OBJECT-BASED RANDOM FOREST CLASSIFICATION FOR MAPPING FLOODPLAIN VEGETATION STRUCTURE FROM NATION-WIDE CIR AND LIDAR DATASETS

L. Kooistra [1]*, E.T. Kuilder*[1] *& C.A. Mücher*[2]

1: Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands
2: Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands

ABSTRACT

Very high resolution aerial images and LiDAR (AHN2) datasets with a national coverage provide opportunities to produce vegetation maps automatically. As such the entire area of the river floodplains in the Netherlands may be mapped with high accuracy and regular updates, capturing the dynamic state of the vegetation. In this study, these fused datasets are used to map the vegetation of 936 ha of the floodplain on the north-side of the river Nederrijn near Wageningen into ten vegetation structure classes. The method follows object-based image analysis principles. Objects are defined in segmentation and subsequently labeled using the ensemble-tree classifier random forest. The mapping scale is controlled by selecting segmentation parameters from quantified discrepancies between reference polygons and segmented objects. Effects on the mapping scale of different reference polygons and different segmentation data is investigated. The results show that it is important to be able to select the right segmentation parameters to control the mapping scale. A discrepancy measure with reference polygons is a suitable method to do this objectively. The use of random forest classification on the objects resulted in an estimated classification accuracy of 86% on the basis of the built-in cross-validation estimate of random forest. Variable importance measures of random forest showed that the AHN2 lidar dataset is a valuable addition to the spectral information contained in the aerial images in the classification.

Index Terms— Object Based Image Analysis (OBIA), reference polygons, segmentation optimization, variable importance, vegetation structure classes

1. INTRODUCTION

Vegetation in river floodplains exerts friction on water and obstructs flow, limiting the capacity of floodplains to discharge water. Predictions of the discharge capacity of river systems require data on the state and distribution of river floodplain vegetation. Uncertainty in vegetation presence and state propagates to incorrect roughness coefficients. Monte Carlo simulations showed that in the case of the Dutch river system the uncertainty in vegetation maps leads to uncertainty in expected flood levels in the order of decimeters [1].

Earlier studies have shown that combining structural information from airborne Light Detection and Ranging (LiDAR) with spectral information from either airborne or spaceborne sensors has proven to be a suitable method to map and monitor floodplain vegetation for hydrological models [2-3]. Nation-wide datasets such as the high point-density elevation dataset for structural information: the Actual Dutch digital elevation model (AHN2) and national aerial photograph data archives provide opportunities for up scaling to nation-wide products. However, to take full advantage of these datasets, (semi)-automated mapping approaches are required which allow both controllability of the delineation of vegetation boundaries and of assigning labels to the delineated areas. Object Based Image Analysis (OBIA) gives the user control over the mapping scale and can handle the implicit variability that comes with very-high resolution imagery [4-5]. More important for applications where reliability is more important than accuracy, OBIA separates the identification from the classification which is in line with the manual approach of delineation of boundaries and the assignment of labels in the field.

This paper describes an objective (semi-)automated approach for object-based segmentation combined with random forest classification of very high resolution spectral and LiDAR derived elevation and surface data to map river floodplain vegetation structure.

2. METHODOLOGY

2.1. Study Area

The study area (936 ha) is located in the center of the Netherlands and covering the northern floodplains of the Nederrijn river West from the city of Wageningen to the railway bridge of the city of Arnhem in the East (Fig. 1). The western part of the area consists of managed natural vegetation, mainly grazed grass with herbaceous vegetation and patches of bush. This area also contains some old river arms and restored oxbow lakes. To the east, the parcels are larger and consist of meadows and agricultural fields. In the study area, ten main land use and vegetation structure classes are identified relevant for water discharge modeling: water, forest, orchards, bush, built-up, field, sand, herbaceous, grass, and pioneer.

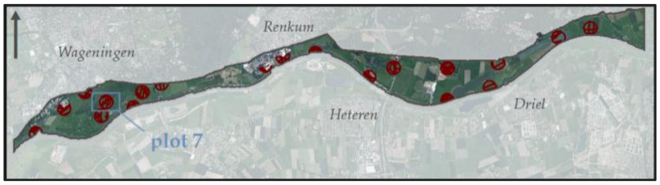


Fig. 1. True-color aerial image for 2008 of the study area from Wageningen in the West to the city of Arnhem in the East and located in the center of the Netherlands. The red circles indicate the reference polygons used to select segmentation parameters.

2.2. Datasets

Starting point for this research was to use nationally available datasets with a regular update frequency as input for the classification procedure. In this way the complete floodplain area in the Netherlands can be covered while also a monitoring approach can be developed. To map vegetation structure two relevant datasets were used as starting point for method development. Nationwide color infrared (CIR) aerial photographs (Cyclomedia) are acquired yearly with a spatial resolution of 25 cm and consisting of four spectral bands: Blue, Green, Red and Near-Infrared. In this study the CIR images acquired in 2008 were used which coincide with the acquisition of airborne laser scanning altimetry for the study area. The latter is part of a nationwide dataset called Actueel Hoogtebestand Nederland (AHN2) which has an average point density of 10 points per square meter. Filter algorithms have been applied to filter out all the points which are not part of the ground area. Both the remaining points and the filtered points are used to construct a grid with a spatial resolution of 0.5 m resulting in a Digital Elevation Model (DEM) and a Digital Surface Model (DSM), respectively. During classification an additional layer was used to characterize the vegetation height, the Digital Canopy Model (DCM).

2.3. Segmentation

The segmentation approach used for this research is Fractal Net Evolution Approach (FNEA) which can be considered as a region merging approach. It involves two steps: a multi-resolution segmentation and subsequently a spectral difference merge. Equal weights have been assigned to the CIR and AHN2 data. The spectral information of the CIR has the same influence to the increased heterogeneity of a merge of two objects as the structural information of the AHN2. Altering these weights might have a positive effect on the segmentation but this has not been pursued in this study. Within the segmentation procedure three parameters need to be set: scale, shape, and compactness. were assumed to be most important. To investigate the influence of these parameters on the segmentation result, shape (50-450 with increment of 100) and scale (10%-30% with increment of 5%) were varied while compactness was set to 50%. From the resulting 25 segmentation parameter settings, the optimal result was chosen according to the method of Möller et al. [6]. For this 19 randomly selected points were used to create circular reference plots with a radius of 150 m (Fig. 1). For all plots the average comparability (C) is calculated of the segmentation intersection to the reference (over segmentation) and to the segmentation (under segmentation). The optimum segmentation parameter is the scale and shape where under and over segmentation are equal and the comparability is highest. This point is selected from plots of over and under segmentation. Segmentation was performed in Definions eCognition developer 8.

2.3. Classification

Random forest (RF) was adopted for classification of the segmented objects into ten vegetation structure classes on the basis of 181 variables. Five categories of variables were calculated: spectral, topographical, textural and geometrical. The variables have been computed using eCognition for all the objects of the segmentations which have been classified. Gray level co-occurrence matrices and gray level difference vectors have been calculated for all the data layers used in the segmentation. The random forest is based on 113 training objects. Labeling has been done manually on the basis of visual interpretation of the RGB layers of the CIR dataset. The 113 training objects are from the segmentation of the CIR and AHN2 with scale 150 and shape 25%. The abundance of some classes (e.g., orchard, field) was low in the study area which made it not possible to get an even amount of training objects for every class. Classification through RF [7] was implemented in the statistical language R via the package randomForest [8].

3. RESULTS AND DISCUSSION

For the 25 evaluated segmentation parameter settings, the segmented objects have been intersected with the 206 reference objects of the 19 plots distributed over the study area (Fig. 1). The average relative position of the gravity centers and the average size of the intersections are compared to the segmented objects to get the under segmentation, indicated in red in Fig. 2. The blue line in Fig. 2 shows over segmentation: the comparability between the intersections and the reference polygons. As can be seen in Fig. 2 the lines intersect around the parameter setting with a scale of 150 and shape 25%. This setting shows the most resemblance to the reference plots in size, shape and position (Fig. 3).

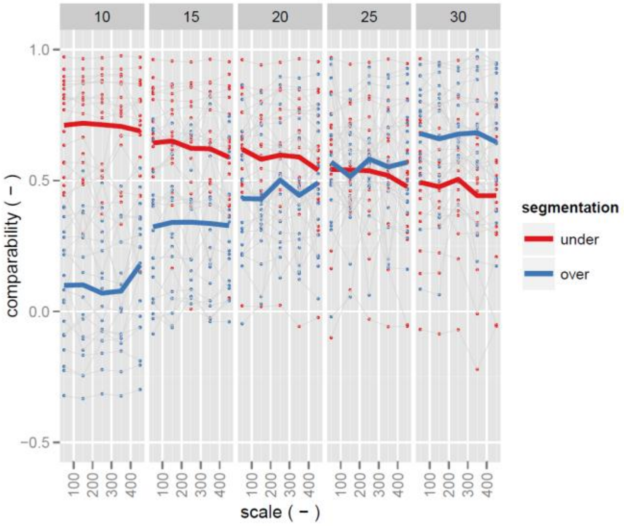


Fig. 2. Average comparability between circular reference plots and segmentation result for five shape settings between 10% till 30% (indicated at the top) and scale between 50 and 450 (indicated at the bottom) of the 2008 color-infrared and 2011 AHN2 dataset.

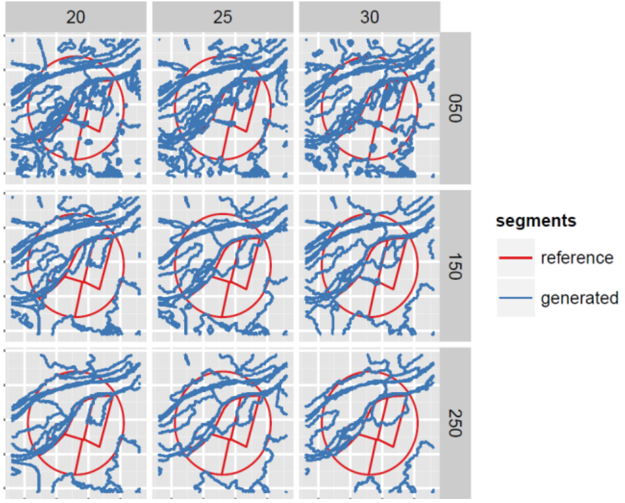


Fig. 3. Representation of segmentation results of 2008 color-infrared and 2011 AHN2 dataset with optimal (center, shape 25% and scale 150) and sub-optimal (periphery) parameter settings with the circular reference plot 7 as reference (Fig. 1).

Every subplot of Fig. 3 shows the result of one combination of segmentation parameters and corresponds to one point on the blue and red lines of Fig. 2. As such Fig. 3 gives a representative insight in how polygons created with optimal and sub-optimal parameter settings visually compare to the reference plots. With a lower scale setting there are more and smaller objects and with less emphasize on shape, these objects are more oddly shaped. Fig. 3 shows that the 150/250 scale and 25% shape gives the most comparable result. Visually this seems due to the two square reference-objects in the middle of the plot, which are reasonably bounded by the segmentation results of 150/250 and shape 25%. All results contain a lot of small noise objects and all except the 150/250 scale and shape 25% do not segregate between two square reference-objects.

Fig. 4 shows the result of a prediction of the forest on the area around plot number seven. The area is the same as in Fig. 3, though slightly larger. Fig. 4 clearly shows that the same dataset may be used to map the area for different map scales. At a scale of 50, there are a lot of small patches of grass within the herbaceous area. These areas are all mapped as herbaceous for a higher scale setting. The same is observed for patches of forest within the bush areas.

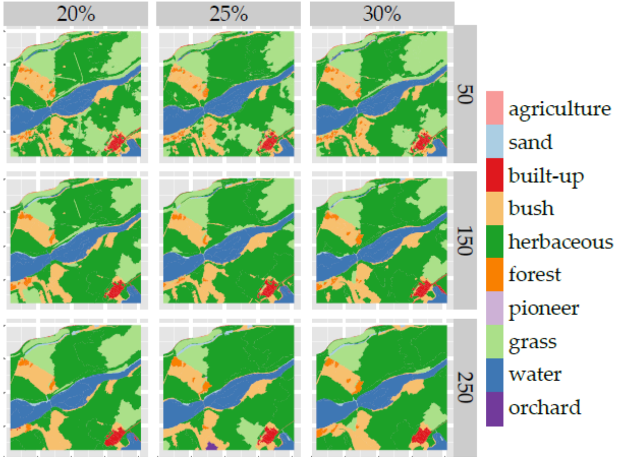
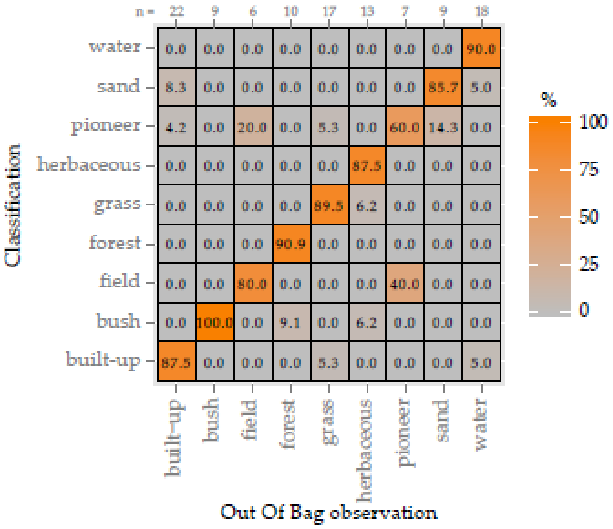


Fig. 4. Representation of classification results of 2008 color-infrared and 2011 AHN2 dataset with optimal (center, shape 25% and scale 150) and sub-optimal (periphery) segmentation parameter settings.



	built-up	bush	field	forest	grass	herbaceous	pioneer	sand	water
n=	22	9	6	10	17	13	7	9	18
water	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.0
sand	8.3	0.0	0.0	0.0	0.0	0.0	0.0	85.7	5.0
pioneer	4.2	0.0	20.0	0.0	5.3	0.0	60.0	14.3	0.0
herbaceous	0.0	0.0	0.0	0.0	0.0	87.5	0.0	0.0	0.0
grass	0.0	0.0	0.0	0.0	89.5	6.2	0.0	0.0	0.0
forest	0.0	0.0	0.0	90.9	0.0	0.0	0.0	0.0	0.0
field	0.0	0.0	80.0	0.0	0.0	0.0	40.0	0.0	0.0
bush	0.0	100.0	0.0	9.1	0.0	6.2	0.0	0.0	0.0
built-up	87.5	0.0	0.0	0.0	5.3	0.0	0.0	0.0	5.0

Fig. 5. Heatmap of the confusion matrix of the random forest based on cross-validation of a forest constructed with 1000 trees and 15 of the 181 variables deduced from the 2008 color-infrared and 2011AHN2 data.

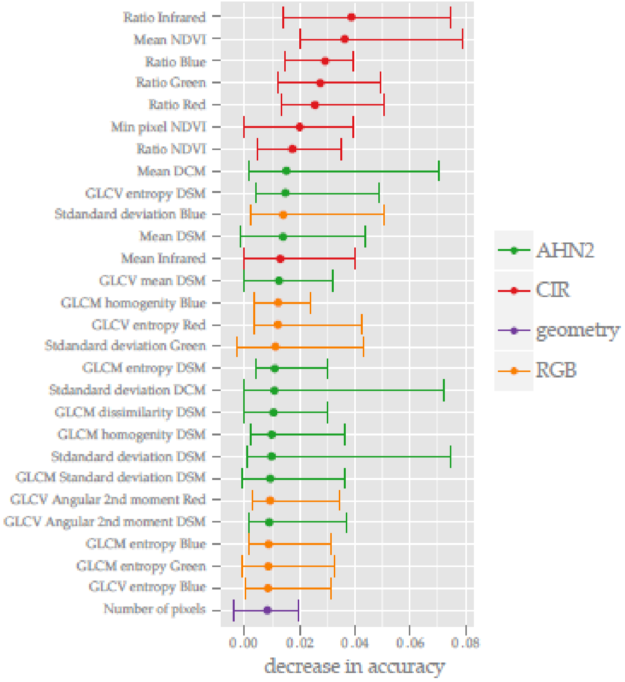


Fig. 6. Decrease in object classification accuracy for the 28 most important variables colored according to the source data layer.

Fig. 5 shows that except for the classes field and pioneer all the classes are well distinguishable and the cross-validation accuracy shows that the performance of the forest is high. On average 86% of the objects are predicted correctly.

To further investigate the value of all the variables for the classification, Fig. 6 gives the 28 most important variables in respect to the increase of object classification error as calculated by random forest. As can be expected from an image with many vegetated features, infrared is the most important variable in the classification. After around the tenth most important variable, the next variable has very low effect on the decrease in accuracy. This indicates that the variables are very correlative. Summarizing, CIR is the most valuable information source for classifying the study area in vegetation structure classes, AHN2 is a valuable dataset to include to classify specific classes and geometry of objects contains very little information.

4. CONCLUSIONS

The current study demonstrates that nation-wide CIR and LiDAR datasets may serve as input into an object-based procedure to map river floodplains into vegetation structure classes and that a high map accuracy is feasible. The empirical goodness measure used in this study has shown to be a suitable method to select segmentation parameters. By quantifying the discrepancies between reference polygons and segmentation results, the optimal scale and shape parameter may be selected. As such it is possible to objectively select appropriate segmentation parameters on the basis of reference polygons. Random forest based classification in combination with the datasets used in this study has shown to be a capable classifier. The internal validation measure of RF showed a producers accuracy of 86% on the basis of the cross-validation with the training samples. The variable importance measure of RF has shown that CIR is by far the most important source of information to distinct different vegetation objects. Including structural information of a LiDAR dataset has proven to be beneficiary, especially in the classification of high and woody vegetation such as forest and bush. Vegetation types which exert a high amount of friction on water flow. The importance measure depicted that including an IHS color transformation or geometric object features does not increase classification accuracy. While including GLCM and GLDV derived texture metrics did show to have slight positive effect on the classification accuracy. This suggests that texture of vegetated objects is a valuable part of the information contained in high resolution data.

4. ACKNOWLEDGEMENTS

This research is part of the project URD-Delta Oost funded by the program Urban Regions in the Delta of the Netherlands Organization for Scientific Research (NWO).

5. REFERENCES

[1] M. Straatsma and F. Huthoff. Uncertainty in 2d hydrodynamic models from errors in roughness parameterization based on aerial images. *Physics and Chemistry of the Earth*, 36(7-8):324–334, 2011.

[2] G. W. Geerling, M. J. Vreeken-Buijs, P. Jesse, A. M. J. Ragas, and A. J. M. Smits. Mapping river floodplain ecotopes by segmentation of spectral (casi) and structural (lidar) remote sensing data. *River Research and Applications*, 25(7):795–813, 2009.

[3] M. W. Straatsma and M. Baptist. Floodplain roughness parameterization using airborne laser scanning and spectral remote sensing. *Remote Sensing of Environment*, 112(3):1062–1080, 2008.

[4] B.N. Jyothi, G.R. Babu, and I.V.M. Krishna. Object oriented and multi-scale image analysis: strengths, weaknesses, opportunities and threats-a review. *Journal of Computer Science*, 4(9):706–712, 2008.

[5] D. Liu and F. Xia. Assessing object-based classification: advantages and limitations. *Remote Sensing Letters*, 1(4):187–194, 2010.

[6] M. Möller, L. Lymburner, and M. Volk. The comparison index: A tool for assessing the accuracy of image segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 9(3):311–321, 2007.

[7] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[8] A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.