features in a classifier can serve as a robust surrogate for computational
329 time, model complexity, data storage and processing. For example, an overall
330 classification accuracy of 80% with ten features as input might be considered more
331 beneficial than a classification of 80.4% obtained with a total of 50 selected features, if
332 parsimony is one of desirable goals.

333 Consequently, to quantify classification performance by means of parsimony, we
334 propose a Classification Optimization Score (COS) based on a variant of the F-score
335 (Witten et al. 2016). The F-score is an accuracy metric that has been used to evaluate
336 accuracy per-class (Grippa et al. 2017) and segmentation performance (Johnson et al.
337 2015), among others. The metric allows for weighting two components and in our case,
338 can be derived as

$$COS = (1 + a^2)\frac{N_n * OA_n}{a^2 * N_n + OA_n}, \quad (2)$$

where N_n is the normalized value of the number of features of a classification model (a single feature having a score of 1, and all features combined, a score of 0), OA_n is the normalized classification accuracy across all classifiers and FS methods (lowest classification accuracy receives a score of 0 and highest a score of 1), and a is the relative weight factor. Since classification accuracy is presumably a more beneficial factor than the number of features, we propose a value of 3, meaning that the OA_n will receive more weighting in the formula. In such manner, a parsimonious model ranking may be derived that takes under consideration the FS method, the specific classifier and the number of features that were used, at the same time.

3. Results

3.1. Classification accuracy trends

The accuracy of each classifier varied as a function of the type and the number of selected features. To explore trends, we computed the Overall Accuracy (OA) for each suggested set of features, starting with the first 5 and incrementing by 1 until a model with all variables was reached. Figure 2 depicts the variability of OA based on feature rankings of the four different FS methods suggested for the Xgboost classifier. The wrapper algorithms reached maximum OA earlier, with smaller feature subsets. In detail, VSURF peaked at 79.5% and 31 features, while RFE at 79.8% with 23 features, respectively. MDA performed poorly in comparison to the other three methods, as its reached its maxima on a model consisted of 136 features. By using VSURF and RFE, we can achieve statistically significant improvements when comparing with a model using all features, as suggested by a one tailed McNemar test of similarity at the 95% significance level (Table 3).

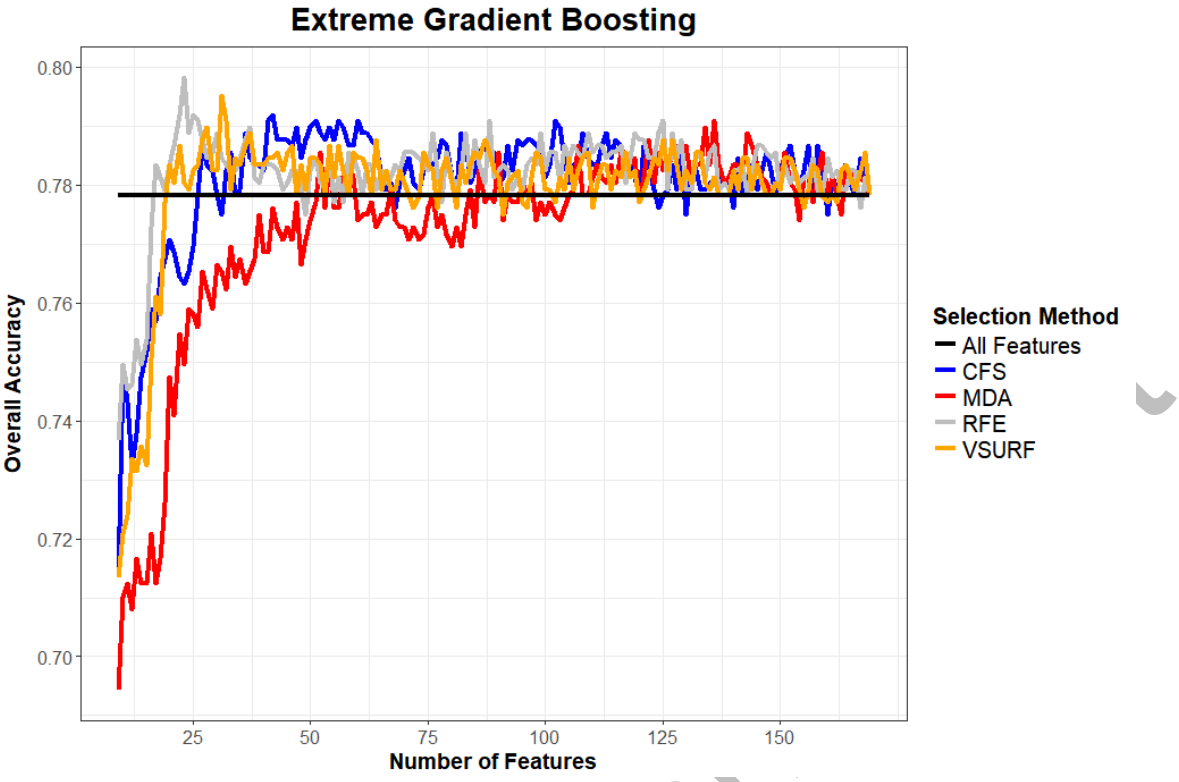


Figure 2 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the Xgboost classifier. The model including all features considered for the classification is represented by the flat horizontal line.

In the case of SVM, the sensitivity of the classifier to high dimensionality was more evident. Critically, the RF feature rankings did not produce any significant improvement with any subset. Surprisingly, the best FS method by far for SVM was CFS as it reached the maximum OA (80.1%) with 37 variables, which was also the highest accuracy that was reached including all different classifiers examined in this study (Figure 3). VSURF performed better than RFE (79.7%, 79.2%, respectively), but nonetheless, both algorithms produced improvements over the full feature model with relatively small feature subsets (75, 32 features, respectively). Undertaking rigorous FS appears to be quite beneficial for SVM, as most CSF and VSURF subsets produce large and significant deviations in OA (~2%) compared to using all features.

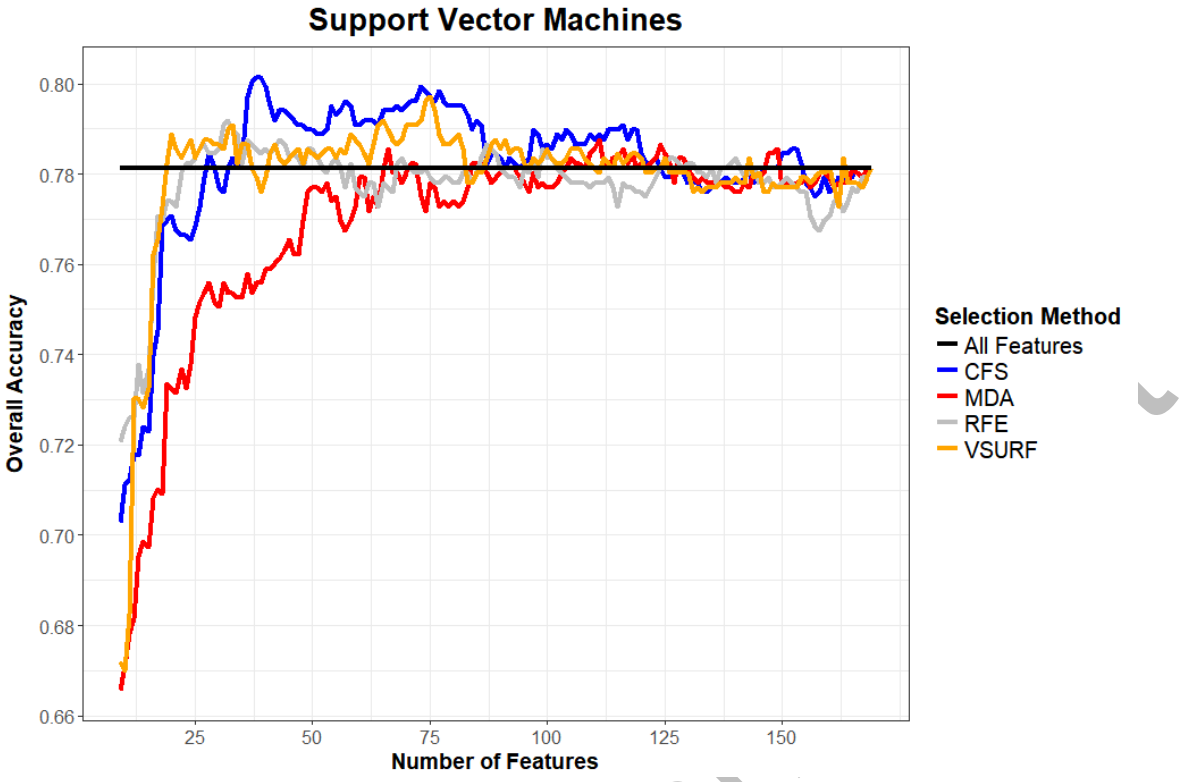


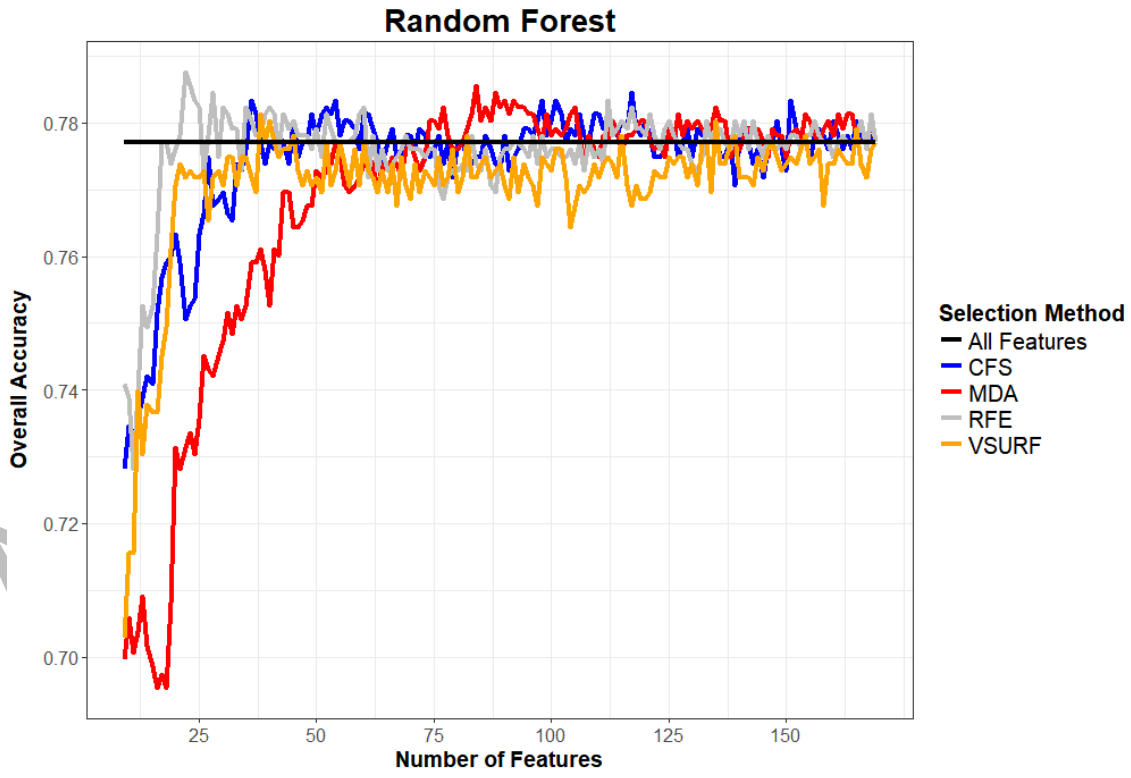
Figure 3 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the SVM classifier. The model including all features considered for the classification is represented by the flat horizontal line.

RF appears to be a robust classifier with respect to increasing feature space in this study (Figure 4) as none of the FS methods produced significant improvements in OA over the full feature model. Here, MDA selection produced better results than in comparison to other classifiers (78.6%, 84 features) but as a disadvantage it peaked with a large number of included predictors. From this we can infer that RF is quite resistant to the addition of redundant and irrelevant features. Finally, the CFS selection performed adequately, especially considering the big computational differences than a more complicated wrapper algorithm such as VSURF or RFE.

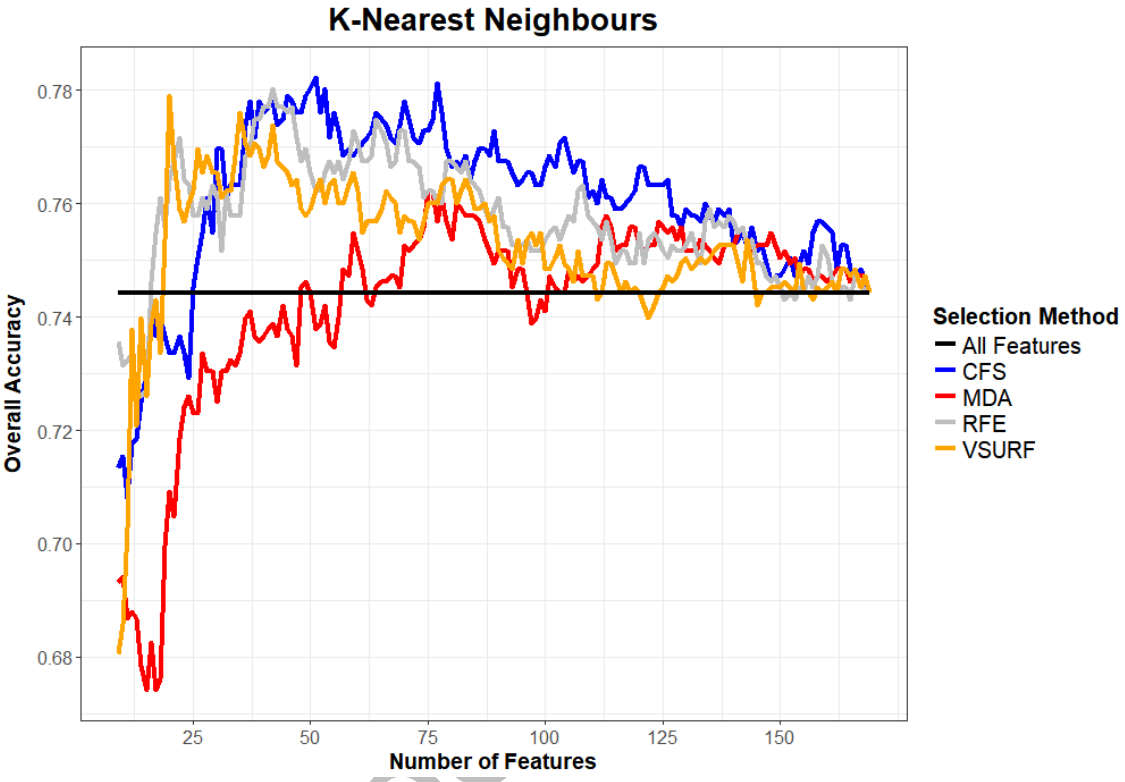
The KNN classifier appears the most sensitive to the results of applying FS (Figure 5). CFS, RFE and VSURF selection signified very high OA improvements with various

small feature subsets. Impressively, VSURF produced a 4% increase in OA with only a subset consisted of 20 features whereas RFE achieved the highest accuracy with 80% and 42 variables. On the contrary MDA did not manage to achieve a similar performance proving rather weak for the KNN classifier. Nonetheless, FS is the most beneficial for KNN which could be explained by its fundamental operating procedure – training on nearest neighbors in feature space. At about 100 features, the trends stabilize irrespective of the FS method.

Considering the classification accuracy trends, RPART proved to be the most robust classifier to high dimensionality, from those that were investigated (Figure 6). None of the proposed FS methods managed to produce significant improvements concerning classification accuracy. However, except for the RF based feature ranking, similar results can be attained with minimal feature subsets.



407 Figure 4 Relationship between Overall Accuracy (OA), number of features and feature
408 rankings based on each FS method for the RF classifier. The model including all
409 features considered for the classification is represented by the flat horizontal line.
410



411

412 Figure 5 Relationship between Overall Accuracy (OA), number of features and feature
413 rankings based on each FS method for the KNN classifier. The model including all
414 features considered for the classification is represented by the flat horizontal line.
415

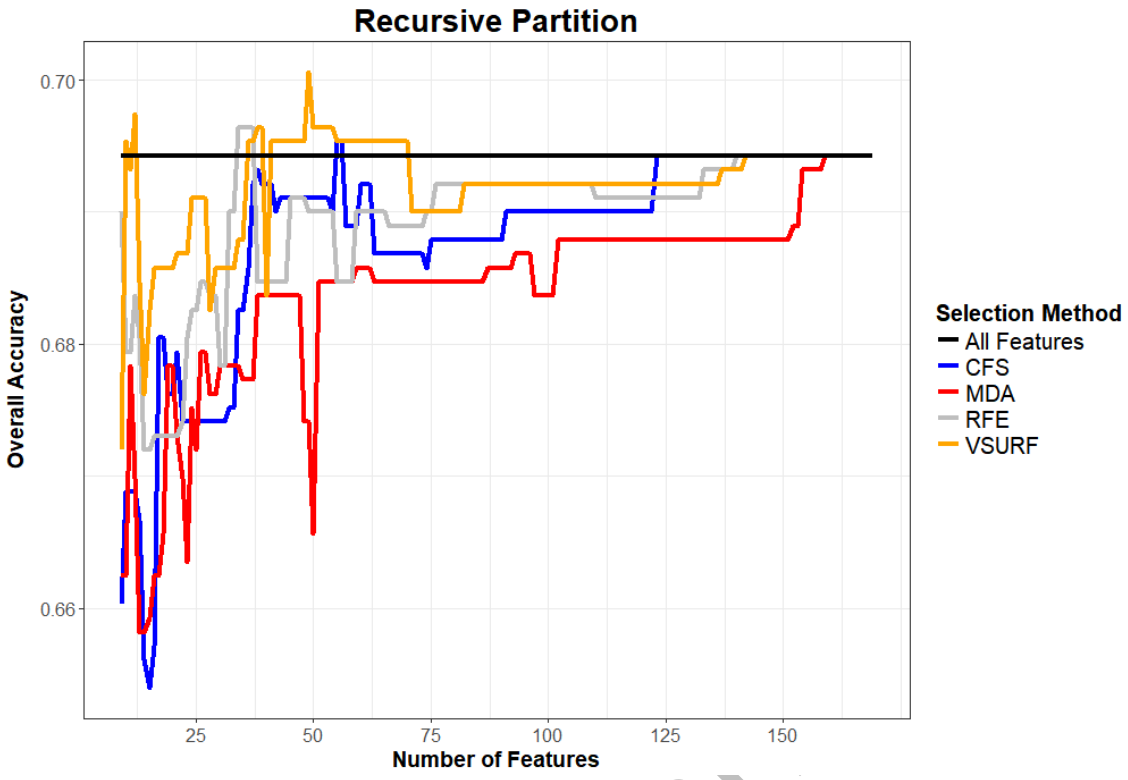


Figure 6 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the RPART classifier. The model including all features considered or the classification is represented by the flat horizontal line.

Table 3 Peak Overall Accuracy (%) for each classifier, based on different FS methods. Numbers in parenthesis () depict the number of features to achieve the reported result. Values with * imply significant improvement over the model with all features, as suggested by a one tailed McNemar test of similarity.

Algorithm	CFS	MDA	RFE	VSURF	All Features
Xgboost	79.2 (42)	79.1 (136)	79.8* (23)	79.5* (31)	77.8 (169)
RF	78.5 (117)	78.6 (84)	78.9 (22)	78.1 (38)	77.7 (169)
SVM	80.1* (37)	78.8 (111)	79.2 (32)	79.7* (75)	78.1 (169)

KNN	78.2* (51)	76.2* (75)	78.0* (42)	77.9* (20)	74 (169)
Rpart	69.5 (53)	69.4 (123)	69.6 (34)	70.1 (49)	69.4 (169)

426

427 ***3.2. Selected features among each method***

428 To investigate the different feature rankings, Table 4 presents the first 30 features
429 derived from each FS application, which arguably have the most impact on the
430 classification performance as shown in the previous section. The results demonstrate
431 that except for the NDVI, there are significant differences between the selection
432 obtained by each FS algorithm. The CFS method focuses on information from the
433 green, near infrared, NDVI and nDSM and geometrical features such as compactness
434 and fractal dimension. The wrapper algorithms VSURF and RFE produce more
435 sophisticated selections, by including textural features and statistical measures such as
436 the coefficient of variation and standard deviation to describe the distribution of training
437 data within the objects. Although these variables might not be particularly predictive by
438 themselves alone, they increase model accuracy through their interactions with other
439 variables (Guyon and Elisseeff 2003; Gheyas and Smith 2010). It seems that these
440 algorithms are particularly efficient in detecting these presumably beneficial
441 interactions since they construct several models for training. Finally, RF selection
442 appears to produce redundant rankings using very similar measures in consequent
443 rankings (e.g. Mean, Median).

444

445 Table 4 Top 30 features according to each feature selection method.

CFS	MDA	RFE	VSURF
compact_circle	NDVI_first_quart	NDVI_mean	NDVI_third_quart
fd	NDWI_third_quart	nDSM_median	NDVI_mean
opt_green_first_quart	NDWI_median	NDVI_median	NDVI_median

opt_green_median	NDWI_mean	NDVI_third_quart	NDVI_perc_90
opt_green_perc_90	NDWI_perc_90	NDVI_first_quart	NDVI_first_quart
opt_nir_min	NDVI_mean	NDWI_perc_90	NDWI_third_quart
opt_nir_max	NDVI_median	opt_nir_first_quart	NDWI_median
opt_nir_mean	Brightness_perc_90	NDWI_third_quart	NDWI_mean
opt_nir_coeff_var	opt_red_median	nDSM_first_quart	Brightness_first_quart
opt_nir_first_quart	Brightness_third_quart	NDVI_perc_90	NDWI_perc_90
opt_nir_median	opt_red_first_quart	opt_nir_mean	opt_green_median
opt_nir_perc_90	NDWI_first_quart	NDWI_coeff_var	nDSM_median
opt_red_max	Brightness_mean	texture_asm5_blue_ASM_min	Brightness_median
opt_red_first_quart	Brightness_median	opt_nir_median	opt_green_first_quart
opt_red_median	opt_red_third_quart	nDSM_mean	opt_nir_first_quart
NDVI_min	Brightness_first_quart	texture_asm5_blue_ASM_first_quart	texture_asm5_blue_ASM_first_quart
NDVI_mean	opt_red_mean	compact_circle	nDSM_mean
NDVI_stddev	NDVI_third_quart	opt_nir_max	NDVI_stddev
NDVI_first_quart	nDSM_median	NDWI_median	Fractal_d
NDVI_median	NDVI_sum	NDWI_mean	compact_circle
NDVI_third_quart	opt_red_perc_90	nDSM_third_quart	texture_asm5_green_ASM_first_quart
NDVI_perc_90	compact_circle	NDVI_min	opt_nir_coeff_var
Brightness_first_quart	NDWI_sum	NDVI_stddev	Brightness_third_quart
Brightness_median	opt_green_perc_90	Brightness_first_quart	NDWI_first_quart
nDSM_coeff_var	NDVI_min	NDWI_stddev	opt_nir_mean
nDSM_median	opt_green_mean	opt_green_median	nDSM_third_quart
nDSM_third_quart	nDSM_third_quart	opt_nir_coeff_var	opt_nir_max
NDWI_max	opt_green_median	Brightness_median	opt_nir_median
NDWI_stddev	perimeter	opt_nir_third_quart	opt_green_mean

3.3. Classification optimization score (COS)

Figure 7 demonstrates the positive relationship between the increase in number of features and the intensification of the computational and data managing burden. Table 5 depicts the highest COS scores as derived by applying the F-score based formula, among all classification algorithms, FS methods and possible number of features (a total of 3300 models). The Xgboost classifier with RFE is suggested as the best model in this

study where a very high classification accuracy is reached with a set of only 22 predictors (Figure 8). If we would decide to use the best model to apply the classification in the whole image of Ouagadougou, we would save more than 24 hours of computing time and 27 GB of storage space just to extract feature information for the rest of the objects other than the training set, arguably important recourses. Moreover, the transferability and interpretation of the model becomes easier. The second-best model in the COS rankings is achieved using the SVM classifier and the filter method, CFS.

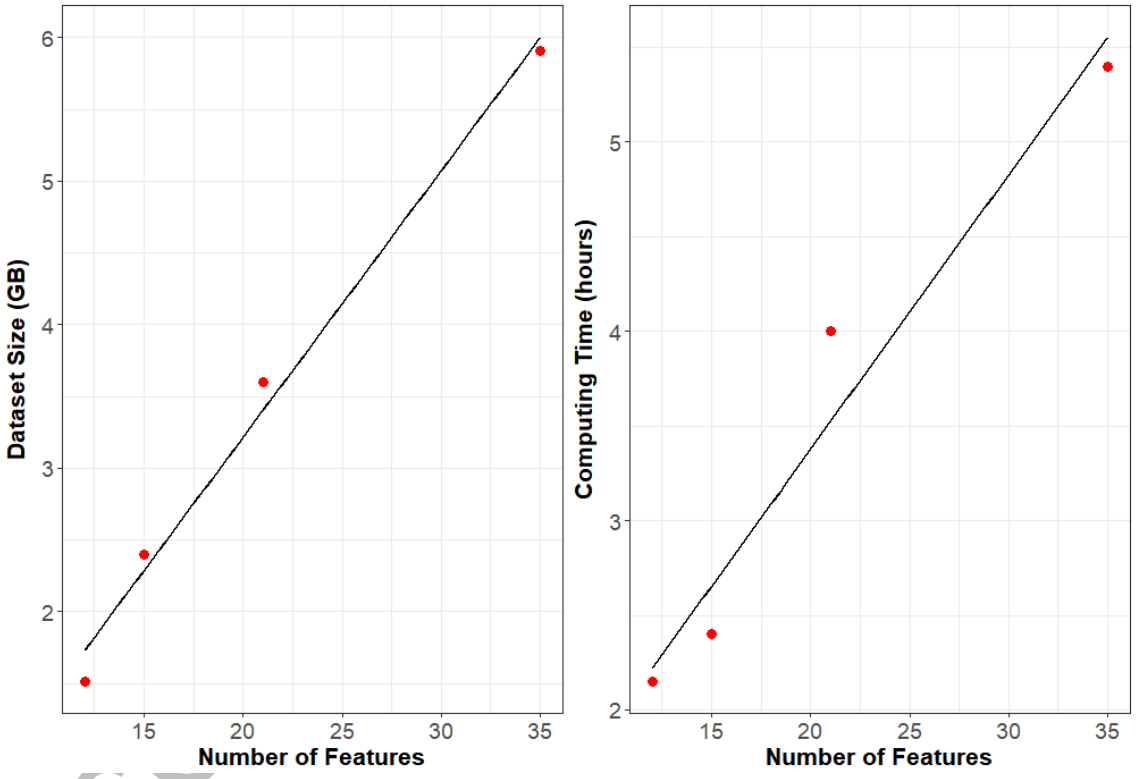


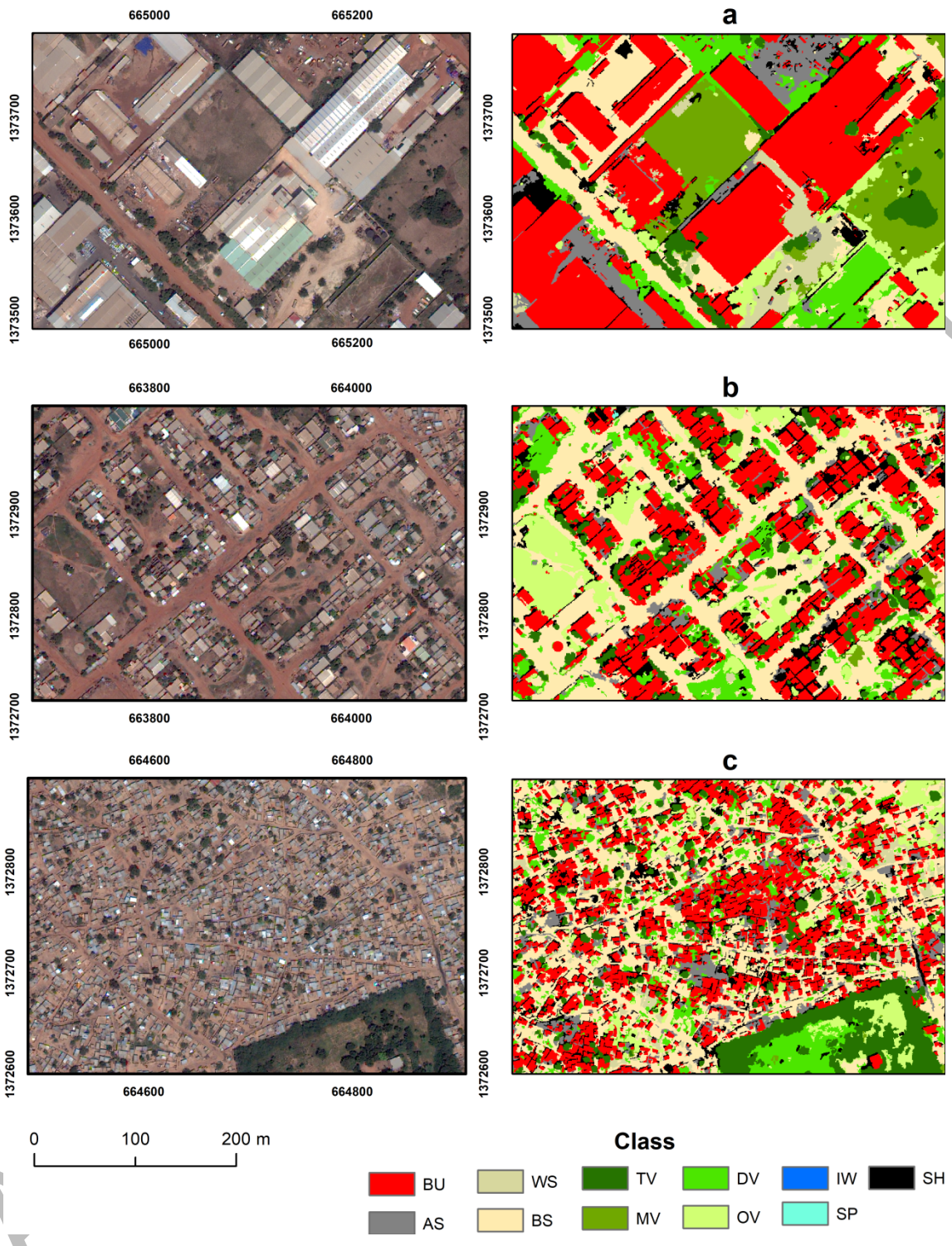
Figure 7 Relationship between the number of features in the classification and the storage size (left) and the computing time (right) to extract the selected predictors for the whole dataset of Ouagadougou, at an object based format. The study area consists of ~ 12 million segments.

In general, CFS rankings are the most beneficial for SVM while both VSURF and RFE appear to be very potent for Xgboost variable selection. MDA importance as a FS

method, underperformed in comparison to the rest. Although the approach of using all models considered for classification for scoring in the same formula is biased towards the classifier who is in general more accurate than others (e.g. SVM, Xgboost), it can alternatively be applied for each classifier separately. Since our aim is to evaluate FS methods for each classifier separately but ultimately decide on the best model between different classifiers, we found the first approach more intuitive and informative.

Table 5 Top 10 highest ranked models according to the COS

Classification Model	COS	Overall Accuracy	Number of Features
RFE_XGBOOST	0.982	0.798	23
CFS_SVM	0.980	0.801	38
CFS_SVM	0.975	0.800	37
CFS_SVM	0.975	0.801	39
RFE_XGBOOST	0.972	0.792	22
VSURF_XGBOOST	0.971	0.795	31
RFE_XGBOOST	0.970	0.792	25
CFS_SVM	0.970	0.797	36
CFS_SVM	0.970	0.799	40
VSURF_SVM	0.969	0.789	20



476

477 Figure 8 Classification results from the highest ranked model of the COS metric
478 (Xgboost, 22 features) for different regions in Ouagadougou. a) Industrial, b) planned
479 residential and c) unplanned residential neighborhoods.

4. Discussion

The results of this study agree with previous research which concluded that the selection of an appropriate FS method can have a crucial impact on the performance of the classification using several ML algorithms (Laliberte, Browning, and Rango 2012; Cánovas-Garcia and Alonso-Sarria 2015; Pal and Foody 2010). However, our understanding of the sensitivity of each classifier to the increase of feature space might be hindered if we use inappropriate FS methods such as with MDA in this research case. As such, the consequences of the curse of dimensionality, overfitting and noisy features might appear only with specific FS methods. In fact, the added value of using a wide set of features might not manifest if no or inappropriate FS is performed. As an example, if we would follow an MDA selection in our dataset, we would assume that SVM or even KNN is quite robust to the previously mentioned phenomena. Notably, SVM performance appears to significantly increase when only the most discriminant features are selected with other FS techniques such as CFS or VSURF.

Cánovas-Garcia and Alonso-Sarria (2015) used four different FS methods, namely Average Correlation, Maximum Correlation, Random Forest importance based on Mean Decrease in Gini (MDG) and Jeffries-Matusita distance and concluded that RF importance rankings are promising regarding FS as it was the better performing method. It should be noted that the rest of selected FS methods in their study were simple filter techniques assuming data normal distribution and/or linearity, which potentially can be the reason for their underperformance in comparison with MDG. Our findings imply that MDA has difficulties to detect discriminant features in heavily redundant datasets, which negatively affects classification accuracy, model parsimony and processing requirements while the true manifestation of underlying issues related to overfitting and high dimensionality might be visible when more advanced/relevant methods are used.

Li et al. (2016) and Ma et al. (2017a) concluded that RF was very robust when the feature space was increased whereas Cánovas-Garcia and Alonso-Sarria, (2015) demonstrated SVM was negatively affected. The results agree with the mentioned studies as RF was indeed, very resistant to the addition of features with respect to classification accuracy while SVM was negatively affected. Nonetheless, in the case of RF, the wrapper RFE selection managed to increase OA marginally by using very small and discriminant feature subsets. Therefore, the selection of an appropriate FS method can be critical in order to draw more realistic conclusions. Similarly, Xgboost appeared to benefit from wrapper based FS such as RFE and VSURF by almost 2% gains in OA in some cases. As boosting classifiers in general have issues with overfitting, FS can be of merit for maximizing their performance.

Another topic this study tackled was the implementation of feature subset methods such as CFS, to provide a type of feature ranking. So far in the OBIA literature, CFS is considered a one way feature subset evaluation method, which provides only one set of features, regardless of the input data (Li et al. 2016; Ma et al. 2017). In this study, we demonstrated, that by using CFS in a wrapper-like fashion, in which the algorithm is applied iteratively, we can rank feature subsets which consequently, is a form of individual feature hierarchy. The results implied that CFS can be a very successful feature ranking method considering its complexity and that future research is needed to validate these results.

It is evident that there is no 'one method fits all' concept. CFS, being a relatively simple filter method consistently overperformed in specific classifiers such as SVM, while maintaining good results with others. The algorithms VSURF and RFE used RF as the base classifier, and as such, are naturally biased towards CART-based methods. In

general, VSURF was more consistent with RPART, Xgboost and SVM, while RFE with Xgboost and RF. This implies that the selection of FS methods should be done in a way that captures the whole variability of the classifier characteristics that are considered for modelling, to remove biased selections in favour of specific classifiers. Arguably, the results of this study might be dependent on the sample size or the number of classes, but both factors were selected to represent a typical application in VHR OBIA RS and thus, to provide practical solutions.

The proposed COS was effectively used to increase model performance while at the same time reducing its complexity. Given that the use of big data is becoming increasingly more common in RS, developing methods for reducing computational demands and for selecting only the most predictive yet simple models appears to be of benefit. The COS can be employed both as a method to evaluate different FS methods and classifiers but also to reduce the computational burden of a classification in large datasets. A point of further investigation would be to examine the applicability of the proposed metric in different RS applications with varying data sizes, computational, requirements and input features.

Further research should consider the application of additional FS algorithms such as Genetic Algorithms (Li et al. 2015; Van Coillie, Verbeke, and De Wulf 2007), Regularized Random Forests (Adam et al. 2017), as well as feature importance derived from boosting classifiers such as Xgboost (Chen and Guestrin 2016) or Adaboost (Chan and Paelinckx 2008). Finally, alternative implementations of popular classifiers such as Rotation Forest (Kavzoglu and Colkesen 2013), and Support Vector Machine Ensembles (Huang and Zhang 2013) should be investigated towards their sensitivity to FS.

5. Conclusion

In this study and in an OBIA environment, the sensitivity of various ML classifiers to high dimensionality was examined and a measure to perform parsimonious model selection was proposed. Through selecting subsets of the most discriminant features most classifiers improved their performance, while requiring less computational resources and data management. Xgboost and SVM were the best performing classification methods in terms of accuracy. However, Xgboost required only half of the features than SVM for equivalent performance. MDA underperformed in comparison to the other FS methods and should be viewed more critically in situations where a lot of correlated and/or noisy features are used as input.

The proposed COS index provided practical solutions in the task of model selection by rewarding simple models that perform well. The metric can be applied in cases where the number of explanatory variables is particularly large or when computational requirements need to be reduced. As such, the COS can be considered as a complementary or alternative evaluation measure to solely accuracy based metrics such as Overall Accuracy or the Kappa index.

Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments and suggestions which improved the quality of the manuscript.

Disclosure statement

The authors report no conflicts of interest.

References

Adam, Elhadi, Houtao Deng, John Odindi, Elfatih M Abdel-Rahman, and Onisimo Mutanga. 2017. “Detecting the Early Stage of Phaeosphaeria Leaf Spot

Infestations in Maize Crop Using In Situ Hyperspectral Data and Guided Regularized Random Forest Algorithm." *Journal of Spectroscopy* 2017. Hindawi Publishing Corporation.

Belgiu, Mariana, and Lucian Drăgu. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

Blaschke, T. 2010. "Object Based Image Analysis for Remote Sensing." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1). Elsevier B.V.: 2–16. doi:10.1016/j.isprsjprs.2009.06.004.

Blaschke, Thomas, Geoffrey J. Hay, Maggi Kelly, Stefan Lang, Peter Hofmann, Elisabeth Addink, Raul Queiroz Feitosa, et al. 2014. "Geographic Object-Based Image Analysis - Towards a New Paradigm." *ISPRS Journal of Photogrammetry and Remote Sensing* 87. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS): 180–191. doi:10.1016/j.isprsjprs.2013.09.014.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Cánovas-García, Fulgencio, and Francisco Alonso-Sarria. 2015. "Optimal Combination of Classification Algorithms and Feature Ranking Methods for Object-Based Classification of Submeter Resolution Z/I-Imaging DMC Imagery." *Remote Sensing* 7 (4): 4651–4677. doi:10.3390/rs70404651.

Carleer, Alexandre. P, and Eleonore Wolff. 2006. "Urban Land Cover Multi-Level Region-Based Classification of VHR Data by Selecting Relevant Features." *International Journal of Remote Sensing* 27 (6): 1035–1051.

601 doi:10.1080/01431160500297956.

602 Chan, Jonathan Cheung Wai, and Desiré Paelinckx. 2008. “Evaluation of Random
603 Forest and Adaboost Tree-Based Ensemble Classification and Spectral Band
604 Selection for Ecotope Mapping Using Airborne Hyperspectral Imagery.” *Remote*
605 *Sensing of Environment* 112 (6): 2999–3011. doi:10.1016/j.rse.2008.02.011.

606 Che, Jinxing, Youlong Yang, Li Li, Xuying Bai, Shenghu Zhang, and Chengzhi Deng.
607 2017. “Maximum Relevance Minimum Common Redundancy Feature Selection
608 for Nonlinear Data.” *Information Sciences* 409–410: 68–86.
609 doi:10.1016/j.ins.2017.05.013.

610 Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost : Reliable Large-Scale Tree
611 Boosting System.” *arXiv*, 1–6. doi:10.1145/2939672.2939785.

612 Chen, Tianqi, and Tong He. 2015. “Xgboost: Extreme Gradient Boosting.” *R Package*
613 *Version 0.4-2*.

614 Cheng, Jiehai, Yanchen Bo, Yuxin Zhu, and Xiaole Ji. 2014. “A Novel Method for
615 Assessing the Segmentation Quality of High-Spatial Resolution Remote-Sensing
616 Images.” *International Journal of Remote Sensing* 35 (10). Taylor & Francis:
617 3816–3839. doi:10.1080/01431161.2014.919678.

618 Chubey, Michael S., Steven E. Franklin, and Michael A. Wulder. 2006. “Object-Based
619 Analysis of Ikonos-2 Imagery for Extraction of Forest Inventory Parameters.”
620 *Photogrammetric Engineering & Remote Sensing* 72 (4): 383–394.
621 doi:10.14358/PERS.72.4.383.

622 Conchedda, Giulia, Laurent Durieux, and Philippe Mayaux. 2008. “An Object-Based

623 Method for Mapping and Change Analysis in Mangrove Ecosystems." *ISPRS*
624 *Journal of Photogrammetry and Remote Sensing* 63 (5): 578–589.
625 doi:10.1016/j.isprsjprs.2008.04.002.

626 Demarchi, Luca, Frank Canters, Claude Cariou, Giorgio Licciardi, and Jonathan
627 Cheung Wai Chan. 2014. "Assessing the Performance of Two Unsupervised
628 Dimensionality Reduction Techniques on Hyperspectral APEX Data for High
629 Resolution Urban Land-Cover Mapping." *ISPRS Journal of Photogrammetry and*
630 *Remote Sensing* 87. International Society for Photogrammetry and Remote
631 Sensing, Inc. (ISPRS): 166–179. doi:10.1016/j.isprsjprs.2013.10.012.

632 Duro, Dennis C., Steven E. Franklin, and Monique G. Dubé. 2012. "Multi-Scale Object-
633 Based Image Analysis and Feature Selection of Multi-Sensor Earth Observation
634 Imagery Using Random Forests." *International Journal of Remote Sensing* 33 (14):
635 4502–4526. doi:10.1080/01431161.2011.649864.

636 Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2015. "VSURF: An R
637 Package for Variable Selection Using Random Forests." *The R Journal* 7 (2): 19–
638 33.

639 Georganos, Stefanos, Abdulhakim M. Abdi, David E. Tenenbaum, and Stamatis
640 Kalogirou. 2017. "Examining the NDVI-Rainfall Relationship in the Semi-Arid
641 Sahel Using Geographically Weighted Regression." *Journal of Arid Environments*,
642 July. doi:10.1016/j.jaridenv.2017.06.004.

643 Georganos, Stefanos, Tais Grippa, Sabine G Vanhuysse, Moritz Lennert, and Eléonore
644 Wolff. "Optimizing Classification Performance in an Object-Based Very-High-
645 Resolution Land Use-Land Cover Urban Application." In *Remote Sensing*

Technologies and Applications in Urban Environments II. Vol. 10431.

Gheyas, Iffat A., and Leslie S. Smith. 2010. "Feature Subset Selection in Large Dimensionality Domains." *Pattern Recognition* 43 (1). Elsevier: 5–13. doi:10.1016/j.patcog.2009.06.009.

Giardina, Federica, Jonas Franke, and Penelope Vounatsou. 2015. "Geostatistical Modelling of the Malaria Risk in Mozambique: Effect of the Spatial Resolution When Using Remotely-Sensed Imagery." *Geospatial Health* 10 (2): 333. doi:10.4081/gh.2015.333.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. 2016. "Correlation and Variable Importance in Random Forests." *Statistics and Computing* 27 (3). Springer US: 1–20. doi:10.1007/s11222-016-9646-1.

Grippa, Taïs, Moritz Lennert, Benjamin Beaumont, Sabine Vanhuysse, Nathalie Stephenne, and Eléonore Wolff. 2017. "An Open-Source Semi-Automated Processing Chain for Urban Object-Based Classification." *Remote Sensing* 9 (4): 358. doi:10.3390/rs9040358.

Grippa, T, M Lennert, B Beaumont, S Vanhuysse, N Stephenne, and E Wolff. 2016. "An Open-Source Semi-Automated Processing Chain for Uban OBIA Classification." *Proceedings of the GEOBIA 2016 Conference: Mach-2: Machine Learning & Automation II*, 1–6. http://proceedings.utwente.nl/367/1/Grippa-An Open-Source Semi-Automated Processing Chain For Urban OBIA Classification-28.pdf.

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3 (Mar): 1157–1182.

669 Huang, Xin, and Liangpei Zhang. 2013. “An SVM Ensemble Approach Combining
670 Spectral, Structural, and Semantic Features for the Classification of High-
671 Resolution Remotely Sensed Imagery.” *IEEE Transactions on Geoscience and*
672 *Remote Sensing* 51 (1): 257–272. doi:10.1109/TGRS.2012.2202912.

673 Hughes, Gordon. 1968. “On the Mean Accuracy of Statistical Pattern Recognizers.”
674 *IEEE Transactions on Information Theory* 14 (1). IEEE: 55–63.

675 Janecek, Andreas, Wilfried N Wn Gansterer, Michael Demel, and Gerhard Ecker. 2008.
676 “On the Relationship Between Feature Selection and Classification Accuracy.”
677 *Fsdm* 4: 90–105.

678 Johnson, Brian, Milben Bragais, Isao Endo, Damasa Magcale-Macandog, and Paula
679 Macandog. 2015. “Image Segmentation Parameter Optimization Considering
680 Within- and Between-Segment Heterogeneity at Multiple Scale Levels: Test Case
681 for Mapping Residential Areas Using Landsat Imagery.” *ISPRS International*
682 *Journal of Geo-Information* 4 (4): 2292–2305. doi:10.3390/ijgi4042292.

683 Kavzoglu, Taskin, and Ismail Colkesen. 2013. “An Assessment of the Effectiveness of a
684 Rotation Forest Ensemble for Land-Use and Land-Cover Mapping.” *International*
685 *Journal of Remote Sensing* 34 (12): 4224–4241.
686 doi:10.1080/01431161.2013.774099.

687 Kavzoglu, Taskin, Ismail Colkesen, and Tahsin Yomralioglu. 2015. “Object-Based
688 Classification with Rotation Forest Ensemble Learning Algorithm Using Very-
689 High-Resolution WorldView-2 Image.” *Remote Sensing Letters* 6 (11). Taylor &
690 Francis: 834–843. doi:10.1080/2150704X.2015.1084550.

691 Kuhn, M, Jed Wing, Stew Weston, Andre Williams, Chris Keefer, Allan Engelhardt,

692 Tony Cooper, et al. 2014. “Caret: Classification and Regression Training. R
693 Package Version 6.0--21.” *CRAN: Wien, Austria*.

694 Laliberte, Andrea S., and Albert Rango. 2009. “Texture and Scale in Object-Based
695 Analysis of Subdecimeter Resolution Unmanned Aerial Vehicle (UAV) Imagery.”
696 *IEEE Transactions on Geoscience and Remote Sensing* 47 (3): 1–10.
697 doi:10.1109/TGRS.2008.2009355.

698 Laliberte, Andrea S, D M Browning, and Albert Rango. 2012. “A Comparison of Three
699 Feature Selection Methods for Object-Based Classification of Sub-Decimeter
700 Resolution UltraCam-L Imagery.” *International Journal of Applied Earth*
701 *Observation and Geoinformation* 15. Elsevier: 70–78.

702 Lennert, M.;, and GRASS Development Team. 2016. “Addon I.segment.uspo.”
703 *Geographic Resources Analysis Support System (GRASS) SoftwareVersion 7.3*.
704 Chicago, IL, USA,: Open Source Geospatial Foundation.

705 Li, Linyi, Yun Chen, Tingbao Xu, Rui Liu, Kaifang Shi, and Chang Huang. 2015.
706 “Super-Resolution Mapping of Wetland Inundation from Remote Sensing Imagery
707 Based on Integration of Back-Propagation Neural Network and Genetic
708 Algorithm.” *Remote Sensing of Environment* 164. Elsevier: 142–154.

709 Liaw, Andy, Matthew Wiener, and others. 2002. “Classification and Regression by
710 randomForest.” *R News* 2 (3): 18–22.

711 Löw, Fabian, U Michel, Stefan Dech, and Christopher Conrad. 2013. “Impact of
712 Feature Selection on the Accuracy and Spatial Uncertainty of per-Field Crop
713 Classification Using Support Vector Machines.” *ISPRS Journal of*
714 *Photogrammetry and Remote Sensing* 85. Elsevier: 102–119.

715 Ma, Lei, Tengyu Fu, Thomas Blaschke, Manchun Li, Dirk Tiede, Zhenjin Zhou,
716 Xiaoxue Ma, and Deliang Chen. 2017. “Evaluation of Feature Selection Methods
717 for Object-Based Land Cover Mapping of Unmanned Aerial Vehicle Imagery
718 Using Random Forest and Support Vector Machine Classifiers.” *ISPRS*
719 *International Journal of Geo-Information* 6 (2): 51. doi:10.3390/ijgi6020051.

720 Ma, Lei, Manchun Li, Xiaoxue Ma, Liang Cheng, Peijun Du, and Yongxue Liu. 2017.
721 “A Review of Supervised Object-Based Land-Cover Image Classification.” *ISPRS*
722 *Journal of Photogrammetry and Remote Sensing* 130. The Authors: 277–293.
723 doi:10.1016/j.isprsjprs.2017.06.001.

724 Manchun Li, Lei Ma, Thomas Blaschke, Liang Cheng, Dirk Tiede. 2016. “A Systematic
725 Comparison of Different Object- Based Classification Techniques Using High
726 Spatial Resolution Imagery in Agricultural Environments.” *International Journal*
727 *of Applied Earth Observation and Geoinformation* 49 (April). Elsevier B.V.: 87–
728 98. doi:10.1016/j.jag.2016.01.011.

729 Marpu, Prashanth Reddy, Irmgard Niemeyer, S Nussbaum, and R Gloaguen. 2008. “A
730 Procedure for Automatic Object-Based Classification.” *Object-Based Image*
731 *Analysis*, 169–184. doi:10.1007/978-3-540-77058-9.

732 Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich
733 Leisch. 2014. “e1071: Misc Functions of the Department of Statistics (e1071), TU
734 Wien. R Package Version 1.6-3.” *Retrieved from*.

735 Mountrakis, Giorgos, Jungho Im, and Caesar Ogole. 2011. “Support Vector Machines
736 in Remote Sensing: A Review.” *ISPRS Journal of Photogrammetry and Remote*
737 *Sensing* 66 (3). Elsevier B.V.: 247–259. doi:10.1016/j.isprsjprs.2010.11.001.

738 Mustapha, Ismail Babajide, and Faisal Saeed. 2016. “Bioactive Molecule Prediction
739 Using Extreme Gradient Boosting.” *Molecules* 21 (8): 1–12.
740 doi:10.3390/molecules21080983.

741 Nations, United. 2014. “World Urbanization Prospects: The 2014 Revision, Highlights.
742 Department of Economic and Social Affairs.” *Population Division, United*
743 *Nations*.

744 Neteler, Markus, M Hamish Bowman, Martin Landa, and Markus Metz. 2012. “GRASS
745 GIS: A Multi-Purpose Open Source GIS.” *Environmental Modelling & Software*
746 31. Elsevier: 124–130.

747 Nguyen, Thanh Tung, Joshua Zhexue Huang, and Thuy Thi Nguyen. 2015. “Unbiased
748 Feature Selection in Learning Random Forests for High-Dimensional Data.”
749 *Scientific World Journal* 2015 (December 2014). doi:10.1155/2015/471371.

750 Nicodemus, Kristin K, James D Malley, Carolin Strobl, and Andreas Ziegler. 2010.
751 “The Behaviour of Random Forest Permutation-Based Variable Importance
752 Measures under Predictor Correlation.” *BMC Bioinformatics* 11: 110.
753 doi:10.1186/1471-2105-11-110.

754 Pal, Mahesh, and Giles M. Foody. 2010. “Feature Selection for Classification of
755 Hyperspectral Data by SVM.” *IEEE Transactions on Geoscience and Remote*
756 *Sensing* 48 (5): 2297–2307. doi:10.1109/TGRS.2009.2039484.

757 Pradhan, Biswajeet, and Haleh Nampak. 2017. “Integration of LiDAR and QuickBird
758 Data for Automatic Landslide Detection Using Object-Based Analysis and
759 Random Forests,” no. May: 69–81.

760 Puissant, Anne, Simon Rougier, and André Stumpf. 2014. “Object-Oriented Mapping of
761 Urban Trees Using Random Forest Classifiers.” *International Journal of Applied*
762 *Earth Observation and Geoinformation* 26. Elsevier B.V.: 235–245.
763 doi:10.1016/j.jag.2013.07.002.

764 Rodriguez-Galiano, V. F., M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson, and
765 C. Jeganathan. 2012. “Random Forest Classification of Mediterranean Land Cover
766 Using Multi-Seasonal Imagery and Multi-Seasonal Texture.” *Remote Sensing of*
767 *Environment* 121. Elsevier Inc.: 93–107. doi:10.1016/j.rse.2011.12.003.

768 Schliep, Klaus, and Klaus Hechenbichler. 2014. “Kknn: Weighted K-Nearest
769 Neighbors. R Package. Version 1.2-5.”

770 Song Rongwei, Chen Siding, Deng Bailong and Li Li. 2016. “eXtreme Gradient
771 Boosting for Identifying Individual Users across Different Digital Devices.” *In*
772 *International Conference on Web-Age Information Management, Pp. 43-*
773 *54.Springer International Publishing* 9098: 43–54. doi:10.1007/978-3-319-21042-
774 1.

775 Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007.
776 “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and
777 a Solution.” *BMC Bioinformatics* 8: 25. doi:10.1186/1471-2105-8-25.

778 Therneau, Terry, Beth Atkinson, and Brian Ripley. 2015. “Rpart: Recursive Partitioning
779 and Regression Trees. R Package Version 4.1--10.”

780 Van Coillie, Frieke M B, Lieven P C Verbeke, and Robert R. De Wulf. 2007. “Feature
781 Selection by Genetic Algorithms in Object-Based Classification of IKONOS
782 Imagery for Forest Mapping in Flanders, Belgium.” *Remote Sensing of*

Environment 110 (4): 476–487. doi:10.1016/j.rse.2007.03.020.

Waske, Björn, Sebastian van der Linden, Jón Atli Benediktsson, Andreas Rabe, and Patrick Hostert. 2010. "Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data." *IEEE Transactions on Geoscience and Remote Sensing* 48 (7). IEEE: 2880–2889.

Witten, Ian H, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Xia, Yufei, Chuanzhe Liu, YuYing Li, and Nana Liu. 2017. "A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring." *Expert Systems with Applications* 78. Elsevier Ltd: 225–241. doi:10.1016/j.eswa.2017.02.017.

Yu, Lei, and Huan Liu. 2003. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution." In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 856–863.

806

807

808 **Tables**

809 Table 1 Classification scheme, train and test data for Ouagadougou.

LULC	Training Set Size	Test Set Size	Segments Median Area (m^2)
Buildings	157	157	273
Swimming Pools	68	69	119
Asphalt	56	56	30240
Brown/Red Bare Soil	90	91	55450
White/Grey Bare Soil	72	72	11300
Trees	77	77	380
Mixed Bare Soil/Vegetation	76	75	24390
Dry Vegetation	70	70	30420
Other Vegetation	139	141	7464
Inland Waters	75	75	2073
Shadow	58	59	72

810

811

812

813

814

815

816

817

818

819

820

Table 2 Optimized parameters for the Xgboost, RF and SVM, and KNN models.

Classifier	Parameters	Value
Xgboost	nrounds	600
	eta	0.015
	gamma	0
	depth	4
	child_weight	2
	subsample	0.4
	column_sample	0.3
RF	ntree	2000
	mtry	1/3 of input features
SVM	gamma	0.003
	C	15
KNN	k	10

Table 3 Peak Overall Accuracy (%) for each classifier, based on different FS methods. Numbers in parenthesis () depict the number of features to achieve the reported result. Values with * imply significant improvement over the model with all features, as suggested by a one tailed McNemar test of similarity.

Algorithm	CFS	MDA	RFE	VSURF	All Features
Xgboost	79.2 (42)	79.1 (136)	79.8* (23)	79.5* (31)	77.8 (169)
RF	78.5 (117)	78.6 (84)	78.9 (22)	78.1 (38)	77.7 (169)
SVM	80.1* (37)	78.8 (111)	79.2 (32)	79.7* (75)	78.1 (169)
KNN	78.2* (51)	76.2* (75)	78.0* (42)	77.9* (20)	74 (169)
Rpart	69.5 (53)	69.4 (123)	69.6 (34)	70.1 (49)	69.4 (169)

855

856 Table 4 Top 30 features according to each feature selection method.

CFS	MDA	RFE	VSURF
compact_circle	NDVI_first_quart	NDVI_mean	NDVI_third_quart
fd	NDWI_third_quart	nDSM_median	NDVI_mean
opt_green_first_quart	NDWI_median	NDVI_median	NDVI_median
opt_green_median	NDWI_mean	NDVI_third_quart	NDVI_perc_90
opt_green_perc_90	NDWI_perc_90	NDVI_first_quart	NDVI_first_quart
opt_nir_min	NDVI_mean	NDWI_perc_90	NDWI_third_quart
opt_nir_max	NDVI_median	opt_nir_first_quart	NDWI_median
opt_nir_mean	Brightness_perc_90	NDWI_third_quart	NDWI_mean
opt_nir_coeff_var	opt_red_median	nDSM_first_quart	Brightness_first_quart
opt_nir_first_quart	Brightness_third_quart	NDVI_perc_90	NDWI_perc_90
opt_nir_median	opt_red_first_quart	opt_nir_mean	opt_green_median
opt_nir_perc_90	NDWI_first_quart	NDWI_coeff_var	nDSM_median
opt_red_max	Brightness_mean	texture_asm5_blue_ASM_min	Brightness_median
opt_red_first_quart	Brightness_median	opt_nir_median	opt_green_first_quart
opt_red_median	opt_red_third_quart	nDSM_mean	opt_nir_first_quart
NDVI_min	Brightness_first_quart	texture_asm5_blue_ASM_first_quart	texture_asm5_blue_ASM_first_quart
NDVI_mean	opt_red_mean	compact_circle	nDSM_mean
NDVI_stddev	NDVI_third_quart	opt_nir_max	NDVI_stddev
NDVI_first_quart	nDSM_median	NDWI_median	Fractal_d
NDVI_median	NDVI_sum	NDWI_mean	compact_circle
NDVI_third_quart	opt_red_perc_90	nDSM_third_quart	texture_asm5_green_ASM_first_quart
NDVI_perc_90	compact_circle	NDVI_min	opt_nir_coeff_var
Brightness_first_quart	NDWI_sum	NDVI_stddev	Brightness_third_quart
Brightness_median	opt_green_perc_90	Brightness_first_quart	NDWI_first_quart
nDSM_coeff_var	NDVI_min	NDWI_stddev	opt_nir_mean
nDSM_median	opt_green_mean	opt_green_median	nDSM_third_quart
nDSM_third_quart	nDSM_third_quart	opt_nir_coeff_var	opt_nir_max
NDWI_max	opt_green_median	Brightness_median	opt_nir_median
NDWI_stddev	perimeter	opt_nir_third_quart	opt_green_mean

857

858

Table 5 Top 10 highest ranked models according to the COS

Classification Model	S	CO	Overall Accuracy	Number of Features
RFE_XGBOOST	2	0.98	0.798	23
CFS_SVM	0	0.98	0.801	38
CFS_SVM	5	0.97	0.800	37
CFS_SVM	5	0.97	0.801	39
RFE_XGBOOST	2	0.97	0.792	22
VSURF_XGBOOST	1	0.97	0.795	31
RFE_XGBOOST	0	0.97	0.792	25
CFS_SVM	0	0.97	0.797	36
CFS_SVM	0	0.97	0.799	40
VSURF_SVM	9	0.96	0.789	20

List of Figures

- Figure 1 Ouagadougou, Burkina Faso. (a) RGB composite of the whole study area, (b) reference map of Ouagadougou within the African continent, (c) high density built up area and (d) nDSM for the same area.
- Figure 2 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the Xgboost classifier. The model including all features considered for the classification is represented by the flat horizontal line.
- Figure 3 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the SVM classifier. The model including all features considered for the classification is represented by the flat horizontal line.
- Figure 4 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the RF classifier. The model including all features considered for the classification is represented by the flat horizontal line.
- Figure 5 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the KNN classifier. The model including all features considered for the classification is represented by the flat horizontal line.

- Figure 6 Relationship between Overall Accuracy (OA), number of features and feature rankings based on each FS method for the RPART classifier. The model including all features considered or the classification is represented by the flat horizontal line.
- Figure 7 Relationship between the number of features in the classification and the storage size (left) and the computing time (right) to extract the selected predictors for the whole dataset of Ouagadougou, at an object based format. The study area consists of ~ 12 million segments.
- Figure 8 Classification results from the highest ranked COS model (Xgboost, 22 features) for different regions in Ouagadougou. a) Industrial, b) planned and c) unplanned neighborhoods.