PFG 2011 / 3, 135 – 144
Stuttgart, June 2011

Article

Feature Evaluation for a Transferable Approach of Object-based Land Cover Classification Based on Ikonos and QuickBird Satellite Data

Nils Wolf, Bochum

Keywords: Random Forests, Feature Selection, QuickBird, Ikonos, Object-based Image Analysis

Summary: This contribution aims at revealing features that can be used for generic object-based land cover classification of Ikonos and QuickBird satellite data. On seven satellite scenes the Random Forest algorithm – a tree-based ensemble classifier – is applied as it provides an internal measure to obtain feature importance scores. This measure quantifies in how far a feature contributes to the reduction of entropy (here based on the Gain Ratio criterion) when constructing a tree. The features under investigation comprise textures and variances, obtained on an image segmentation level, and the brightness values of single bands, respectively their ratios and differences, obtained on pixel level. As an outcome, the features are ranked with respect to their robustness. Top-ranked are those features which show good overall performance on each of the seven scenes.

Zusammenfassung: Bewertung von Merkmalen für einen übertragbaren objektbasierten Ansatz zur Landbedeckungsklassifikation basierend auf Ikonos und QuickBird Satellitenbilddaten. Dieser Beitrag beschreibt ein Verfahren zur Bewertung von Merkmalen hinsichtlich ihrer Eignung für einen übertragbaren objektbasierten Ansatz zur Landbedeckungsklassifikation. Gesucht sind demnach Merkmale, welche sich gegenüber spezifischen Einflussfaktoren verschiedener Eingangsdaten – hier sieben Ikonos und QuickBird Szenen – robust verhalten. Die Bewertung erfolgt über die Induktion von Verbünden de-korrelierter Entscheidungsbäume, sogenannter Random Forests. Der Informationsgewinn einzelner Merkmale an den Knotenpunkten der Entscheidungsbäume wird durch das Gain Ratio Maß ermittelt und quantifiziert in der Aggregation über den Verbund die Wichtigkeit der Merkmale. Der zu untersuchende Merkmalsraum setzt sich aus Texturen und Varianzen auf Segmentierungsebene sowie Grauwerten einzelner Bänder (bzw. deren Kombination in Ratios und Differenzen) auf Pixelebene zusammen. Das Ergebnis dieser Arbeit ist eine Bewertung der Merkmale hinsichtlich ihrer datensatzübergreifenden Qualität.

1 Introduction

Land cover classification is a common remote sensing scenario and applied for many purposes, such as urban growth mapping, parameterization of hydrological or climate models, or mobile network planning. In recent years, an increasing number of optical very high resolution (VHR) satellite systems are put into operation, providing repetitively amounts of data which raises the need of automating their interpretation. Given a common classification task, it is questioned in how far classification can be generic and robust to variances of the input provided by such VHR systems.

Ikonos and QuickBird share a nearly identical Kodak sensor system (Jacobsen 2003) and deliver products with comparable spectral characteristics. The main difference of both systems lies in the spatial resolution capacities caused by different orbits. Regardless of the sensor system, in fact every scene is affected

www.schweizerbart.de
DOI: 10.1127/1432-8364/2011/0078
1432-8364/11/0078 $ 2.50

by specific properties, such as atmospheric conditions, sun elevation or off-nadir angle. Furthermore, adapting a classification method to another study area further requires a stable semantic scheme and an even more generalizing, but though precise model. Wolf et al. (2010) propose a generic framework for object-based classification (rule-based expert system) which is adaptable to different Ikonos and QuickBird scenes by tuning a set of key-parameters (visual on-screen inspection). The present study aims at improving this framework by identifying features that yield good performance on different scenes as they are less affected by scene-specific conditions.

In order to measure the importance of features with respect to their robustness, they are evaluated separately on training data of seven Ikonos and QuickBird scenes with footprints across a highly urbanized region, the Ruhr Area, in Germany. Comparing and averaging the features' performances over the different training sets is assumed to reveal robust features that can be used for a generic object-based land cover classification framework. The importance scores are obtained by measuring in how far particular features contribute to the construction of decision boundaries applying a Random Forest (RF) classifier. RF are ensemble classifiers that induce a large number of de-correlated decision trees which give their votes for unknown instances.

The feature space under investigation is generated within an object-based environment. Object-based approaches for image analysis take advantage of a feature space extended beyond the n-dimensional matrix space of an image because a local neighborhood of pixels – created by image segmentation – can be utilized in order to obtain also variances, textures and further metrics (Benz et al. 2004). In comparison to moving window approaches which include local neighborhoods as well, objects based on segmentation represent homogeneous image regions that are assumed to exhibit features which better describe the real-world entities (Castilla & Hay 2008).

This paper is organized as follows: In Sections 2 and 3, decision trees and the RF ensemble method are introduced. Section 4 describes how the feature importance measures – technically a by-product of the RF model induction – is obtained. In Section 5, the method is applied for the investigation of two binary classification problems: firstly, the separation of built-up (i. e., buildings/roofs, pavements, and other artificial materials) from bare surfaces (e. g., rocks or exposed soils) and, secondly, the separation of water bodies and shadowed non-vegetation surfaces. For sake of simplicity, these categories are referred to in the following as: *built-up*, *bare*, *water* and *shadow*. For consistency of the land cover classification scheme, the class vegetation constitutes the complement class. In Section 6, the results are presented and discussed while Section 7 concludes the work.

2 Tree Induction

The most common decision tree recursively partitions a feature space by axis-parallel linear splits whereas the decision boundaries ideally enclose instances representing only one class. Prominent algorithms are ID3 and its successor C4.5 (Quinlan 1986), or *Classification and Regression Trees* (CART – Breiman 1984). Trees are non-parametric classifiers; thus they do not rely on assumptions regarding the distribution of the data.

The RF classifier applied for this study utilizes a modified CART algorithm to construct the ensemble of trees. CART produces binary trees with univariate splits. The splits aim at reducing the impurity of the child nodes, which is evaluated by a best-split criterion. As best-split criterion, CART employs the Gini Index while the present study uses the Gain Ratio instead. The Gain Ratio measure tends to create smaller trees (Quinlan 1986, Mingers 1998) and led in some test runs of this study to slightly better performances.

The best-split criterion based on the Gain Ratio is derived in the following (Kohavi 1999). Let $C = \{0,1,...,a\}$ be the class attribute (here only C_0 and C_1, e. g., *water* and *shadow*) and let S be a set of n training instances, attributed with p features; respectively, $S_i\{0,1,...,b\}$ are partitions of S (here restricted to S_0 and S_1). Further, $RF(C_i,S)$ denotes the relative frequency of instances that belong to class C_i.

The entropy (Shannon) is a measure of uncertainty associated with a feature. For a discrete set of instances it is defined as:

$$H = -\sum_{i=a}^{i=1} RF\,(C_i, S)\,\log_2\,(RF(C_i, S)) \quad (1)$$

Imagine S to be arranged as a sequence, ordered with respect to a particular feature. The first binary split possible is the one after the first entry of the sequence, the second split after the second entry, and so on. The maximum number of splits is determined by n-1*p, resulting each time in two candidate partitions with its own entropy $H(S_i)$. The difference of entropy before and after the split is called Information Gain and defined as:

$$G(S\,B) = H(S) - \sum_{j=1}^{b} \frac{|S_j|}{|S|} H(S_j) \quad (2)$$

where B denotes the test on a particular split.

Moreover, the Gain Ratio is introduces to normalize Information Gain by the number and the sizes of generated child nodes from a candidate split. The Gain Ratio is defined as $G(S,B)/P(S,B)$, where $P(S,B)$ denotes the intrinsic entropy:

$$P(S\,B) = -\sum_{j=1}^{b} \frac{|S_j|}{|S|} \log_2\left(\frac{|S_j|}{|S|}\right) \quad (3)$$

The test B which maximizes the Gain Ratio finally constitutes a particular node of a tree and the whole process is iterated for the child nodes, until pure leafs or other stopping criterions are reached.

3 Random Forests

Random Forests (RF), invented by Breiman (2001), have been become popular as they are simple to tune and also applicable to high-dimensional problems with only few training instances (*"n < p"*-problem) (Hastie 2007). They show good predictive accuracy for problems with highly correlated features, which is relevant to this study that is based on multi-spectral satellite data and various correlating derivatives. Even though RF are known for well dealing with high-order interactions (Strobl 2008), it is assumed that associating interactions in artificial features (such as band ratios) is advantageous for keeping decision rules simple. Furthermore, RF allow fast computation on large datasets and lead to accuracies comparable to the well approved boosting or support vector machine learners (Breiman 2001, Diaz-Uriarte & Alvarez de Andres 2005, Hastie 2007).

RF belong to the category of ensemble classifiers which share the concept of constructing several base learners and combining their outputs to a committee with superior performance. Fully grown and unpruned trees have approximately no bias but they suffer from high variance. They are grown in a greedy manner and form unstable models which track every instance by a branch, hence even noise is memorized. By averaging over a group of de-correlated trees, RF counteract the variance problem while keeping a low bias (Diaz-Uriarte & Alvarez de Andres 2005). Breiman (2001) stated that overfitting is not an issue for this type of classifier.

The performance of RF highly depends on the de-correlation of the trees, which Breiman (2001) solves by combining his idea of bagging (Breiman 1996) with Ho's (1998) concept of random feature subspaces. Hereby, randomness is injected at several stages of the learning process. Firstly, bagging introduces random variation for the training datasets by bootstrap resampling. Given a training dataset S with n instances, bagging generates m new training datasets S_i by sampling instances from S uniformly and with replacement. Thus, a particular instance has the following probability of being in a dataset S_i:

$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} \approx 0.638 \quad (4)$$

A tree is fit to each bootstrap training set S_i, resulting in m de-correlated trees, while instances not belonging to S_i, about one third of the dataset, form the so-called out of bag (oob) data. The oob data constitutes an independent test set for each tree which can be used to estimate the prediction error without need for external tests, such as cross-validation. It can be stated that the error converges

as m increases. An empirical test about the setting of m is given in HASTIE et al. (2007).

A further injection of randomness comes from sub-sampling the feature space. At each node of a tree, a feature set p_s (with $p_s < p$) is drawn randomly and evaluated by the best-split-criterion. BREIMAN (2001) shows that small p_s (about $p^{-0.5}$+1) yield good results because the prediction error largely depends on the trees' de-correlation.

4 Feature Scoring Based on Random Forests

Feature selection is a commonly applied preprocessing step in machine learning and aims at finding an efficient subset of features. Reducing the dimensionality can improve models, save computational costs or gain a better understanding of an underlying problem (GUYON et al. 2003). Feature selection approaches can be categorized in filter, wrapper and embedded methods, depending on their interaction with a classifier. Filter methods are cheap to obtain because they directly operate on the data. They assess features individually (e. g., by Gain Ratio, Fisher Score or Relief-F) and therefore suffer from ignoring the importance of a feature in presence of another. Wrapper methods evaluate different feature subspaces by the performances of an applied classifier. Generally, they provide good results, however alternating through combinations of features (e. g., by genetic algorithms, greedy forward selection or greedy backward elimination) as well as the mandatory repetitive performance estimation (e. g., by cross-validation) make wrappers computationally expensive. Embedded methods often provide a reasonable trade-off between quality and computational costs. They derive feature importance directly from a classification model (e. g., by using the weight vector in support vector machines).

In this study, an embedded method is used, based on the application of the RF classifier (similar as described in MENZE 2009). For each feature, an importance score is obtained by accumulating its Gain Ratio values over all nodes of the forest. Thus, the score describes the number but also the quality of the feature's occurrences. By applying this method on the different input spaces of the seven satellite scenes, the features can be evaluated regarding their robustness. Good features would show a high average performance and – at the same time – no negative outlier for any of the tests.

5 Data and Application

5.1 *Satellite Image Data*

Fig. 1 gives an overview of the used satellite data, comprising three QuickBird (*Ortho-ready*) and four Ikonos (*Standard Geometrically Corrected*) scenes with a total footprint size of 1130 km². The study site is located in the Ruhr Area, Germany. The images are characterized by manifold urban structures, adjacent to agricultural and some forested and (semi-)natural areas.

The metadata of the images, listed in Tab. 1, reveal some of the scene-specific properties which are just a few of the factors that potentially complicate the adaption of image analysis rules from one scene to another. The acquisition dates vary between April and September, the sun elevation angles between 41 and 51 degrees and the off-nadir angles between 8 and 28 degrees. Furthermore, the images comprise different pixel sizes, i. e., 1 m (panchromatic)/4 m (multispectral) for Ikonos and 60 cm/2.4 m for QuickBird.

The image data pool is further extended by pansharpened datasets, using the *Subtractive Resolution Merge* (SR-Merge) and the *Principle Component Resolution Merge* (PC-Merge) algorithms, both implemented in the *ERDAS Imagine 10* software package.

5.2 *Sample Selection*

Given is a land cover classification task (referred to the urban land cover classification scheme by HEROLD (2004)) with the focus on two binary classification problems which are frequently reported as the core problems: built-up vs. bare and water vs. shadow (on non-vegetation surfaces). Accordingly, training samples are collected on each scene by the

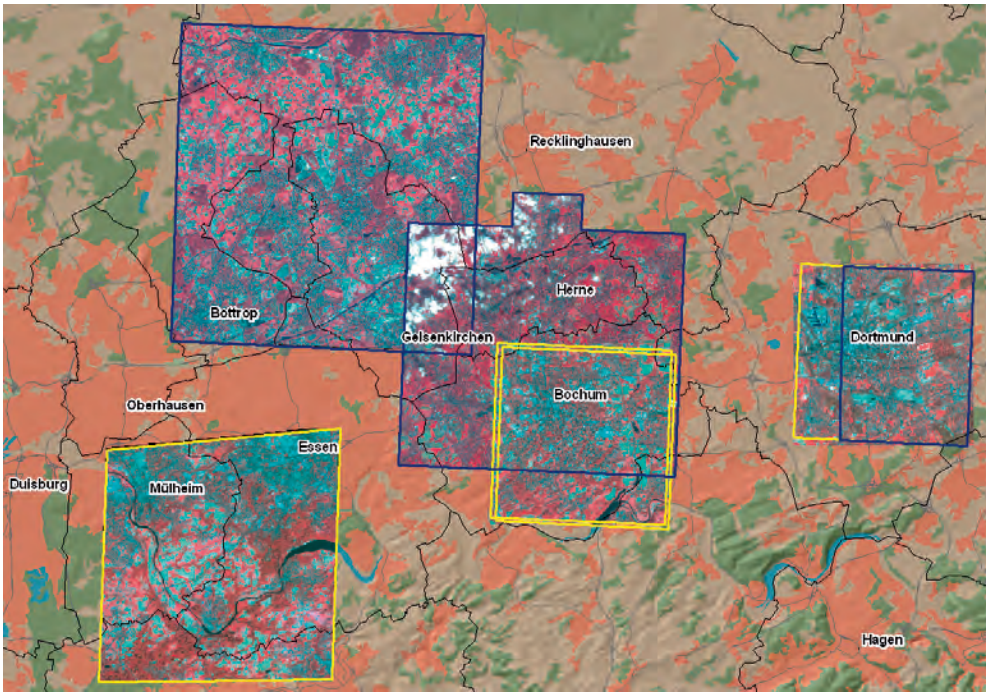

Fig. 1: The study site, Ruhr Area, and the coverage by satellite data, comprising four Ikonos scenes (framed in yellow) and three QuickBird scenes (framed in blue). Each footprint consists of the original multispectral and panchromatic layers, plus two further 4-layer stacks created by pan-sharpening (PC-Merge and SR-Merge). Background map: SRTM hillshade, CORINE Land Cover (CLC2000), basemap (districts and highways).

Tab. 1: Metadata of the Ikonos and QuickBird satellite scenes.

Satellite Scene	GSD [m]	Clouds [%]	Off-nadir [degree]	Sun Elev. [degree]	Acquisition [M/D/Y]
Ikonos (Bochum)	1.0 / 4.0	1	27.9	48.8	04/21/2005
Ikonos (Bochum)	1.0 / 4.0	0	26.2	42.7	09/11/2008
Ikonos (Dortmund)	1.0 / 4.0	1	22.7	41.7	04/02/2007
Ikonos (Essen - Mülheim)	1.0 / 4.0	0	23.3	44.5	09/06/2005
QuickBird (Bochum)	0.6 / 2.4	14	13.3	51.1	08/16/2009
QuickBird (Dortmund)	0.6 / 2.4	0	7.6	51.2	04/25/2006
QuickBird (Gelsenkirchen - Bottrop)	0.6 / 2.4	0	13.5	42.2	09/10/2004

labeling of image objects. The image objects are created using the Multiresolution Segmentation (Baatz & Shäpe 2000, implemented in eCognition Developer 8) on the panchromatic layers with the following parameter setting: scale = 20/(pixel size[m] * 1.4); shape = 0.2; compct: 0.8. The scale parameter value is normalized by the pixel size in order to obtain objects of comparable size (Ikonos and QuickBird provide different spatial resolutions). In a subsequent step, small objects of less than 10 pixels are merged into their spectrally most similar neighbor object. However, no merge is conducted if none of the neighbors offer a considerable similarity (here the difference of the spectral mean values is restricted to a maxi-

mum of 50). The remaining small objects of less than 10 pixels are excluded from the sampling process as they are at risk to exhibit statistically less reliable or even undefined feature values. The overall premise guiding the process of object creation is to avoid under-segmentation, i. e., objects representing more than one class of interest, but to ensure at the same time that objects are large enough to exhibit significant attributes.

Obligatory, land cover mapping raises questions about dichotomies and semantics of classes, especially for crisp classification schemes. Land cover categories are formed out of a continuum of different materials and always remain vague. The question is about the trade-off between reality, their projection into image data, and the highly subjective conceptual reality of the thematic background. In this study, semantic highly ambiguous objects are not considered for the sampling because they are more likely to inject noise than to supply the indeed valuable class extremes. Here, examples for semantically highly ambiguous objects are: very marshy or even silted up water bodies mainly covered by plants, swimming pools, or farm tracks where the pavement is only partially visible.

5.3 *Features*

61 features are generated on an image object level and 65 features on pixel level (for an overview see Tab. 4). The former feature set contains textures (after HARALICK 1973) and variances while the latter represents the digital numbers of individual layers, respectively their combinations (e. g., ratios and differences). Projecting the training regions also on a pixel level brings several advantages: Firstly, pixel values and their distribution remain pure as they are not averaged per object. Secondly, there is no need to counteract the under-representation of homogeneous image regions which tend to be aggregated in larger objects. Furthermore, the number of training instances increases significantly.

Tab. 2: Formal description of features.

Feature	**Description**	
GLCM Contrast	$\sum_{i\,j=0}^{N-1} \frac{P_{i\,j}}{1+(i-j)^2}$	**i:** row number **j:** column number **$P_{i,j}$:** normalized value in the cell i,j **N:** number of rows or column **V_k:** normalized grey level difference vector V(k) = SUM(i,j=0,N-1 and |i-j|=k) P(i,j) **B_1:** BLUE band **B_2:** GREEN band **B_3:** RED band **B_4:** NIR band Ratios: B_x / B_y (e.g., vis / B2)
GLCM Homogeneity	$\sum_{i\,j=0}^{N-1} P_{i\,j}(i-j)^2$	
GLCM Entropy	$\sum_{i,j=0}^{N-1} P_{i,j}(-\ln P_{i,j})$	
GLDV Entropy	$\sum_{k=0}^{N-1} V_k(-\ln V_k)$	
ndvi	$(B_4 - B_3)/(B_4 + B_3)$	
ndwi	$(B_4 - B_2)/(B_4 + B_2)$	
bndvi	$(B_4 - B_1)/(B_4 + B_1)$	
sd	$(B_1 - B_2)^2+(B_2 - B_3)^2+(B_3 - B_4)2$	
brightness	$B_1 + B_2 + B_3 + B_4$	
vis	$B_1 + B_2 + B_3$	
ssi	abs $(B_1+B_3-2B_2)$	

A formal description for the texture features as well as ratios, differences and other indices is provided in Tab. 2. All feature values are calculated in the software *eCognition Developer 8* and exported to a spreadsheet format.

5.4 *Feature Importance Analysis*

The RF algorithm and the associated feature scoring used in this work are implemented in the software *RapidMiner 5*. The RF are constructed as described in Sections 2 and 3. The remaining tuning parameters are set as follows: The random subspaces p_s are restricted to $p^{-0.5}$+1 features and the number of trees (all fully grown) constituting a forest is set to 200, which should approximate an equal representation of features in the total of p_s.

For each of the seven training sets, feature scores are obtained on four scenarios as both binary classification problems are treated separately and, moreover, further split in individual runs for the object level and pixel level feature spaces. Thus, each training set leads to four preliminary results where features are ranked according to their accumulated Gain Ratio values. These scores are normalized ([0-1]) to avoid a bias towards training sets where classes are more easily separable. The overall importance is obtained by averaging the normalized scores obtained on the different training sets – again with regard to the four scenarios.

The method is applied in two stages. In the first stage, scores are obtained by restricting the feature spaces to the pansharpened images. In doing so, both pansharpening methods are evaluated against each other by a direct competition. To confirm this outcome, the oob error is calculated on scenarios with feature spaces restricted to the particular pansharpening method. Features stemming from the inferior pansharpened dataset are excluded from any further investigation. Consequently, in the second stage, scores are obtained on features stemming from the original images and the superior pansharpening layers.

The problem of class imbalances is treated by down-sampling the majority class. Object level scenarios are represented by at least ≈ 200 instances per class and pixel level scenarios by at least ≈ 2800. Class imbalances are a frequently reported problem for many classifiers and also RF are concerned by this as they are constructed to minimize the overall error, which can leads to poor accuracies for the minority class (Chen 2004).

6 Results and Discussion

6.1 *Pansharpening Competition*

Tab. 3 shows the results of the pansharpening competition by comparing the scores (quotient: *PC-Merge*/*SR-Merge*) according to training sets and scenarios. In the majority of the cases the PC-Merge outperforms the SR-Merge, resulting in an overall average quotient of 1.074. The oob error for feature spaces that exclusively refer to one of the pansharpened sets confirms this tendency. The PC-Merge scenarios obtain an average error of 0.155 in comparison to 0.178 for the SR-Merge scenarios.

Tab. 3: Values obtained by dividing the PC-Merge scores by the SR-Merge scores, with respect to the four scenarios (ws= *water* vs. *shadow*, bb = *built-up* vs. *bare*) and the seven training sets (denoted by their corresponding satellite scenes: IK = Ikonos, QB = QuickBird). Green cells indicate a superior performance by the PC-Merge (i. e., cell value > 1).

	QB_a	QB_b	QB_c	IK_a	IK_b	IK_c	IK_d	Average
ws_ob	1.270	1.042	1.270	0.893	1.073	1.033	1.198	1.111
ws_px	0.876	0.891	1.071	0.781	1.564	1.012	1.096	1.041
bb_ob	1.120	1.201	0.994	1.045	1.101	1.214	1.144	1.117
bb_px	1.183	1.151	0.983	0.923	1.003	0.847	1.101	1.027
Average	1.112	1.071	1.079	0.910	1.185	1.026	1.135	1.074

Tab. 4: Overall feature importance scores (OFIS), calculated by averaging the results obtained on the seven training sets. The prefix *ps_* denotes the PC-Merge layers; if no prefix is given, the original multispectral and panchromatic layers are addressed.

built-up vs bare			
pixel level		object level	
feature	OFIS	feature	OFIS
vis/red	.681	Standard deviation nir	.817
vis/nir	.552	Standard deviation blue	.664
bndvi	.548	Standard deviation ps_nir	.649
blue/red	.533	Standard deviation ps_blue	.600
ps_vis/ps_green	.528	Standard deviation pan	.539
ps_vis/ps_red	.528	Standard deviation green	.524
blue/green	.523	GLCM Entropy ps_green	.466
Mean ps_blue	.499	GLCM Entropy ps_red	.461
green/red	.474	GLDV Entropy ps_nir	.444
vis/blue	.467	GLCM Homogeneity ps_nir	.432
ndwi	.449	GLCM Homogeneity green	.432
ps_sd	.436	GLCM Homogeneity ps_green	.431
ps_green/ps_red	.434	GLDV Entropy ps_red	.406
ps_vis/ps_nir	.425	GLCM Contrast pan	.406
ps_blue/ps_nir	.420	GLCM Contrast ps_nir	.403
green/nir	.418	GLCM Homogeneity nir	.395
ndvi	.417	Standard deviation red	.391
blue/nir	.417	GLCM Homogeneity red	.386
ps_ssi	.416	GLDV Entropy ps_blue	.379
vis/green	.388	Standard deviation ps_green	.376
ps_bndvi	.383	GLCM Homogeneity ps_blue	.373
ps_blue/ps_red	.365	GLDV Entropy red	.343
Mean ps_green	.346	GLCM Contrast red	.342
sd	.340	GLCM Contrast blue	.341
ps_vis	.337	GLCM Contrast green	.338
ps_green/ps_nir	.335	GLCM Contrast ps_green	.337
ps_blue/ps_green	.332	GLCM Contrast nir	.333
Mean ps_nir	.323	GLCM Contrast ps_blue	.331
ps_vis/ps_blue	.321	GLDV Entropy green	.328
ps_ndwi	.317	GLCM Entropy ps_blue	.327
ps_ndvi	.315	GLDV Entropy pan	.326
brightness	.290	GLCM Entropy green	.319
vis	.263	GLDV Entropy blue	.315
ps_brightness	.259	GLDV Entropy ps_green	.314
ssi	.249	GLCM Homogeneity blue	.303
Mean blue	.241	GLCM Entropy nir	.285
Mean pan	.240	GLCM Entropy pan	.273
Mean nir	.188	GLDV Entropy nir	.267
Mean red	.170	GLCM Homogeneity ps_red	.257
Mean green	.140	Standard deviation ps_red	.253
Mean ps_red	.088	GLCM Entropy ps_nir	.246
		GLCM Entropy blue	.237
		GLCM Entropy red	.215
		GLCM Contrast ps_red	.208
		GLCM Homogeneity pan	.200

water vs shadow			
pixel level		object level	
feature	OFIS	feature	OFIS
ps_blue/ps_green	.679	GLCM Homogeneity ps_blue	.702
ps_blue/ps_nir	.614	Standard deviation green	.473
bndvi	.571	Standard deviation blue	.462
ndwi	.560	GLDV Entropy ps_blue	.457
sd	.555	Standard deviation nir	.442
ps_vis	.551	GLCM Homogeneity ps_red	.431
green/nir	.550	GLDV Entropy ps_red	.403
ps_vis/ps_nir	.550	GLCM Contrast ps_blue	.402
ps_vis/ps_green	.547	Standard deviation red	.400
vis/nir	.545	GLCM Contrast ps_nir	.399
ps_sd	.541	GLCM Homogeneity ps_green	.380
blue/red	.526	GLDV Entropy ps_nir	.359
ps_ndwi	.506	GLCM Contrast nir	.358
ps_ndvi	.492	Standard deviation ps_blue	.350
ps_green/ps_red	.466	Standard deviation ps_nir	.350
ps_green/ps_nir	.463	GLCM Entropy nir	.349
ps_bndvi	.418	Standard deviation pan	.343
blue/nir	.411	GLCM Homogeneity pan	.327
vis/blue	.409	GLCM Homogeneity ps_nir	.317
ps_vis/ps_red	.404	GLCM Contrast green	.296
ssi	.391	GLCM Entropy blue	.278
green/red	.375	GLCM Contrast red	.271
ps_blue/ps_red	.365	GLCM Entropy ps_green	.259
ndvi	.360	GLCM Entropy ps_nir	.250
vis/green	.330	GLDV Entropy ps_green	.250
ps_vis/ps_blue	.326	GLCM Homogeneity blue	.238
vis/red	.313	GLCM Homogeneity nir	.233
Mean ps_red	.308	GLCM Homogeneity green	.227
ps_ssi	.307	Standard deviation ps_red	.220
brightness	.296	GLCM Homogeneity red	.217
blue/green	.288	GLCM Entropy ps_red	.215
ps_brightness	.285	GLCM Entropy pan	.208
Mean ps_nir	.261	GLCM Contrast pan	.207
Mean ps_blue	.259	GLCM Entropy green	.207
Mean pan	.247	GLCM Contrast blue	.202
Mean nir	.247	GLCM Entropy ps_blue	.184
vis	.227	GLDV Entropy nir	.181
Mean ps_green	.193	Standard deviation ps_green	.179
Mean green	.172	GLCM Entropy red	.160
Mean red	.165	GLCM Contrast ps_red	.151
Mean blue	.133	GLCM Contrast ps_green	.149
		GLDV Entropy pan	.139
		GLDV Entropy red	.094
		GLDV Entropy blue	.076
		GLDV Entropy green	.074

6.2 *Overall Importance Scores – Robust Features*

Tab. 4 ranks the features by their overall importance scores (averaged over the results of the different training sets). The preliminary results obtained on the individual training sets though can be important when also considering a minimum score criterion, i. e., a feature with a fair average performance but with one poor result on any of the scenes might not be seen as robust. The oob error estimates for all RF models are below 0.38, with an average of 0.09 for the *water* vs. *shadow* scenarios and 0.16 for the *built-up* vs. *bare* scenarios.

Scenario *built-up* vs. *bare*, pixel level: The feature *vis/red*, calculated by (*BLUE band* + *GREEN band* + *RED band*)/*RED band*, obtains the highest score. It refers to pigment contents of the visual spectrum and seems to address the separation of rather bluish urban

materials (e. g., asphalt, tar, concrete or gray shingle roofs) from the rather reddish tones of nun-urban bare surfaces (dark brown irrigated acres or strong reflecting sandy soils on construction and dump sites). Furthermore, it can be noted that the most discrimination stems from the lower resolution multispectral bands (here the four best-ranked features are based on the 2.4 respectively 4.0 m resolution multispectral layers).

Scenario *built-up* vs. *bare*, object level: This scenario unambiguously votes for standard deviation features which obtained higher scores than textures based on GLCM and GLDV. The greatest separability stems from *Standard deviation nir*, leaving some gap to the next best-ranked *Standard deviation blue*.

Scenario *water* vs. *shadow*, pixel level: In contrast to the object-based scenario, the discrimination is foremost supplied by the pan-sharpened layers: the features *ps_blue/ps_nir* and *ps_blue/ps_green* are best-ranked. This result hints at the mixed pixel problem of the lower resolution multispectral layers, which becomes relevant for small-sized entities of a few square meters, here numerous shadows, mainly caused by buildings. Furthermore, some bias might result from the segmentation of the reference map which is created on high-resolution layers; thus the higher resolution layers are more likely to exhibit strong features. It has to be noted that the features *ps_blue*/*ps_green* and *ps_vis*/*ps_nir* (forth ranked) have a negative performance outlier with a normalized score below 0.1.

Scenario *water* vs. *shadow*, object level: For this scenario, the texture features significantly contribute to the classification. The feature *GLCM Homogeneity ps_blue* obtains the highest score by far.

According the pixel level scenarios, it can be stated that ratios and differences, i. e., features that incorporate the interaction of layers, achieves the highest scores. This does not necessarily imply that a RF classifier without such artificial features would perform worse. But it shows that a good performance can be obtained on a shorter way, literally described for a greedy tree. For the purpose of a generic framework for land cover classification, a small set of features is preferable as it leads to decision rules that are simpler and better to tune.

7 Conclusion

The development of rule-based expert systems for classification is often time consuming as rules are developed by forward or backward chaining, usually accompanied by trials and errors. This process requires knowledge about the data and the underlying problem and lacks of an explicit theoretical grounding (such as provided by supervised methods). This work aims at supporting the stage of rule set development by identifying robust features for object-based land cover classification.

A multivariate feature importance analysis – based on the Random Forest classifier – has been applied on training sets of several Ikonos and QuickBird satellite scenes in order to reveal features that show a good overall performance. The method has been conducted for two common problems in land cover classification, namely the separation of *water* from *shadow* and *built-up* from *bare* surfaces.

The outcome of this study can be used to develop/improve a generic object-based land cover classification framework (rule-based expert system) by incorporating a subset of the top-ranked features. However, so far it has not been evaluated how those top-ranked features work in collaboration and how many of them are required to obtain good results. Those questions could be addressed in further research.

References

Baatz, M. & Schäpe, A., 2000: Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. – Angewandte Geographische Informationsverarbeitung **XII:** 12–23.

Benz, U., Hofmann, P., Willhauck, G., Lingenfelder, I. & Heynen, M., 2004: Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. – ISPRS Journal of Photogrammetry & Remote Sensing **58:** 239–258.

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

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., 1984: Classification and Regression Trees. – Wadsworth and Brooks/Cole, Monterey, USA.

Castilla, G. & Hay, G.J., 2008: Image objects and geographic objects. – Object-Based Image Analysis. Spatial Concepts for Knowledge-Driven Remote Sensing Applications: 91–110; Springer-Verlag, Berlin.

Chen, C., Liaw, A. & Breiman, L., 2004: Using Random Forest to Learn Imbalanced Data. – Tech. Report, www.stat.berkeley.edu/tech-reports/666.pdf.

Diaz-Uriarte, R. & Alvarez de Andres, S., 2005: Variable selection from random forests: application to gene expression data. – Tech. Report, ligarto.org/rdiaz/Papers/rfVS/rfVarSel.pdf.

Guyon, I. & Elisseeff, A., 2003: An Introduction to Variable and Feature Selection. – Journal of Machine Learning Research **3:** 1157–1182.

Haralick, R.M., Shanmugam, K. & Dinstein, I., 1973: Textural Features for Image Classification. – Transactions on Systems, Man and Cybernetics **3** (6): 610–621.

Hastie, T., Tibshirani, R. & Friedmann, J., 2009: The Elements of Statistical Learning. Data Mining, Inference, and Prediction. – Second Edition; Springer-Verlag, New York.

Herold, M., Roberts, D.A., Gardner, M.E. & Dennison, P.E., 2004: Spectrometry for urban area remote sensing - Development and analysis of a spectral library from 350 to 2400 nm. – Remote Sensing of Environment **91:** 304–319.

Ho, T.K., 1998: The Random Subspace Method for Constructing Decision Forests. – IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (8): 832–844.

Jacobsen, K., 2003: Geometric Potential of IKONOS- and QuickBird-Images. – Photogrammetric Week 2003, Germany, on CD.

Kohavi, R. & Quinlan, R. 1999: Decision Tree Discovery. – Handbook of Data Mining and Knowledge Discovery: 267–276; Oxford University Press, New York, USA.

Menze, B.H., Kelm, M., Masuch, R., Himmelreich, U., Bachert, P., Petrich W. & Hamprecht, F.A., 2009: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. – BMC Bioinformatics **10** (213).

Mingers, J., 1989: An Empirical Comparison of Selection Measures for Decision-Tree Induction. – Machine Learning **3:** 319–342.

Quinlan, J.R., 1986: Induction of Decision Trees. – Machine Learning **1:** 81–106.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A., 2008: Conditional variable importance for random forests. – BMC Bioinformatics **9** (307).

Wolf, N., Thunig, H. & Kalkan, K., 2010: Extraction of potential areas for inner-urban development. – Elsevier Editorial System for International Journal of Applied Earth Observation and Geoinformation (manuscript submitted).

Address of the Author:

MSc Geography Nils Wolf, Ruhr-Universität Bochum, Geographisches Institut, D-44801 Bochum, Tel.: +49-234-32-23380, Fax: -14180, e-mail: nils.wolf@rub.de

Manuskript eingereicht: Februar 2011
Angenommen: März 2011