What makes segmentation good? A case study in boreal forest habitat mapping

Aleksi Räsänen*, Antti Rusanen, Markku Kuitunen and Anssi Lensu

Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

*Corresponding author. Department of Biological and Environmental Science, P.O. Box 35, FI-40014 University of Jyväskylä, Email: t.aleksi.rasanen@jyu.fi

What makes segmentation good? A case study in boreal forest habitat mapping

Segmentation goodness evaluation is a set of approaches meant for deciding which segmentation is good. In this study, we tested different supervised segmentation evaluation measures reviewed by Clinton et al. (2010) and visual interpretation in the case of boreal forest habitat mapping in Southern Finland. Used data were WorldView-2 satellite imagery and LiDAR digital elevation model (DEM) and canopy height model (CHM) in 2-m resolution. Tested segmentation methods were Fractal Net Evolution Approach (FNEA) and IDRISI watershed segmentation. Overall, 252 different segmentation method, layer and parameter combinations were tested. We also used eight different habitat delineations as reference polygons against which 252 different segmentations were tested. The ranking order of segmentations depended on the chosen supervised evaluation measure; hence, no single segmentation could be ranked as the best. In visual interpretation, we found out that several different segmentations were rather good and selected one of them as the best. In literature, it has been noted that better segmentation leads to higher classification accuracy. We tested this argument by classifying 12 of our segmentations with the random forest classifier. It was found out that there is no straightforward answer to the argument, since the definition of good segmentation is inconsistent. Highest classification accuracy (0.72) was obtained with segmentation which was regarded as one of the best in visual interpretation. However, almost as high classification accuracies were obtained with other segmentations. We conclude that one has to decide what she/he wants from segmentation and use segmentation evaluation measures with care.

1 Introduction

Since the early 2000s, with the rise of object-based image analysis (OBIA) methodology (Blaschke 2010), segmentation goodness evaluation has been an emerging topic within the remote sensing literature (Clinton et al. 2010; Marpu et al. 2010). Evaluation has been concentrated on segmentation method development and comparison as well as parameter optimization.

Generally in segmentation, the goal is to partition imagery into regions that are meaningful; and thus, either mimic real world objects (Zhang, Fritts, and Goldman 2008; Clinton et al. 2010) or minimize intrasegment and maximize intersegment heterogeneity (Zhang, Fritts, and Goldman 2008; Hou et al. 2013). Remote sensing segmentation methods are a special case of more general image segmentation methods, which can be divided into two complementary groups: similarity or region based segmentation and discontinuity based segmentation. In region based segmentation, a similarity measure is used to find suitable regions. In discontinuity based segmentation, discontinuities of the images, usually boundaries, are detected (Zhang 1997; Gonzales and Woods 2002). Some of the methods combine concepts from both groups. For instance, in watershed segmentation, dividing lines between basin areas are sought by flooding the image (Gonzales and Woods 2002).

Within remote sensing, many of the segmentation implementations have been region based. Arguably, the most widely used remote sensing segmentation method has

been Fractal Net Evolution Approach (FNEA) developed by Baatz and Schäpe (2000), and implemented in eCognition software. FNEA has been used as a benchmark segmentation, against which other methods have been compared. Although some authors have claimed to have developed better methods (Derivaux et al. 2010; H. Li et al. 2010; N. Li, Huo, and Fang 2010; Z. Wang, Sousa, and Gong 2010), FNEA has been a good performer in method comparisons (Neubert and Meinel 2003; Meinel and Neubert 2004; Carleer, Debeir, and Wolff 2005; Neubert, Herold, and Meinel 2008; Marpu et al. 2010) and seems to still be the standard method (e.g. Bar Massada et al. 2012; Duro, Franklin, and Dubé 2012). Also many other standard remote sensing analysis products such as ENVI, ERDAS Imagine and IDRISI Selva have included segmentation methods in their newer versions. Yet, there are also numerous other methods, algorithms and software applications for segmenting remotely sensed data. To analyse the goodness of these methods, segmentation evaluation has been performed (Zhang 1996; Clinton et al. 2010; Marpu et al. 2010).

Segmentation evaluation can be divided into two major categories: subjective (visual) and objective evaluation. Objective evaluation can be further divided into system-level, which evaluates the overall system in which segmentation is performed, and direct evaluation (Zhang, Fritts, and Goldman 2008). As an example, final classification output can be regarded as a system when assessing segmentation quality (sensu H. Li et al. 2010; Smith 2010; Z. Wang, Sousa, and Gong 2010; Gao et al. 2011). Direct evaluation can be either analytical or empirical, of which the former evaluates the method itself and the latter its results. Empirical methods consist of supervised and unsupervised methods, i.e. if ground truth is used as a reference or not (Zhang, Fritts, and Goldman 2008).

In remote sensing, analytical methods (Hay et al. 2003) and unsupervised methods (Espindola et al. 2006; Corcoran, Winstanley, and Mooney 2010; Drăguţ, Tiede, and Levick 2010; Yue et al. 2012; Hou et al. 2013) have been used in segmentation evaluation. For instance, Corcoran, Winstanley, and Mooney (2010) evaluate segmentation goodness by measuring contrast between segments that share a boundary. However, evaluation has largely been performed using supervised methods, more specifically either area based or location based measures. From these two, area based measures evaluate either if segmentation is too coarse (undersegmentation) or too fine (oversegmentation). Over- and undersegmentation measures can also be combined. Location based measures, on the other hand, are based on distances between segment centroids and reference polygon centroids or distances between boundary pixels. For a good review of these measures and an evaluation of different measures, see Clinton et al. (2010). These goodness measures are applicable especially in mapping clearly bordered urban features (Tian and Chen 2007; Zhan et al. 2005; Weidner 2008; Clinton et al. 2010), agricultural areas (Lucieer and Stein 2002; Möller, Lymburner, and Volk 2007; Z. Wang, Sousa, and Gong 2010) or larger land use / land cover types (Weidner 2008). Yet, goodness measures have also been used in natural area segmentations (Carleer, Debeir, and Wolff 2005; Ke, Quackenbush, and Im 2010; Bar Massada et al. 2012). Some of the supervised segmentation evaluation methods use a larger set of reference polygons inside a larger area (e.g. Clinton et al. 2010) whereas some of the methods use only a couple of distinct reference polygons and semi-automated approaches (e.g. Marpu et al. 2010).

Segmentation evaluation can be used to compare different types of segmentation methods, i.e. boundary against region-based (Carleer, Debeir, and Wolff 2005). As well,

evaluation can be made between different segmentation methods or software, or inside a segmentation method as parameter optimization (Marpu et al. 2010). Although similar methods can be used in all these problems, also task-specific methodology for the problems has been developed, especially to parameter optimization. For instance, genetic algorithms have been used in optimizing segmentation to match reference delineation (Feitosa et al. 2006; Chabrier et al. 2008). Optimization has also been made without supervised goodness measures based on unsupervised evaluation. For instance, Drăguţ, Tiede, and Levick (2010) developed a scale parameter optimization tool which measures rate of change of local variance inside a scene. Optimal scale parameters are those which have local maximum of the rate of change. In a bit similar vein, Espindola et al. (2006), Gao et al. (2011), and Yue et al. (2012) tried to combine low intrasegment variance and low intersegment autocorrelation. On the other hand, Kim, Madden, and Warner (2008, 2009) hypothesized that optimal scales should only have low spatial autocorrelation between segments; whereas L. Wang, Sousa, and Gong (2004) maximized Battacharya distance between candidate segments. Finally, Smith (2010) optimized segmentation scale by minimizing classification error in a random forest classifier.

Forest inventory or forest habitat mapping is only one instance of where segmentation is often used. Yet, forest inventories are more and more dependent on automatic segmentations (Pekkarinen 2002; Hay et al. 2005; Castilla, Hay, and Ruiz-Gallardo 2008; Mustonen, Packalén, and Kangas 2008; Wulder et al. 2008; Falkowski et al. 2009; Kim, Madden, and Warner 2009; Ke, Quackenbush, and Im 2010; Hou et al. 2013). Segmentations used in forest inventory are usually made with feature values calculated from aerial or satellite images; however, the usage of light detection and ranging (LiDAR) data has recently become popular (Mustonen, Packalén, and Kangas 2008; Ke, Quackenbush, and Im 2010; Breidenbach et al. 2011; Eysn et al. 2012; Hou et al. 2013). It has been noted that synergy of imagery and LiDAR provide promising segmentation results and the selection of input data has an effect on segmentation quality (Geerling et al. 2007, 2009; Mustonen, Packalén, and Kangas 2008; Ke, Quackenbush, and Im 2010; Hou et al. 2013). Despite of this, studies incorporating different datasets remain scarce; both in forest inventory and in other applications.

In forest inventory or habitat mapping, segmentation goodness evaluations have been performed both qualitatively (Leckie et al. 2003; Wulder et al. 2008) and quantitatively using unsupervised (Kim, Madden, and Warner 2008, 2009; Hou et al. 2013) or supervised methods (Radoux and Defourny 2007; Ke, Quackenbush, and Im 2010). Some evaluations have been based on thematic quality of segments against reference polygons (Pekkarinen 2002; Mustonen, Packalén, and Kangas 2008). Wulder et al. (2008) criticize quantitative evaluation because used ground truth is a subjective delineation of forest patches; hence, real truth does not exist. Therefore, objects in forests are not as clearly separable as e.g. urban features. Furthermore, also in urban features, supervised evaluation has been criticized due to inaccurate ground truth (Corcoran, Winstanley, and Mooney 2010). In this work, we wanted to test if supervised segmentation evaluation methods are applicable to forested areas on a larger set of reference polygons.

It has been stated and tested that segmentation goodness affects directly classification accuracy in OBIA classification (Kim, Madden, and Warner 2009; Clinton et al. 2010; Ke, Quackenbush, and Im 2010; Gao et al. 2011). Although there are many different approaches and measures for segmentation goodness evaluation, the

evaluation of goodness measures has not been thorough. In this paper, we will test if widely used supervised segmentation goodness measures are applicable in boreal forest habitat type mapping and if best segmentation leads to best classification accuracy. We test supervised methods instead of unsupervised methods, because our goal is to find a segmentation which matches with habitat type patches that are delineated using field work. Furthermore, we try to find a segmentation method, parameter value and image/data layer combination which suits our purposes. To do this, we test two different methods (FNEA region based segmentation and IDRISI watershed segmentation), several parameter combinations and different layers derived from WorldView-2 imagery and LiDAR data. We also test if different methods are habitat type, reference polygon, or area sensitive.

2 Methods

2.1 Study area and reference polygons

We studied a 7 km^2 rural-forested area southwest of the city of Jyväskylä located in Southern Finland. The area belongs to southern boreal vegetation zone (Ahti, Hämet-Ahti, and Jalas 1968). The geographic coordinates (WGS84) of the site are 62° 10´30´´–62° 13´30´´ N and 25° 29'0´´–25° 38´0´´ E. The study area mainly consists of both coniferous and deciduous forest habitats, mires and agricultural area. The main tree species of the study area are the Scots Pine (*Pinus sylvestris*), Norwegian Spruce (*Picea abies*) and Birches (*Betula pubescens* and *B. verrucosa*). The study area was divided into three sub-areas with slightly varying land cover. The sub-areas were classified into 25 different habitat types (Table 1), which were mapped by field work during June-August 2011. Habitat patches were used as reference polygons in segmentation goodness measures.

(Table 1 should be inserted here)

Most of the studied forest area is under heavy forestry and clear-cuts which create a human induced dynamic. Yet, two of the three delineated sub-areas included also one protected area which covered 100 ha and 25 ha of these sub-areas. Protected areas were dominated by semi-natural, over 100-year-old, forest. The larger protected area is part of a NATURA 2000 area. Inside the NATURA area and our study area, several different NATURA 2000 habitats are found. NATURA 2000 habitats were not mapped per se, but they were included in some of our mapped habitat types.

The three studied sub-areas were selected from different parts of a larger area southwest of Jyväskylä so that they included many different habitat types and different landscape configurations. Each sub-area was delineated into habitat patches whom there were 628 in total. Sub-area 1 (Sallaajärvi) included small to medium sized patches of different age, mostly mesic forest, some spruce mires, small streams, meadows, lakes and yards. In the area, there are also some old fields, which have been afforested. Moreover, in the middle of the area, there is a 250 ha conservation area with semi-natural forest. Sub-area 2 (Kuusimäki) included large areas of protected old mesic forest with spruce mire patches. As well, this sub-area had some open and pine mires, small lakes and fields, yards, and different aged forest around the old forest. Sub-area 3 (Lapinmäki) included areas of bare rock surrounded by mesic forest, and with yards, fields and lakes on the fringes.

During field work, patches were drawn into paper printouts of orthophotos, which included 5 meter contour lines derived from a topographic map. Additionally, Trimble GeoXT and Juno SB GPS devices with differential location correction were used for checking accurate location and in delineating patches, which were difficult to distinguish from aerial images. ArcGIS 9.3.1 editor was used when patches were manually drawn into digital format. Patches were initially mapped as they were in the terrain. Afterwards, some recent clear-cuts were modified to be of the same age and forest type as neighbouring forest patches to match the state of the forest in the used satellite image and LiDAR data.

As alternative reference polygons, we used a forestry planning dataset created for the City of Jyväskylä and a biotope classification dataset created by Finnish Forest and Park Service (FFPS) (Vesterbacka 2010). In forestry planning dataset, polygons are drawn first from aerial imagery and after initial drawings; polygons are double checked using field work. This data was from sub-area 1 only. . FFPS biotope data is also generated using field work and aerial imagery and was from sub-area 2 only.

2.2 Remotely sensed data

Our primary data consisted of 8-band multispectral 2-meter resolution WorldView-2 (WV-2) satellite image taken in July 14th 2010 and LiDAR data with a minimum of 0.5 points per 1 m^2 from May 2010. Additionally, we used 20 cm resolution aerial images (orthophotos) taken in 2007 in assisting the drawing of reference polygons.

WV-2 image, taken by Digital Globe Inc., consists of 8 bands: coastal blue (band 1, 400–450 nm), blue (2, 450–510 nm), green (3, 510–580 nm), yellow (4, 585–625 nm), red (5, 630–690 nm), red-edge (6, 705–745 nm), NIR1 (7, 770–895 nm), NIR2 (8, 860–1040 nm) in 2 meter resolution and a panchromatic band (450–800 nm) in 50 cm resolution. Image was delivered radiometrically and sensor corrected, projected to a plane with average terrain elevation. In our preprocessing phase, the image was first orthorectified using 5-meter resolution digital elevation model derived from LiDAR data. In georeferencing, 13 ground control points from block features (buildings etc.) which were scattered all over study area were taken from orthophotos and nearest neighbour sampling was used. In visual interpretation, the differences between orthophotos, LiDAR data and orthorectified WV-2 were at maximum a couple of meters. From WV-2, we used all multispectral bands in 2 m resolution.

LiDAR data was created by National Land Survey of Finland. Flying altitude is on average 2000 meters. Used scan angle was ± 20˚ and laser pulse footprint on the ground approximately 50 cm. Mean error in elevation information is at maximum 15 centimetres and in planar information at maximum 60 cm. Data was delivered automatically classified to ground hits, low vegetation hits, low error hits and unclassified hits.

LiDAR point clouds were first triangulated and after that rasterized using LASTOOLS (Isenburg 2011). We first derived two layers in 2 m resolution from LiDAR: digital terrain model (DTM) and digital surface model (DSM). In DTM, only ground hits were used whereas in DSM, point cloud was first thinned to one meter resolution to include highest hits. Then we subtracted DTM from DSM to create canopy height model (CHM). CHM was further manipulated to include values only between 0 and 40 meters to filter out unrealistic values. The CHM still had some wrong values below 40 meters but these could not be corrected easily.

From DTM, we calculated also Saga Wetness Index (SWI) in 2 m resolution using SAGA-GIS to model soil moisture; and thus, potential places for mires. SWI is a modification of topographic wetness index (TWI). It has been noted that in wetland mapping standard TWI performed worse than some other models; however, mainly because in these studies TWI underestimated extent and contiguity of wetlands (Grabs et al. 2009, Murphy, Ogilvie, and Arp 2009). This might be due to that the standard TWI concentrates large values to stream networks where water flow is concentrated. This underestimation problem is overcome in SWI, which assumes homogenous hydrologic conditions in flat areas and predicts larger moisture values for cells with small vertical distance to streams (Böhner and Selige 2006, Equations 1 and 2).

$$\alpha_M = \alpha_{max} t^{-\beta \exp(t^\beta)} \quad \text{for} \quad \alpha < \alpha_{max} t^{-\beta \exp(t^\beta)} \tag{1}$$

Specific catchment area (α) used in TWI is defined as the pixels upslope contributing area per contour width whereas α_M is modified catchment area used in SWI. In calculating α, slope angle β (in radians) and neighbouring cell maximum α_{max} are taken into account unless results remain unchanged. Parameter t is a value for suction, so that lower values, e.g. under 10, lead to stronger suction and stronger spreading of large α values, and higher values lead to weaker suction. After counting α_M, SWI is calculated with the standard equation given in Equation (2).

$$\text{SWI} = \ln\left(\frac{\alpha_M}{\tan\beta}\right) \tag{2}$$

Before calculations, DTM was filled to remove uncertainties, missing values and false values from the data. Before the filling, values in DTM in known and evident places of bridges and culverts were manipulated to let imagined water to flow through road banks in those locations. To angle β, 0.0174532 rad was added so that division by 0 was avoided. In flow direction calculations, we used multiple flow direction method by Freeman (1991). In this method, the slope value is raised to the power of 1.1. Thus, steeper slopes are weighted only a bit. It has been noted that in relatively flat areas multiple flow direction methods, in which slope value is raised by a low exponent (e.g. 0.5 to 2), give good results in TWI calculation (Güntner, Seibert, and Uhlenbrook 2004; Sørensen, Zinko, and Seibert 2006; Kopecký and Čížková 2010). Furthermore, parameter t in Equation 1 was decided to be default 10 after visual interpretation of SWI with different t values.

Before segmentation, SWI was quantized to 32 classes using equal intervals and CHM was quantized to 40 classes (nearest integer). WV-2 layers were first filtered using a 3×3 window and a median filter. After filtering, layers were quantized to 256 classes.

2.3 Segmentation methods

Data was segmented using different datasets, methods and parameters. To compare different types of segmentation methods, two segmentation methods were used: one watershed segmentation and one region based segmentation method. Next, brief introductions of the used segmentation methods and their parameters are given.

Watershed segmentation was implemented in IDRISI Taiga software. In IDRISI segmentation, a variance image is derived from each layer by moving window analysis.

A weighted average of variance images is the final surface image for watershed delineation. Both the size of the moving window as well as the weights of averaging can be adjusted by the user. The values of this surface image are treated as elevation values like in a DEM, and pixels are grouped into watersheds. After watershed delineation, watersheds are merged iteratively. Pairs of segments are merged if they are most similar segments to each other in the neighbourhood and if their difference is smaller than a similarity tolerance adjusted by the user. Difference is evaluated by two aspects: the mean value and the standard deviation. The weights for the mean and for the standard deviation are set by the user.

Our region based segmentation was the widely used segmentation method of eCognition software, Fractal Net Evolution Approach (FNEA) (Baatz and Schäpe 2001; Benz et al. 2004). FNEA segmentation was carried out using TerraLib 4.2.0 C++-GIS-library (Câmara et al. 2008). In FNEA, regions are formed by merging pixels; i.e. in the beginning, each pixel is treated as a region. In segmentation, three user parameters can be adjusted: scale parameter and weights between colour and shape ($w_{color} + w_{shape} = 1$) as well as smoothness and compactness ($w_{smooth} + w_{compt} = w_{shape}$). Scale parameter controls the average object size. The more weight is given to colour (or spectral) homogeneity, the less weight is given to a specific shape i.e. spatial homogeneity. Smoothness and compactness define the shape as follows. Smoothness is the ratio of the border length of the segment and border length of the bounding box of the segment. Compactness, on the other hand, is the ratio of the border length of the segment and the square root of the number of pixels in the segment. Hence, they are not antagonistic but the weight is defined between them. Finally, the weights for the different layers are set by the user.

2.4 Initial work for segmentation goodness evaluation

In segmentation, several issues affect the final segmentation goodness: segmentation method, parameterization including weights for the layers (e.g. Marpu et al. 2010), used layers (e.g. Ke, Quackenbush, and Im 2010), (re)classification of the layers, transformations made for the layers and filtering of the layers (e.g. Carleer, Debeir, and Wolff 2005). Easily thousands of different combinations can be tested. Therefore, we first did initial trial-and-error testing and visual interpretation for different types of segmentation. We segmented single layers, reclassified and filtered the layers and tried different parameter combinations and segmentation methods. In our initial analysis, the goal was to find good segmentation methods that could be further evaluated using the evaluation measures. As well, we wanted to scale our layers so that they could be used in same segmentations; in other words, the segments that are produced in single layer segmentations should be approximately of same size. Needless to say, our initial evaluation was not thorough but it was good enough to find good segmentation methods. All possible combinations could not be tested but we found a set of segmentations that were probably among the best that are available.

2.5 Used parameter and layer combinations

Four different layer combinations were tested: (a) WV-2 layers only, (b) LiDAR layers only, (c) WV-2 bands 2, 3, 5, 7 (blue, green, red, NIR1) and LiDAR layers, and (d) all layers. IDRISI segmentation was performed using a window size of 5. Similarity

tolerance was varied between 20 and 70 with intervals of 5. Three different combinations of mean and variance weights were used: mean 0.5, variance 0.5; mean 0.9, variance 0.1; and mean 0.1, variance 0.9. Hence, overall 33 IDRISI segmentations were performed for all layer combinations. FNEA segmentation was performed by varying the scale parameter between 5 and 50 with intervals of 5, and using colour parameter values of 0.25, 0.5 and 0.75. Therefore, 30 different FNEA segmentations were done for all layer combinations. In all segmentations, all layers were given equal weight.

2.6 Different reference polygons

We tested different segmentation methods using eight different reference polygon sets. First, all reference polygons from the whole study area were used. Second, three sets included all reference polygons from three different sub-areas separately. Third, two sets included reference polygons of only one habitat type: one set included all mires and one set water. Finally, we tested segmentation quality against two other reference polygon sets (FFPS biotope and forestry planning data) (Figure 1).

(Figure 1 should be inserted here)

2.7 Goodness evaluation measures

Segmentation goodness was evaluated using several different supervised measures (Table 2) reviewed by Clinton et al. (2010) with a Java tool that they developed. For more clarification and equations, please refer to Clinton et al. (2010) and original publications listed in Table 2. All measures were calculated as a mean of all reference polygons inside a reference polygon set. The value for a specific reference polygon was calculated as a mean (or standard deviation) of the values of those segments that met at least one out of four criteria. Criteria were: (1) the centroid of the segment is inside the reference polygon, (2) the centroid of the reference polygon is inside the segment, (3) the shared area of the segment and the reference polygon is over 0.5 of the segment area, and (4) the shared area of the segment and the reference polygon is over 0.5 of the reference polygon area (Clinton et al. 2010). Some of the measures were weighted by the reference objects (Table 2). Furthermore, we calculated combined measures which were proposed by Clinton et al. (2010) and which all included measures from single authors only (Table 3). Some of the combined measures were calculated as root mean square (RMS) individual criterion values whereas some of them were simple sum calculations. In RMS calculations, all measures were adjusted so that ideal segmentation was set to 0. Finally, a combined measure COMBINED was calculated which was a RMS of all basic area and location-based measures as suggested by Clinton et al. (2010). However, QLoc was not included since it was the same measure as RPsub. Before RMS calculation in COMBINED, all measures were scaled to [0,1] by dividing each value with the maximum and setting ideal segmentation to 0.

(Tables 2 and 3 should be inserted here)

Furthermore, we measured segmentation goodness using visual interpretation. In visual interpretation, we paid detail especially to if the segmentation methods find the boundaries of some reference polygons and habitat types. Hence, we were more worried about undersegmentation than oversegmentation. Additionally, we checked if different kinds of habitat types are segmented and if the segmentation produces objects that are

meaningful entities and can be easily used in classification and planning (Hay et al. 2005). Therefore, segments should not be too complex (Mustonen, Packalén, and Kangas 2008). Due to the large number of different segmentations, our visual interpretation was not thorough; instead, we tried to find some general trends from different segmentation methods as well as layer and parameter combinations.

2.8 Classifications

After segmentation evaluation, we selected 12 segmentations for classification. Segmentations were selected using subjective evaluation; so that meaningful evaluation of segmentation performance versus classification accuracy could be made and some of the segmentations could be compared to each other. Both good and not as good segmentations, based on evaluation measures and visual interpretation, were selected. In classification, we calculated mean values of each layer per segment. In all classifications, all layers were always used regardless of which layer combination a-d was used in the segmentation phase.

Supervised classification was performed using the random forest classifier (Breiman 2001) with R package randomForest (Liaw and Wiener 2004) in R version 2.15.2 (R Development Core Team 2012). Random forest classification has been used in remote sensing and OBIA with good experience (Lawrence, Wood, and Sheley 2006; Rodriguez-Galiano et al. 2012). Random forest is an ensemble classifier, which combines several bootstrapped classification trees. In the final classification majority vote over all trees is made. Trees are randomized at each node by selecting only a subset of variables of which the best split is chosen. When a tree is built, approximately 2/3 of the data is selected for training the classifier and the rest is called out of bag (OOB) test data. OOB data is used for error rate estimation, which is averaged over all trees to get an error rate for the whole classification. Because of the OOB, independent test data or cross-validation is not needed when random forest is used (Breiman 2001; Breiman and Cutler 2007) which has been confirmed in remote sensing studies (Lawrence, Wood, and Sheley 2006; Rodriguez-Galiano et al. 2012).

When random forest was performed, 500 trees were built and the number of features at each split was given the default value of square root of all features. We used our own reference polygons over all three sub-areas as training data, not the FP nor the FFPS data. Training set in random forest run was all those segments that had a minimum of 60 % coverage of one reference habitat type. Classification accuracies were calculated using all reference polygons with simple cross-tabulation matrices.

3 Results

3.1 Segmentation goodness based on evaluation measures

Based on all area and COMBINED measure, best segmentation was FNEA with layer set b, scale parameter 25 and colour parameter 0.5 (Table 4, Table 5). However, choosing this segmentation as the best was contradictory, since no other segmentation evaluation measure ranked it as the best method. As well, its rank was between 4 and 94 when COMBINED measure and other reference polygons than all area were used. Hence, different segmentations were chosen as the best or being among the best, when different goodness measures or reference polygons were used (Table 4). Some of the

measures (UnderMerging, OverMerging, CountOver, SimSize sd, RAsuper, RAsub, OverSegmentation, UnderSegmentation), nonetheless, gave rather consistent results, i.e., the same segmentation was the best or one of the best using different reference polygon sets. Other measures, on the other hand, had larger variation in their results. Consistency in results can be seen as a downside, since reference sets were different as illustrated in Figure 1. For instance, individual measures may prefer segmentation result, which is as fine or as coarse as possible. On the other hand, consistency can also be seen as an asset if some of the segmentations truly are better despite of the reference set used, i.e. those segmentations contain almost all meaningful patch boundaries.

(Tables 4 and 5 should be inserted here)

Some overall evaluations can be made from segmentations ranked as the best (Table 6). First, FNEA segmentation outperformed IDRISI segmentation, since FNEA was ranked best 140 times against 44 times of IDRISI. Second, layer set b outperformed other layer sets. Therefore, it could be thought that FNEA with layer set b provides the best results. If undersegmentation is wanted to be avoided, low scale parameter brings good results. Vice versa, high scale parameter should be selected when oversegmentation is not desired. Segmentations with intermediate scale or similarity parameter value were not ranked as best as often as segmentations with high or low parameter value. Yet, different combined measures as well as AFI, RP measures, SimSize mean, QLoc mean, QLoc sd and QR usually preferred intermediate scale parameter values. For instance, when all area reference polygons were used, the COMBINED measure favoured intermediate scale parameter values; whereas it ranked those segmentations with low scale parameter value as the worst (Table 5). In FNEA segmentations, high value for colour parameter gave more often best segmentations than low or intermediate value for colour. In IDRISI segmentations, on the other hand, high mean, low variance combination gave the largest number of best segmentations.

(Table 6 should be inserted here)

When correlations between different goodness measure results were evaluated (Table 7), it was found out that correlations range from large negative correlations to high positive correlations. Hence, measures did give different results and preferred different issues in segmentation. It can also be seen that measures that measure oversegmentation had positive correlations with the COMBINED measure whereas undersegmentation measures had negative correlations (for over- and undersegmentation measures, see Table 2). Some measures (RPsuper, MergeSum, M, ZH1) had even both positive and negative correlations. Correlations were dependent on reference polygons used; but correlations between the COMBINED measure based on different reference polygons were rather high and positive (Table 8). Only water has correlations below 0.75.

(Tables 7 and 8 should be inserted here)

3.2 Segmentation goodness based on visual interpretation

In visual interpretation, it was found out that segmentations based on layer set b (LiDAR data only) were especially successful in delineating mires and small streams. Also the problem of shadow effect in WV-2 imagery was overcome when LiDAR data was used. On the other hand, the shorelines of water bodies were insufficiently delineated with LiDAR data only. As well, boundaries between deciduous and coniferous forests were better delineated using WV-2 imagery. However, more gradual

boundaries, for instance between mesic and xeric forests, could not be easily segmented using any method or layer combination. In visual interpretation, we could not make a preference between layer sets c and d. Although segmentation outputs were slightly different, differences were minor. Same kinds of observations were made, when different IDRISI mean/variance weight alternatives were compared. Furthermore, to delineate some small objects, small values of scale or similarity parameters were needed. In finding meaningful and simple entities, it was found out that FNEA segmentation with low (0.25) or intermediate (0.5) weight for colour brought superior results over other segmentation methods. Putting little weight to colour had its downside, on the other hand. In other words, segment boundaries did not necessarily follow natural or data boundaries but segments were equally sized objects with often arbitrary boundaries. Nevertheless, FNEA segmentations with large weight for colour and IDRISI segmentations were unnecessarily complex. Additionally in IDRISI segmentations, boundaries were often crisscrossing reference polygon boundaries. Using visual interpretation, we chose FNEA segmentation with layer combination c, scale parameter 10 and colour parameter 0.5, as the best one (Figure 2c). This selection was, yet, more or less arbitrary, since many different segmentation options gave quite similar results. Furthermore, since there were so many different segmentation options, visual interpretation was not thoroughly reliable in finding the best parameter values. Hence, choosing the best segmentation using visual interpretation was tricky.

3.3 Classification results

Classification accuracies between classifications derived from different segmentations varied a bit (Table 9, some of the segmentations in Figure 2). Best accuracy (0.72) was achieved using best segmentation in visual interpretation (Figure 2c) whereas worst accuracy (0.60) was got using segmentation that was ranked high using some of the measures (Figure 2h). Many of the different segmentations got reasonably good results compared to the best classification. Some of these segmentations were selected based on measures, some by using visual interpretation. On the other hand, best segmentation based on COMBINED measure and all area (Figure 2e), was not among the best segmentations in classification accuracy analysis. It can be seen that fine or moderately fine segmentations led to better classification accuracies. Vice versa, coarse segmentation led to poorer accuracies. On the other hand, too fine segmentation can lead to salt-and-pepper effect (Figure 2b) and thus possibly also make classification accuracy worse. Also in visual interpretation it became evident that classifications performed with FNEA segmentations and scale parameter value 5 suffered from this effect more than classifications performed with segmentations with scale value 10. Best classification accuracies were achieved using segmentations with both LiDAR and WV-2 layers. This might be due to the fact that boundaries were best detected using both data types in segmentation. Yet, classification accuracies using only segmentations performed with LiDAR or WV-2 data were nevertheless little worse. In all, classification accuracy evaluation was not thorough; in other words, good classification accuracies can be got using segmentations, which were not among the 12 segmentations tested here. Moreover, some measures may be good in selecting segmentations that maximize classification accuracy.

(Table 9 and Figure 2 should be inserted here)

4 Discussion

4.1 Segmentation goodness compared to classification accuracy

One of our study objectives was to test if better segmentation leads to better classification accuracy as it has been argued by others (Kim, Madden, and Warner 2009; Clinton et al. 2010; Ke, Quackenbush, and Im 2010; Gao et al. 2011). After our analysis, it is obvious that there is no straightforward answer to this question. Although it is self-evident that good segmentation is needed for good classification, there is no adequate definition of what makes a good segmentation. After classification analysis one can easily state that the best segmentation was the segmentation with the best classification accuracy. There is no method, however, to test before classification, which segmentation will give the best classification accuracy. This is illustrated also by the studies of Kim, Madden, and Warner (2009) and Gao et al. (2011). While they both claim that optimal segmentation produced best classification output, their definitions of optimal segmentation were contradictory. Kim, Madden, and Warner (2009) minimized spatial autocorrelation between different segments, whereas Gao et al. (2011) sought for segmentations that combined low intersegment autocorrelation and intrasegment variance. Furthermore, Gao et al. (2011) had the lowest intersegment autocorrelation at the coarsest scale which did not produce best classification accuracy. On the other hand, the tasks in these studies were different, since Kim, Madden, and Warner (2009) used 4-m resolution IKONOS data in forest type mapping, while Gao et al. (2011) used 25-m Landsat ETM+ data in mixed mountainous shrub-forest-grassland landscape.

Based on ambivalence of what segmentation is good, we propose that the goodness of segmentation should be defined in each case. In other words, one should know and clarify what he/she wants from segmentation; and critically evaluate if the best segmentation can be selected based on evaluation criteria. In our case, good segmentation was segmentation with (1) meaningful and not too complex segments, (2) boundaries parallel to reference polygon boundaries even for the smallest reference polygons but (3) as coarse as possible. Taking the best segmentation based on some measure does not automatically lead to the best classification accuracy, as it has been already noted by Verbeeck, Hermy, and van Orshoven (2012). Nevertheless, the classification accuracies between different classifications were rather small in our case study. This might point out to the robustness of OBIA methodology: good classification accuracy can be obtained even if the segmentation is not the best possible. On the other hand, the classification outputs that had better classification accuracies were visually more appealing. Boundaries were more often in right place, different habitat types could be mapped and patches were not too small.

Many authors have argued that oversegmentation is a smaller problem than undersegmentation in post-segmentation classification (e.g. Weidner 2008; Marpu et al. 2010). However, in the analysis by Verbeeck, Hermy, and van Orshoven (2012), it was found out that more under-segmented output gave better classification accuracy than more over-segmented output. Our results suggest that both arguments are partly correct. In other words, both oversegmentation and undersegmentation are problematic in classification as it was found out also by Kim, Madden, and Warner (2009) and Gao et al. (2011). First, segmentation cannot be very coarse, since smaller objects are thus easily under-segmented. Second, if segmentation is too fine, salt-and-pepper effect is obtained. Salt-and-pepper effect can lead to worse classification accuracy as it has been

found out in OBIA vs. pixel-based classification studies (e.g. Bock et al. 2005; Whiteside, Boggs, and Maier 2011). Also, objects are more meaningful when they are not too small (Blaschke 2010). Yet, the better classification accuracy of OBIA is not automatic and some smaller, but rare, objects may be easily missed in OBIA classification (Dingle Robertson and King 2011). Overall, it has been noted that in single-scale segmentation optimal segmentation is class-dependent, i.e., some classes can be poorly segmented even if overall segmentation is optimal. Hence, multi-scale segmentation has been offered as a solution to this problem (Hay et al. 2003; Kim et al. 2011; dos Santos et al. 2012).

4.2 Object and patch delineation

We found out that some habitat patches were not segmented properly using any of the methods, layers or parameter combinations. For instance stream-sided habitats or mires were often poorly delineated. Therefore, some extra analysis is needed, such as stream network mapping (Räsänen et al., in prep), other ancillary information or expert knowledge (Mustonen, Packalén, and Kangas 2008) or segmentation post-modification to delineate some of the patches correctly. It can even be asked, can the segmentation goodness over difficult patch delineation even be calculated. For instance, Radoux and Defourny (2007) delineated only those patches that could be seen from imagery. Therefore, it is not realistic to expect that segmentation delineates those objects that cannot be easily seen from the data that is segmented. LiDAR data, however, helped in finding some of the tricky features, such as mires. On the other hand, segmentations with LiDAR data and four WV-2 layers were not significantly different compared to segmentations with LiDAR data and eight WV-2 layers. As well, our study reasserted earlier studies that the problematic shadow effect of aerial or satellite imagery can be mitigated using LiDAR data (Geerling et al. 2007; Mustonen, Packalén, and Kangas 2008; Ke, Quackenbush, and Im 2010). Segmentations based on imagery only; nonetheless, produced classifications with almost as high classification accuracies as segmentations based on both imagery and LiDAR data. There can be at least two possible reasons for this small difference. First, segmentation based only on imagery may have other benefits compared to segmentation using both data types. Second, the proportion of shadow areas over all area can be rather small, especially with data resolution not higher than 2 m.

One major question in segmentation evaluation is whether it is better to delineate meaningful objects with meaningful thematic quality and maximum homogeneity (e.g. Mustonen, Packalén, and Kangas 2008) or to find segmentation that mimics field observations. Some authors (Wulder et al. 2008; Corcoran, Winstanley, and Mooney 2010) have questioned the rationality of supervised segmentation evaluation, especially in natural environments. It is true that nature is not easy to interpret. Different mappers classify habitat patches differently and also delineate patch boundaries differently. Yet, according to Cherrill and McClean (1995, 1999) the former type of error was more common in habitat mapping in the UK. Nevertheless, boundaries are not easy to draw and their location depends on the study scale (Lang et al. 2010). In our analysis, there were differences between boundary locations when our field data was compared to either FP or FFPS data. There were some differences between optimal segmentations based on different reference polygons. These differences were mostly minor, and approximately same kinds of segmentations were preferred irrespective of

596 the reference data. As well, correlations between COMBINED measures based on
597 different reference data were rather high (SA1 to FP 0.79 and SA2 to FFPS 0.89).
598 Furthermore, segmentation evaluation based on thematic quality is not unproblematic
599 either. Although the segmentation has good thematic quality, the segmentation
600 boundaries do not necessarily match with habitat type boundaries that exist in nature.
601 This can lead to difficulties in habitat classification, if it is performed, and eventually to
602 differences in planning decisions.
603 In a more general level, one can question if automated segmentation is worse
604 than manual delineation when it cannot find the boundaries that are manually
605 delineated. Delineations are different; that is true, but it is not straightforward to judge
606 either one of them better. For instance, automated delineation often produces more
607 complex objects. Complexity of the objects, however, can be both good and bad.
608 Although complexities hinder the usage in operational context, complexity can be
609 reduced using GIS techniques. Furthermore, complex boundaries can be even truer,
610 since natural boundaries are not always straight (Wulder et al. 2008). Therefore,
611 automated and manual delineations are two different interpretations and both of them
612 can be either good or bad depending on the segmentation method, mapper skills or the
613 operational context. In other words, question is not necessarily if one of them is correct
614 or incorrect but whether it is appropriate or inappropriate (Lang et al. 2010).

615 ***4.3 How to evaluate and measure segmentation goodness?***

616 According to our analysis, the FNEA was better segmentation method than the
617 watershed segmentation in IDRISI Taiga. Still, also IDRISI's segmentation method
618 provided good results. As already noted earlier, FNEA has been a good performer in
619 segmentation evaluations and is a standard method in OBIA studies. However, we
620 cannot give any percentage or any other quantitative evaluation which indicates how
621 much better FNEA is compared to IDRISI contrary to values given e.g. by N. Li, Huo,
622 and Fang (2010). N. Li, Huo, and Fang (2010) classified different types of objects to
623 correctly delineated, acceptably delineated and wrongly delineated. From these
624 classifications, they calculated performances of different segmentations and also the
625 percentage difference of performance. In our framework, such quantitative difference
626 evaluation would be more or less artificial, since in our study different measures of
627 segmentation goodness gave different results. This inconsistency has also been noted by
628 Clinton et al. (2010). Partly this inconsistency can be explained in terms of over- and
629 undersegmentation; i.e., deliberately avoiding one of them results often in getting the
630 other. However, also evaluation measures that should quantify the same phenomenon
631 can give different results. One explanation to this is that we tested several different
632 segmentations of which some were rather similar to each other. Furthermore, measures
633 are a bit dependent on the training data set used. One should, thus, be careful in the
634 selection of the reference data. On the other hand, some of the measures were robust,
635 i.e. produced similar results irrespective of the reference data. Additionally, general
636 picture was more or less similar with different reference polygons.
637 According to classification accuracies derived in our study, the best
638 segmentation was found using visual interpretation. Therefore, it could be argued that
639 supervised segmentation goodness evaluation measures evaluated by Clinton et al.
640 (2010) are not good. On the other hand, we knew what we wanted from visual
641 interpretation and fixed our objectives based on these needs. Automated supervised

segmentation goodness evaluation measures, on the contrary, were just selected based on what has been done before. Therefore, we knew better what we want from segmentation when we evaluated them visually: meaningful objects and boundaries. Yet, we would have included our own automated and supervised evaluation method in our analysis, if we had found a successful way to do automated evaluation. One reason that supervised segmentation goodness evaluation measures partly failed in our analysis could be that we used continuous reference polygon data and objects that were difficult to delineate. On the other hand, segmentation goodness measures did not work that well on water bodies either although water bodies are usually easy to delineate and are not bordered by each other. Based on measures, best segmentations for water bodies were usually segmentations using LiDAR data or segmentations as fine as possible. In our visual interpretation, it was, nonetheless, found out that water bodies cannot be delineated using LiDAR data alone. Only evaluation measure M ranked best a segmentation that could be good in water body delineation. Nevertheless, we cannot say that supervised evaluation measures are completely useless. On the contrary, one should know what evaluation measures favour and what she/he wants from segmentation before using evaluation measures. As well, visual interpretation is subjective, tedious and time-consuming (Zhang, Fritts, and Goldman 2008). It can be even practically impossible if several different segmentations over large areas should be evaluated.

Automated segmentation goodness evaluation could be done using landscape or shape metrics and thus unsupervised evaluation (Neubert and Meinel 2003, Meinel and Neubert 2004, Neubert, Herold, and Meinel 2008, H. Li et al. 2010, Ji et al. 2012). This is problematic, though, since for instance FNEA method uses shape metrics as parameters which the user can modify. Hence, using shape metrics also in evaluation could lead to circular reasoning. Another possible solution in finding good segmentation evaluation measures could be focusing on boundaries. In other words, it could be examined if boundaries drawn in the reference map are found in segmentation. For instance, Neubert and Herold (2008) measured what proportion of segment's perimeter is inside specific reference polygon's buffer zone. In similar vein, Lucieer and Stein (2002) have proposed boundary based measure and Clinton et al. (2010) included a modification of this measure in their analysis. Whereas Lucieer and Stein (2002) calculated shortest distances from reference polygons to any boundary pixel in segmentation, Clinton et al. (2010) averaged all distances to all segments inside reference polygon. Of these measures, original measure by Lucieer and Stein (2002) is more tempting, since boundaries inside reference polygon can disappear in classification but boundaries near reference polygon boundaries cannot be moved. However, Lucieer and Stein (2002) noted that finest segmentations ranked best using this evaluation. Taking this into account, they modified the original measure to take the length of boundary into account. These kinds of modifications, on the other hand, are difficult to design, because they easily favour either undersegmentation or oversegmentation. Furthermore, boundaries of natural objects are not exact. Hence, it is not always meaningful to find the "real" boundaries but boundaries that are visible in data.

Finally, unsupervised segmentation evaluation methods that often measure intersegment and intrasegment homogeneity or heterogeneity have been found useful in segmentation evaluation (Kim, Madden, and Warner 2009; Gao et al. 2011; Yue et al. 2012; Hou et al. 2013). While in our case the goal was to find segmentation that mimics

reference polygons, it could be interesting to test if unsupervised methods work well in this kind of task.

5 Conclusion

We tested different supervised segmentation goodness evaluation measures and visual interpretation to find a good segmentation for boreal forest habitat mapping. While different supervised segmentation goodness measures were fast to calculate from several segmentations, they provided inconsistent results. In other words, different segmentations stood out as being best when different measures were used. Visual interpretation, on the other hand, was tedious and segmentations could not be evaluated thoroughly in reasonable time. Although we selected only one segmentation as being the best based on visual interpretation, also other segmentations were visually good. In classification analysis, the visually selected segmentation gave the best classification accuracy but differences between different segmentations were rather small. Better segmentation may lead to better classification but there are several different definitions for good segmentation. Therefore, the relationship between segmentation and classification is not straightforward. We propose that the goodness of segmentation should be defined in each case separately and evaluation measures should be selected based on that definition. In our case, good segmentation was segmentation with (1) meaningful and not too complex segments, (2) boundaries parallel to reference polygon boundaries even for the smallest reference polygons but (3) as coarse as possible. There were, however, no evaluation measures to find these kinds of segmentations automatically. Overall, the best segmentations were FNEA segmentations with both imagery and LiDAR data. We conclude that different segmentation evaluation methods should be used with care especially in natural environment mappings. When segmentation evaluation is rigorously used; though, it can assist in finding a more optimal segmentation. Quantitative segmentation evaluation might provide better results in urban environments but more thorough testing is needed to support this claim.

6 Acknowledgements

This research was funded by Maj and Tor Nessling foundation. We are grateful to City of Jyväskylä for providing us the reference polygon dataset and the aerial imagery as well as to Finnish Forest and Park Service for the reference polygon dataset. MK received from Jenny and Antti Wihuri foundation a sabbatical scholarship for the year 2013 that was partly funded also by the EU IMPERIA-project (LIFE11 ENV/FI/905) and the University of Jyväskylä.

7 References

Ahti, Teuvo, Leena Hämet-Ahti, and Jaakko Jalas. 1968. “Vegetation Zones and Their Sections in Northwestern Europe.” *Annales Botanici Fennici* 5: 169–211.

Baatz, Martin and Arno Schäpe. 2000. “Multiresolution Segmentation – An Optimization Approach for High Quality Multi-Scale Image Segmentation.” In *Angewandte Geographische Informationsverarbeitung XII*, edited by J. Strobl and G. Griesebner, 12–23. Heidelberg: Wichmann.

Bar Massada, Avi, Rafi Kent, Lior Blank, Avi Perevolotsky, Liat Hadar, and Yohay Carmel. 2012. ”Automated Segmentation of Vegetation Units in a

731 Mediterranean Landscape." *International Journal of Remote Sensing* 33 (2):
732 346–364.
733 Benz, Ursula C., Peter Hofmann, Gregor Willhauck, Iris Lingenfelder, and Markus
734 Heynen. 2004. Multi-Resolution, Object-Oriented Fuzzy Analysis of Remote
735 Sensing Data for GIS-Ready Information. *ISPRS Journal of Photogrammetry*
736 *and Remote Sensing* 58: 239–258.
737 Blaschke, Thomas. 2010. "Object Based Image Analysis for Remote Sensing." *ISPRS*
738 *Journal of Photogrammetry and Remote Sensing* 65: 2–16.
739 Bock, Michael, Panteleimon Xofis, Jonathan Mitchley, Godela Rossner, and Michael
740 Wissen. 2005. "Object-Oriented Methods for Habitat Mapping at Multiple
741 Scales – Case Studies from Northern Germany and Wye Downs, UK." *Journal*
742 *of Nature Conservation* 13: 75–89.
743 Böhner, Jürgen, and Thomas Selige. 2006. "Spatial Prediction of Soil Attributes using
744 Terrain Analysis and Climate Regionalisation." In *SAGA – Analysis and*
745 *Modelling Applications*, edited by Jürgen Böhner, K. R. McCloy, and J. Strobl,
746 13–28. Göttinger Geographische Abhandlungen, Vol.115.
747 Breidenbach, Johannes, Erik Næsset, Vegard Lien, Terje Gobakken, and Svein Solberg.
748 2010. "Prediction of Species Specific Forest Inventory Attributes using a
749 Nonparametric Semi-Individual Tree Crown Approach Based on Fused
750 Airborne Laser Scanning and Multispectral Data." *Remote Sensing of*
751 *Environment* 114: 911–924.
752 Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.
753 Breiman, Leo, and Adele Cutler. 2007. "Random Forest: Classification Description."
754 http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
755 Câmara, Gilberto, Lubia Vinhas, Karine Reis Ferreira, Gilberto Ribeiro de Queiroz,
756 Ricardo Cartaxo Modesto De Souza, Antonio Miguel Vieira Monteiro, Marcelo
757 Tílio de Carvalho, Marco Antonio Casanova, and Ubirajara Moura De Freitas.
758 2008. "TerraLib: An Open Source GIS Library for Large-Scale Environmental
759 and Socio-Economic Applications." In *Open Source Approaches in Spatial Data*
760 *Handling*, edited by G. Brent Hall, and Michael G. Leahy, 247–270. Berlin:
761 Springer.
762 Carleer, A. P., O. Debeir, and E. Wolff. 2005. "Assessment of Very High Resolution
763 Satellite Image Segmentations." *Photogrammetric Engineering & Remote*
764 *Sensing* 71 (11): 1285–1294.
765 Castilla, Guillermo, Geoffrey J. Hay, and Jose R. Ruiz-Gallardo. 2008. "Size-
766 Constrained Region Merging (SCRM): An Automated Delineation Tool for
767 Assisted Photointerpretation." *Photogrammetric Engineering & Remote Sensing*
768 74 (4): 409–419.
769 Chabrier, S., C. Rosenberger, B. Emile, and H. Laurent, H. 2008. "Optimization-Based
770 Image Segmentation by Genetic Algorithms." *EURASIP Journal on Image and*
771 *Video Processing*, Article ID 842029. doi:10.1155/2008/842029.
772 Cherrill, Andrew, and Colin McClean. 1995. An Investigation of Uncertainty in Field
773 Habitat Mapping and the Implications for Detecting Land Cover Change.
774 *Landscape Ecology* 10 (1): 5–21.
775 Cherrill, Andrew, and Colin McClean. 1999. The Reliability of 'Phase 1' Habitat
776 Mapping in the UK: The Extent and Types of Observer Bias. *Landscape and*
777 *Urban Planning* 45: 131–143.

778 Clinton, Nicholas, Ashley Holt, James Scarborough, Li Yan, and Peng Gong. 2010.
779 "Accuracy Assessment Measures for Object-based Image Segmentation
780 Goodness." *Photogrammetric Engineering & Remote Sensing* 76 (3): 289–299.
781 Corcoran, Padraig, Adam Winstanley, and Peter Mooney. 2010. "Segmentation
782 Performance Evaluation for Object-Based Remotely Sensed Image Analysis".
783 *International Journal of Remote Sensing* 31 (3): 617–645.
784 Derivaux, S., G. Forestier, C. Wemmert, and S. Lefévre. 2010. "Supervised Image
785 Segmentation using Watershed Transform, Fuzzy Classification and
786 Evolutionary Computation." *Pattern Recognition Letters* 31 (15): 2364–2374.
787 Dingle Robertson, Laura, and Douglas J. King. 2011. "Comparison of Pixel- and
788 Object-Based Classification in Land Cover Change Mapping." *International*
789 *Journal of Remote Sensing* 32 (6): 1505–1529.
790 dos Santos, Jefferson Alex, Philippe-Henri Gosselin, Sylvie Phillipp-Foliguet, Ricardo
791 da S. Torres, and Alexandre Xavier Falcão. 2012. "Multiscale Classification of
792 Remote Sensing Images". *IEEE Transactions on Geoscience and Remote*
793 *Sensing* 50 (10): 3764–3775.
794 Drăguţ, Lucian, Dirk Tiede, and Shaun R. Levick. 2010. "ESP: A Tool to Estimate
795 Scale Parameter for Multiresolution Image Segmentation of Remotely Sensed
796 Data." *International Journal of Geographical Information Science* 24 (6): 859–
797 871.
798 Duro, Dennis C., Steven E. Franklin, and Monique G. Dubé. 2012. "A Comparison of
799 Pixel-Based and Object-Based Image Analysis with Selected Machine Learning
800 Algorithms for the Classification of Agricultural Landscapes using SPOT-5
801 HRG Imagery." *Remote Sensing of Environment* 118: 259–272.
802 Espindola, G. M., G. Camara, I. A. Reis, L. S. Bins, and A. M. Monteiro. 2006.
803 "Parameter Selection for Region-Growing Image Segmentation Algorithms
804 using Spatial Autocorrelation." *International Journal of Remote Sensing* 27
805 (14): 3035–3040.
806 Eysn, Lothar, Markus Hollaus, Klemens Schadauer, and Norbert Pfeifer. 2012. "Forest
807 Delineation Based on Airborne LIDAR Data." *Remote Sensing* 4: 762–783.
808 doi:10.3390/rs4030762.
809 Falkowski, Michael J., Michael A. Wulder, Joanne C. White, and Mark G. Gillis. 2009.
810 "Supporting Large-Area, Sample-Based Forest Inventories with Very High
811 Spatial Resolution Satellite Imagery." *Progress in Physical Geography* 33 (3):
812 403–423.
813 Feitosa, R. Q., G. A. O. P. Costa, T. B. Cazes, and B. Feijo. 2006. "A Genetic Approach
814 for the Automatic Adaptation of Segmentation Parameters." In *Bridging Remote*
815 *Sensing and GIS: 1st International Conference on Object-based Image Analysis*
816 *(OBIA 2006)*, edited by: Stefan Lang, Thomas Blaschke, and Elisabeth
817 Schöpfer. International Archives of the Photogrammetry, Remote Sensing and
818 Spatial Information Sciences – Volume XXXVI/4-C42.
819 Freeman, T. Graham. 1991. "Calculating catchment area with divergent flow based on a
820 regular grid." *Computers and Geosciences* 17: 413–422.Gao, Yan, Jean Francois
821 Mas, Norman Kerle, and Jose Antonio Navarrete Pacheco. 2011. "Optimal
822 Region Growing Segmentation and Its Effect on Classification Accuracy."
823 *International Journal of Remote Sensing* 32 (13): 3747–3763.
824 Geerling, G. W., M. Labrador-Garcia, J. P. G. W. Clevers, A. M. J. Ragas, and A. J. M.
825 Smits. 2007. "Classification of Floodplain Vegetation by Data-Fusion of

826 Spectral (CASI) and LiDAR Data." *International Journal of Remote Sensing* 28
827 (19): 4263–4284.
828 Geerling, G. W., M. J. Vreeken-Buijs, P. Jesse, A. M. J. Ragas, and A. J. M. Smits.
829 2009. "Mapping River Floodplain Ecotopes by Segmentation of Spectral (CASI)
830 and Structural (LiDAR) Remote Sensing Data." *River Research and*
831 *Applications* 25: 795–813.
832 Gonzales, Rafael C., and Richard E. Woods. 2002. *Digital Image Processing.* 2nd ed.
833 Upper Saddle River: Prentice Hall.
834 Grabs, T., Jan Seibert, K. Bishop, and H. Laudon. 2009. "Modeling Spatial Patterns of
835 Saturated Areas: A Comparison of the Topographic Wetness Index and a
836 Dynamic Distributed Model. *Journal of Hydrology* 373: 15–23.
837 Güntner, Andreas, Jan Seibert, and Stefan Uhlenbrook. 2004. "Modeling Spatial
838 Patterns of Saturated Areas: An Evaluation of Different Terrain Indices." *Water*
839 *Resources Research* 40, W05114. doi:10.1029/2003WR002864.
840 Hay, Geoffrey. J., Thomas Blaschke, Danielle J. Marceau, and André Bouchard. 2003.
841 "A Comparison of Three Image-Object Methods for the Multiscale Analysis of
842 Landscape Structure." *ISPRS Journal of Photogrammetry and Remote Sensing*
843 57: 327–345.
844 Hay, Geoffrey J., Guillermo Castilla, Michael A. Wulder, and Jose R. Ruiz 2005. "An
845 Automated Object-Based Approach for the Multiscale Image Segmentation of
846 Forest Scenes. *International Journal of Applied Earth Observation and*
847 *Geoinformation* 7: 339–359.
848 Hou, Zhengyang, Qing Xu, Tuula Nuutinen, and Timo Tokola. 2013. "Extraction of
849 remote sensing-based forest management units in tropical forests". *Remote*
850 *Sensing of Environment* 130: 1–10.
851 Isenburg, Martin. 2011. *LAStools: Efficient Tools for LiDAR Processing*. Version
852 110815. http://lastools.org.
853 Ji, Xiaole, Yanchen Bo, Jiehai Cheng, Yaqian He, and Xiaolong Liu. 2012. "The
854 Method of Assessing Shape Similarity of Object-Based Classification Result of
855 Remote Sensing Imagery". In *Proceedings of the Second International*
856 *Workshop on Earth Observation and Remote Sensing Applications (EORSA*
857 *2012)*, 157–160. IEEE.
858 http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6248407/.
859 Ke, Yinghai, Lindi J. Quackenbush, and Jungho Im. 2010. "Synergistic Use of
860 QuickBird Multispectral Imagery and LIDAR Data for Object-Based Forest
861 Species Classification." *Remote Sensing of Environment* 114: 1141–1154.
862 Kim, Minho, Marguerite Madden, and Timothy A. Warner. 2008. "Estimation of
863 Optimal Image Object Size for the Segmentation of Forest Stands with
864 Multispectral IKONOS Imagery." In *Object-Based Image Analysis: Spatial*
865 *Concepts for Knowledge-Driven Remote Sensing Applications*, edited by
866 Thomas Blaschke, Stefan Lang, and Geoffrey J. Hay, 291–307. Berlin: Springer.
867 Kim, Minho, Marguerite Madden, and Timothy A. Warner. 2009. "Forest Type
868 Mapping using Object-specific Texture Measures from Multispectral Ikonos
869 Imagery: Segmentation Quality and Image Classification Issues."
870 *Photogrammetric Engineering & Remote Sensing* 75 (7): 819–829.
871 Kim, Minho, Timothy A. Warner, Marguerite Madden, and Douglas S. Atkinson. 2011.
872 "Multi-Scale GEOBIA with Very High Spatial Resolution Digital Aerial

873 Imagery: Scale, Texture and Image Objects." *International Journal of Remote*
874 *Sensing* 32 (10): 2825–2850
875 Kopecký, Martin and Štěpánka Čížková. 2010. "Using Topographic Wetness Index in
876 Vegetation Ecology: Does the Algorithm Matter?" *Applied Vegetation Science*
877 13: 450–459.
878 Lang, Stefan, F. Albrecht, S. Kienberger, and Dirk Tiede. 2010. "Object Validity for
879 Operational Tasks in a Policy Context." *Journal of Spatial Science* 55 (1): 9–22.
880 Lawrence, Rick L., Shana D. Wood, and Roger L. Sheley. 2006. "Mapping Invasive
881 Plants using Hyperspectral Imagery and Breiman Cutler Classifications
882 (RandomForest)." *Remote Sensing of Environment* 100: 356–362.
883 Leckie, Donald G., François A. Gougeon, Nicholas Walsworth, and Dennis Paradine.
884 2003. "Stand Delineation and Composition Estimation using Semi-Automated
885 Individual Tree Crown Analysis". *Remote Sensing of Environment* 85: 355–369.
886 Li, Haitao, Haiyan Gu, Yanshun Han, and Jinghui Yang. 2010. "Object-Oriented
887 Classification of High-Resolution Remote Sensing Imagery based on an
888 Improved Colour Structure Code and a Support Vector Machine." *International*
889 *Journal of Remote Sensing* 31 (6): 1453–1470.
890 Li, Nan, Hong Huo, and Tao Fang. 2010. "A Novel Texture-Preceded Segmentation
891 Algorithm for High-Resolution Imagery." *IEEE Transactions on Geoscience*
892 *and Remote Sensing* 48 (7): 2818–2828.
893 Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by
894 randomForest." *R News* 2/3: 18–22.
895 Lucieer, Arko, and Alfred Stein. 2002. "Existential Uncertainty of Spatial Objects
896 Segmented From Satellite Sensor Imagery." *IEEE Transactions on Geoscience*
897 *and Remote Sensing* 40 (11): 2518–2521.
898 Marpu, P. R., M. Neubert, H. Herold, and I. Niemeyer. 2010. "Enhanced Evaluation of
899 Image Segmentation Results". *Journal of Spatial Science* 55 (1): 55–68.
900 Meinel, G., and M. Neubert. 2004. "A Comparison of Segmentation Programs for High
901 Resolution Remote Sensing Data." *In XX ISPRS Congress: Technical*
902 *Commission IV*, edited by Orhan Altan, 1097–1105. International Archives of
903 the Photogrammetry, Remote Sensing and Spatial Information Sciences –
904 Volume XXXV Part B4.
905 Möller, M., L. Lymburner, and M. Volk. 2007. "The Comparison Index: A Tool for
906 Assessing the Accuracy of Image Segmentation." *International Journal of*
907 *Applied Earth Observation and Geoinformation* 9: 311–321.
908 Mustonen, Jukka, Petteri Packalén, and Annika Kangas. 2008. "Automatic
909 Segmentation of Forest Stands Using a Canopy Height Model and Aerial
910 Photography." *Scandinavian Journal of Forest Research* 23 (6): 534–545.
911 Murphy, P. N. C., J. Ogilvie, and P. Arp. 2009. "Topographic Modeling of Soil
912 Moisture Conditions: a Comparison and Verification of Two Models."
913 *European Journal of Soil Science* 60: 94–109.
914 Neubert, M., and H. Herold. 2008. "Assessment of Remote Sensing Image
915 Segmentation Quality." In *GEOBIA 2008 - Pixels, Objects, Intelligence:*
916 *GEOgraphic Object Based Image Analysis for the 21st Century Proceedings*,
917 edited by Geoffrey J. Hay, Thomas Blaschke and Danielle Marceau.
918 International Archives of Photogrammetry, Remote Sensing and Spatial
919 Information Sciences XXXVIII-4/C1.

Neubert, M., and G. Meinel. 2003. “Evaluation of Segmentation Programs for High Resolution Remote Sensing Applications.” In *Proceedings Joint ISPRS/EARSeL Workshop “High Resolution Mapping from Space 2003”*, edited by M. Schroeder, K. Jacobsen, and C. Heipke. http://www.ipi.uni-hannover.de/fileadmin/institut/pdf/neubert.pdf.

Neubert, M., H. Herold, and G. Meinel. 2008. “Assessing Image Segmentation Quality – Concepts, Methods and Applications.” In *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*, edited by Thomas Blaschke, Stefan Lang, and Geoffrey J. Hay, 769–784. Berlin: Springer.

Pekkarinen, Anssi. 2002. “Image Segment-Based Spectral Features in the Estimation of Timber Volume.” *Remote sensing of Environment* 82: 349–359.

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Radoux, J. and P. Defourny. 2007. “A Quantitative Assessment of Boundaries in Automated Forest Stand Delineation using Very High Resolution Imagery.” *Remote Sensing of Environment* 110: 468–475.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. 2012. “An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification.” *ISPRS Journal of Photogrammetry and Remote Sensing* 67: 93–104.

Smith, A. 2010. “Image Segmentation Scale Parameter Optimization and Land Cover Classification using the Random Forest Algorithm.” *Journal of Spatial Science* 55(1): 69–79.

Sørensen, R., U. Zinko, and Jan Seibert. 2006. “On the Calculation of the Topographic Wetness Index: an Evaluation of Different Methods based on Field Observations.” *Hydrology and Earth System Sciences* 10: 101–112.

Tian, J. and D.-M. Chen. 2007. “Optimization in Multi-Scale Segmentation of High-Resolution Satellite Images for Artificial Feature Recognition.” *International Journal of Remote Sensing* 28 (20): 4625–4644.

Verbeeck, Klaartje, Martin Hermy, and Jos van Orshoven. 2012. “External Geo-Information in the Segmentation of VHR imagery Improves the Detection of Imperviousness in Urban Neighborhoods.” *International Journal of Applied Earth Observation and Geoinformation* 18: 428–435.

Vesterbacka, Raisa. 2010. Luontopalvelujen luontotyyppi-inventoinnin maastotyöohje. Vantaa: Metsähallitus.

Wang, L., W. Sousa, and Peng Gong. 2004. “Integration of Object-Based and Pixel-Based Classification for Mapping Mangroves with IKONOS Imagery.” *International Journal of Remote Sensing* 25 (24): 5655–5668.

Wang, Zhongwu, John R. Jensen, and Jungho Im. 2010. “An Automatic Region-Based Image Segmentation Algorithm for Remote Sensing Applications.” *Environmental Modelling and Software* 25: 1149–1165.

Weidner, Uwe 2008. “Contribution to the Assessment of Segmentation Quality for Remote Sensing Applications.” In *XXIst ISPRS Congress: Technical Commission VII*, edited by Chen Jun, Jiang Jie, and John van Genderen, 479–484. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences – Volume XXXVII Part B7.

967 Whiteside, Timothy G., Guy S. Boggs, and Stefan W. Maier. 2011. “Comparing Object-
968 Based and Pixel-Based Classifications for Mapping Savannas.” *International*
969 *Journal of Applied Earth Observation and Geoinformation* 13: 884–893.
970 Wulder, Michael A., Joanne C. White, Geoffrey J. Hay, and Guillermo Castilla. 2008.
971 “Towards Automated Segmentation of Forest Inventory Polygons on High
972 Spatial Resolution Satellite Imagery.” *The Forestry Chronicle* 84 (2): 221–230.
973 Yang, Luren, Fritz Albregtsen, Tor Lønnestad, and Per Grøttum. 1995. “A Supervised
974 Approach to the Evaluation of Image Segmentation Methods.” *Computer*
975 *Analysis of Images and Patterns: 6thInternational Conference, CAIP ’95*
976 *Prague Czech Republic, September 6-8, 1995 Proceedings*, edited by Václac
977 Hlaváč, and Radim Šára, 759–765. Lecture Notes on Computer Science,
978 Volume 970. Heildelberg: Springer.
979 Yue, Anzhi, Jianyu Yang, Chao Zhang, Wei Su, Wenju Yun, Dehai Zhu, Shunxi Liu,
980 and Zhongwu Wang. 2012. “The Optimal Segmentation Scale Identification
981 Using Multispectral WorldView-2 Images”. *Sensor Letters* 10 (1–2): 285–
982 291(7).
983 Zhan, Qingming, Martien Molenaar, Klaus Tempfli, and Wengzhong Shi. 2005.
984 “Quality Assessment for Geo-Spatial Objects Derived from Remotely Sensed
985 Data.” *International Journal of Remote Sensing* 26 (14): 2593–2974.
986 Zhang, Hui, Jason E. Fritts, and Sally A. Goldman. 2008. “Image Segmentation
987 Rvaluation: A Survey of Unsupervised Methods.” *Computer Vision and Image*
988 *Understanding* 110: 260–280.
989 Zhang, Y. J. 1996. “A Survey on Evaluation Methods for Image Segmentation.” *Pattern*
990 *Recognition* 29 (8): 1335–1346.
991 Zhang, Y. J. 1997. “Evaluation and Comparison of Different Segmentation
992 Algorithms.” *Pattern Recognition Letters* 18 (10): 963–974.

993 Table 1. Different habitat types that were mapped during field work when the reference
994 polygons were drawn and that were used in the classification part of the research.

Habitat type	Number of age groups/management possibilities
xeric (pine dominated) forests	4: clear-cut, sapling stand, young, mature
mesic (spruce dominated) forests	5: clear-cut, sapling stand, young, mature, natural
herb-rich (mixed/deciduous) forests	4: sapling stand, young, mature, natural
bare rock	1
pine mires	1
spruce mires	2: not drained, drained
open mires	1
water (lakes and streams)	1
small creeks	1
springs	1
grasslands	1
fields	1
roads	1
yards	1
sand pits	1

995

996 Table 2. Simple segmentation goodness evaluation measures that were used in
997 segmentation evaluation. Column MEASURES refers to what the method should
998 measure. Column SOURCE refers to the article where the measure was first used.
999 Column WEIGHTED refers to if the measure is weighted by a reference object.

METHOD	MEASURES	SOURCE	WEIGHTED	NOTE
UnderMerging	undersegmentation	Yang et al. (1995)		
OverMerging	oversegmentation	Yang et al. (1995)		
AFI	area match	Lucieer & Stein (2002)		
CountOver	oversegmentation	Lucieer & Stein (2002)		based on AFI
CountUnder	undersegmentation	Lucieer & Stein (2002)		based on AFI
SimSize mean	area match	Zhan et al. (2005)	X	
SimSize sd	area match	Zhan et al. (2005)	X	
RAsuper	undersegmentation	Möller, Lymburner, and Volk (2007)	X	
RAsub	oversegmentation	Möller, Lymburner, and Volk (2007)	X	
QR	area match	Weidner (2008)	X	
OverSegmentation	oversegmentation	Clinton et al. (2010)	X	
UnderSegmentation	undersegmentation	Clinton et al. (2010)	X	
RPsuper	distance to centroid	Möller et al. (2007)	X	
RPsub	distance to centroid	Möller et al. (2007)	X	
QLoc mean	distance to centroid	Zhan et al. (2005)	X	=RPsub
QLoc sd	distance to centroid	Zhan et al. (2005)	X	

1000
1001

1002 Table 3. Combined segmentation goodness evaluation measures that were used in
1003 segmentation evaluation. Column INCLUDES refers to the simple measures that are
1004 included in the respective combined measure. Column CALCULATION refers to how
1005 the combined measure was calculated.

METHOD	INCLUDES	CALCULATION
M	RAsuper, RAsub, RPsuper, RPsub	RMS
ZH1	SimSize mean, SimSize std, QLoc mean, QLoc std	RMS
ZH2	SimSize mean, QLoc mean	RMS
D	OverSegmentation, UnderSegmentation	RMS
OverUnder	CountOver, CountUnder	SUM
MergeSum	OverMerging, UnderMerging	SUM
COMBINED	all other simple measures than QLoc mean	RMS, normalized

Table 4. Best segmentations according to different measures and reference polygons. Segmentation methods are marked as follows. Text refers to method, letter after the text to layer combination a-d (see text), s to scale (FNEA) or similarity parameter (IDRISI), c to colour parameter, m to weight given to mean and v to weight given to variance.

MEASURE	All area	SA1	SA2	SA3	water	mires	FP	FFPS
UnderMerging	FNEA_d_s5_c.75	FNEA_d_s5_c.75	FNEA_a_s5_c.75	FNEA_a_s5_c.75	FNEA_a_s5_c.5	FNEA_d_s5_c.75	FNEA_c_s5_c.75	FNEA_a_s5_c.5
OverMerging	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_b_s50_c.5	FNEA_b_s50_c.5	IDRISI_b_s70_m1_v9	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_c_s50_c.75
AFI	IDRISI_a_s25_m9_v1	IDRISI_a_s25_m1_v9	IDRISI_d_s20_m9_v1	IDRISI_c_s35_m1_v9	IDRISI_d_s70_m5_v5	FNEA_b_s5_c.25	IDRISI_a_s45_m5_v5	IDRISI_b_s35_m9_v1
CountOver	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_c_s50_c.25*	FNEA_b_s50_c.5	FNEA_d_s45_c.25*	FNEA_b_s50_c.5*	FNEA_b_s50_c.75*
CountUnder	IDRISI_a_s20_m1_v9*	IDRISI_a_s25_m1_v9*	IDRISI_a_s20_m9_v1*	IDRISI_b_s50_m9_v1	FNEA_b_s25_c.25*	IDRISI_a_s40_m1_v9*	IDRISI_a_s70_m1_v9*	IDRISI_a_s70_m1_v9*
SimSize mean	FNEA_b_s30_c.75	FNEA_b_s30_c.75	FNEA_d_s30_c.75	IDRISI_c_s65_m9_v1	FNEA_b_s45_c.75	IDRISI_c_s55_m9_v1	FNEA_c_s30_c.25	FNEA_d_s40_c.25
SimSize sd	FNEA_a_s5_c.75	FNEA_b_s5_c.75	FNEA_a_s5_c.75	FNEA_a_s5_c.75	FNEA_a_s5_c.5	FNEA_b_s50_c.25	FNEA_a_s5_c.75	FNEA_d_s5_c.75
RAsuper	FNEA_b_s5_c.75	FNEA_b_s5_c.75	FNEA_b_s5_c.75	FNEA_a_s5_c.75	FNEA_c_s5_c.5	FNEA_d_s5_c.75	FNEA_b_s5_c.75	FNEA_b_s5_c.75
RAsub	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_b_s50_c.5	FNEA_b_s50_c.25	FNEA_b_s50_c.5	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_b_s50_c.25
QR	FNEA_b_s35_c.75	FNEA_b_s30_c.75	FNEA_c_s30_c.75	FNEA_a_s45_c.75	IDRISI_b_s65_m9_v1	IDRISI_a_s45_m1_v9	FNEA_c_s45_c.75	FNEA_d_s40_c.75
OverSegmentation	FNEA_b_s50_c.5	FNEA_b_s50_c.25	FNEA_b_s50_c.5	FNEA_b_s50_c.5	IDRISI_b_s70_m1_v9	FNEA_b_s50_c.25	FNEA_b_s50_c.25	FNEA_c_s50_c.75
UnderSegmentation	FNEA_b_s5_c.75	FNEA_b_s5_c.75	FNEA_b_s5_c.75	FNEA_b_s5_c.75	IDRISI_a_s20_m5_v5	FNEA_b_s5_c.75	FNEA_b_s5_c.75	FNEA_b_s5_c.75
RPsuper	IDRISI_d_s55_m9_v1	FNEA_b_s20_c.5	FNEA_d_s20_c.5	FNEA_b_s15_c.5	IDRISI_b_s65_m9_v1	FNEA_d_s10_c.25	FNEA_d_s30_c.75	IDRISI_d_s60_m9_v1
RPsub	FNEA_b_s35_c.75	IDRISI_b_s55_m5_v5	FNEA_a_s40_c.75	FNEA_c_s40_c.75	IDRISI_b_s65_m9_v1	FNEA_b_s15_c.25	FNEA_c_s30_c.75	FNEA_d_s40_c.75
QLoc mean	FNEA_b_s35_c.75	IDRISI_b_s55_m5_v5	FNEA_a_s40_c.75	FNEA_c_s40_c.75	IDRISI_b_s65_m9_v1	FNEA_b_s15_c.25	FNEA_c_s30_c.75	FNEA_d_s40_c.75
QLoc sd	FNEA_a_s45_c.25	IDRISI_b_s65_m9_v1	FNEA_c_s45_c.75	FNEA_c_s40_c.75	IDRISI_b_s65_m1_v9	FNEA_b_s20_c.25	FNEA_c_s30_c.75	FNEA_c_s50_c.25
M	FNEA_c_s25_c.75	IDRISI_b_s45_m9_v1	FNEA_a_s5_c.5	FNEA_b_s20_c.75	FNEA_c_s45_c.75	FNEA_d_s5_c.75	FNEA_b_s5_c.75	FNEA_a_s25_c.5
ZH1	FNEA_d_s50_c.25	FNEA_d_s50_c.25	FNEA_c_s5_c.25	FNEA_b_s45_c.5	IDRISI_b_s40_m5_v5	FNEA_b_s50_c.25	FNEA_a_s5_c.25	FNEA_a_s45_c.25
ZH2	FNEA_b_s35_c.75	IDRISI_b_s60_m9_v1	FNEA_d_s30_c.75	FNEA_c_s40_c.75	FNEA_b_s45_c.75	IDRISI_c_s55_m9_v1	FNEA_c_s30_c.25	FNEA_d_s45_c.25
D	FNEA_d_s35_c.5	FNEA_d_s35_c.5	FNEA_d_s35_c.5	FNEA_b_s35_c.75	IDRISI_b_s70_m1_v9	FNEA_b_s20_c.75	FNEA_c_s35_c.75	FNEA_d_s40_c.5
OverUnder	IDRISI_b_s70_m9_v1	IDRISI_b_s65_m1_v9	IDRISI_a_s70_m5_v5	IDRISI_b_s70_m9_v1	FNEA_b_s50_c.5	IDRISI_a_s70_m9_v1	FNEA_c_s50_c.75	IDRISI_a_s70_m1_v9
MergeSum	FNEA_b_s15_c.5	FNEA_b_s15_c.5	FNEA_a_s15_c.75	FNEA_b_s20_c.75	IDRISI_b_s65_m9_v1	FNEA_b_s10_c.25	FNEA_c_s25_c.5	FNEA_d_s25_c.25
COMBINED	FNEA_b_s25_c.5	IDRISI_b_s50_m5_v5	FNEA_b_s45_c.75	FNEA_b_s30_c.75	IDRISI_b_s70_m1_v9	FNEA_a_s40_c.5	FNEA_b_s45_c.5	FNEA_a_s35_c.25

*=tie with other segmentations which are not indicated here. Segmentation that is shown ranked best using OverUnder evaluation method

1012

1013 Table 5. 20 best and 10 worst segmentations based on all area reference polygons and
1014 COMBINED measure. Segmentation methods are named as in Table 4.

	SEGMENTATION	COMBINED
1	FNEA_b_s25_c.5	0.646
2	FNEA_b_s25_c.75	0.649
3	FNEA_b_s30_c.75	0.650
4	FNEA_c_s25_c.75	0.650
5	FNEA_d_s25_c.75	0.651
6	IDRISI_b_s55_m9_v1	0.652
7	FNEA_a_s30_c.75	0.652
8	IDRISI_b_s65_m9_v1	0.652
9	IDRISI_d_s65_m9_v1	0.652
10	FNEA_d_s30_c.75	0.652
11	IDRISI_b_s50_m5_v5	0.653
12	FNEA_b_s20_c.5	0.653
13	IDRISI_b_s55_m5_v5	0.653
14	IDRISI_b_s60_m9_v1	0.653
15	IDRISI_b_s50_m9_v1	0.653
16	IDRISI_c_s60_m5_v5	0.653
17	IDRISI_a_s60_m1_v9	0.653
18	IDRISI_a_s70_m9_v1	0.653
19	IDRISI_d_s65_m5_v5	0.654
20	IDRISI_a_s60_m5_v5	0.654
…		
243	IDRISI_c_s20_m9_v1	0.770
244	FNEA_a_s5_c.25	0.771
245	FNEA_c_s5_c.5	0.782
246	FNEA_d_s5_c.5	0.789
247	FNEA_a_s5_c.5	0.796
248	FNEA_b_s5_c.5	0.797
249	FNEA_c_s5_c.75	0.801
250	FNEA_d_s5_c.75	0.805
251	FNEA_a_s5_c.75	0.811
252	FNEA_b_s5_c.75	0.813

1015
1016

Table 6. Best segmentations based on different measures sorted by segmentation method, layer combinations, and different parameter options. Numbers refer to the number of segmentations that were regarded as best.

all segmentations			FNEA segmentations			IDRISI segmentations		
	value	count		value	count		value	count
method	FNEA	140	scale	5	36	similarity	20	4
	IDRISI	44		10	2		25	3
				15	6		30	0
layers	a	35		20	6		35	2
	b	93		25	6		40	2
	c	28		30	13		45	3
	d	28		35	10		50	2
				40	12		55	5
				45	13		60	2
				50	36		65	9
							70	12
			colour	0.25	40			
				0.5	30	mean/var	0.1/0.9	15
				0.75	70		0.5/0.5	8
							0.9/0.1	21

Table 7. Correlations between the COMBINED goodness measure and individual goodness measures based on different reference polygons. SA refers to sub-area, FP to forestry planning data and FFPS for FFPS data.

MEASURE	ALL	SA1	SA2	SA3	water	mires	FP	FFPS
UnderMerging	-0.24	-0.39	-0.63	-0.13	-0.63	-0.61	-0.69	-0.33
OverMerging	0.89	0.90	0.95	0.90	0.88	0.97	0.91	0.90
AFI	-0.22	-0.36	-0.63	-0.10	-0.40	-0.61	-0.54	-0.24
CountOver	0.69	0.78	0.91	0.73	0.87	0.91	0.88	0.71
CountUnder	-0.25	-0.31	-0.73	-0.31	-0.53	-0.62	-0.53	-0.10
SimSize mean	0.89	0.91	0.91	0.89	0.85	0.29	0.92	0.82
SimSize sd	-0.91	-0.93	-0.93	-0.91	-0.94	-0.34	-0.93	-0.85
RAsuper	-0.79	-0.86	-0.94	-0.83	-0.95	-0.93	-0.96	-0.86
RAsub	0.43	0.54	0.80	0.53	0.86	0.75	0.85	0.52
QR	0.91	0.94	0.93	0.92	0.95	0.35	0.95	0.90
OverSegmentation	0.48	0.60	0.79	0.56	0.81	0.84	0.84	0.50
UnderSegmentation	-0.52	-0.65	-0.82	-0.58	-0.79	-0.86	-0.87	-0.63
RPsuper	0.85	0.67	0.52	0.78	0.88	-0.55	0.74	0.96
RPsub	0.86	0.94	0.88	0.94	0.95	0.01	0.92	0.87
QLoc mean	0.86	0.94	0.88	0.94	0.95	0.01	0.92	0.87
QLoc sd	0.63	0.68	0.80	0.79	0.80	0.30	0.80	0.72
M	0.98	0.98	-0.88	0.97	-0.50	-0.72	-0.90	0.97
ZH1	0.66	0.60	-0.83	0.84	0.76	-0.29	-0.87	0.83
ZH2	0.89	0.96	0.92	0.94	0.94	0.29	0.92	0.89
D	0.77	0.82	0.88	0.81	0.83	0.36	0.90	0.65
OverUnder	0.76	0.84	0.90	0.78	0.91	0.86	0.89	0.75
MergeSum	0.41	0.50	-0.34	0.69	0.86	-0.52	0.85	0.90

1025 Table 8. Correlations between the COMBINED measure values based on different
1026 reference polygons.

	ALL	SA1	SA2	SA3	water	mires	FP	FFPS
ALL	1.00	0.89	0.85	0.94	0.64	0.84	0.78	0.94
SA1	0.89	1.00	0.83	0.84	0.59	0.83	0.79	0.88
SA2	0.85	0.83	1.00	0.87	0.82	0.97	0.97	0.89
SA3	0.94	0.84	0.87	1.00	0.73	0.87	0.83	0.88
water	0.64	0.59	0.82	0.73	1.00	0.80	0.80	0.63
mires	0.84	0.83	0.97	0.87	0.80	1.00	0.95	0.85
FP	0.78	0.79	0.97	0.83	0.80	0.95	1.00	0.84
FFPS	0.94	0.88	0.89	0.88	0.63	0.85	0.84	1.00

1027 Table 9. Classification accuracies derived from classifications based on different
1028 segmentations. Segmentations are marked as in Table 2. Also, criteria why each
1029 segmentation was chosen to the classification analyses are given.

SEGMENTATION	ACC	Why segmentation was selected to classification
FNEA_b_s5_c.75	0.69	BEST in avoiding undersegmentation, WORST based on COMBINATION and ALL AREA
FNEA_c_s5_c.75	0.69	Comparison against FNEA_b_s5_c.75
FNEA_a_s10_c.5	0.71	WV-2 layers only, comparison against FNEA_c_s10_c.5
FNEA_c_s10_c.5	0.72	BEST segmentation based on visual interpretation
FNEA_d_s15_c.25	0.71	GOOD in visual interpretation
FNEA_a_s20_c.75	0.69	Segmentation based on WV-2 data only, OK visually
FNEA_b_s25_c.5	0.66	BEST segmentation based on COMBINATION and ALL AREA, OK visually
FNEA_d_s35_c.5	0.66	BEST based on D and ALL AREA
FNEA_c_s50_c.75	0.65	GOOD in avoiding oversegmentation, GOOD in visual boundary evaluation
IDRISI_d_s30_m9v1	0.70	OK in visual interpretation, small segments
IDRISI_c_s40_m5v5	0.69	OK in visual interpretation, quite small segments
IDRISI_b_s70_m9v1	0.60	BEST based on OverUnder and ALL AREA, BAD in visual interpretation

1030

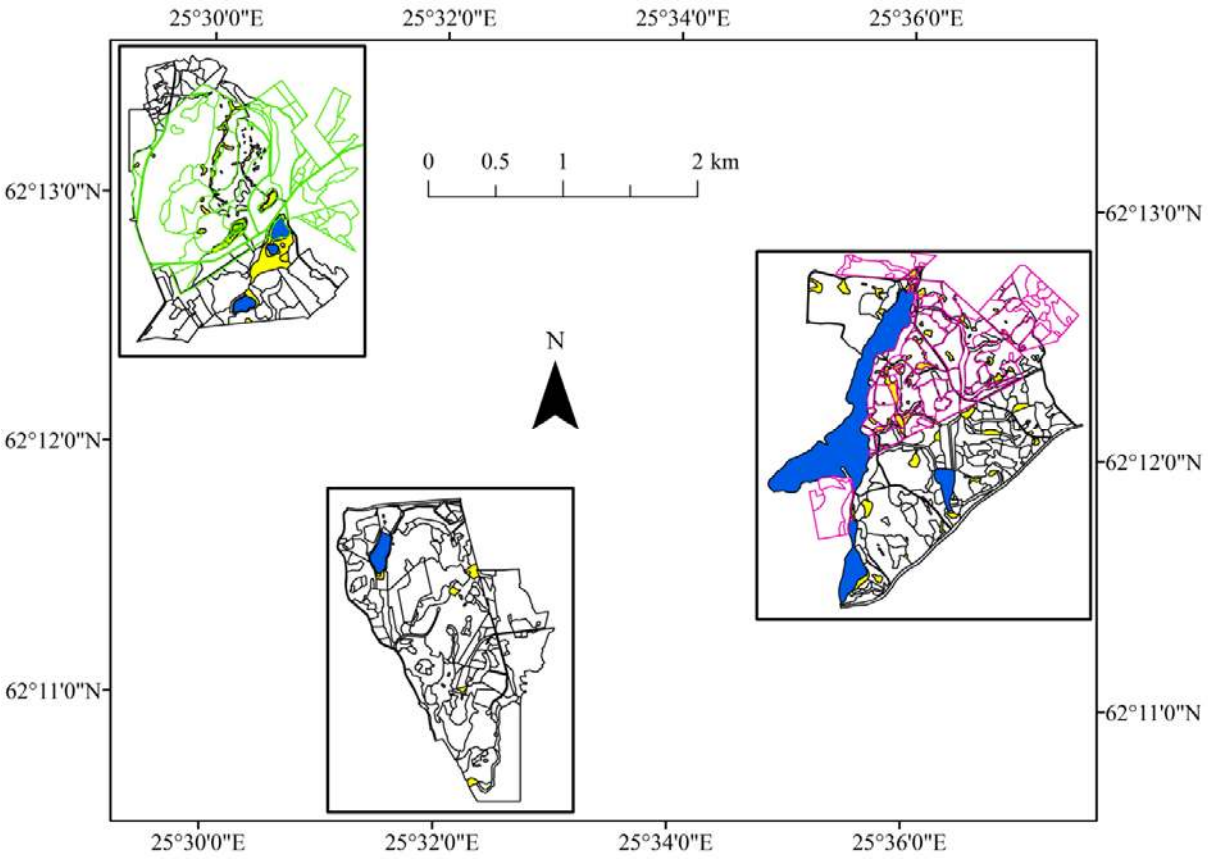

Figure 1. Different reference polygons used. Reference polygons drawn in our field work are marked with black borders. Sub-area 1 is located in the eastern, sub-area 2 in the north-western and sub-area 3 in the southern part of the whole area. Water bodies are drawn with blue colour and mires with yellow colour. Forestry planning polygons are marked with magenta outlines and FFPS polygons with green outlines. FNEA segmentations were performed inside the black rectangles whereas IDRISI segmentations were performed also in the areas between the black rectangles.

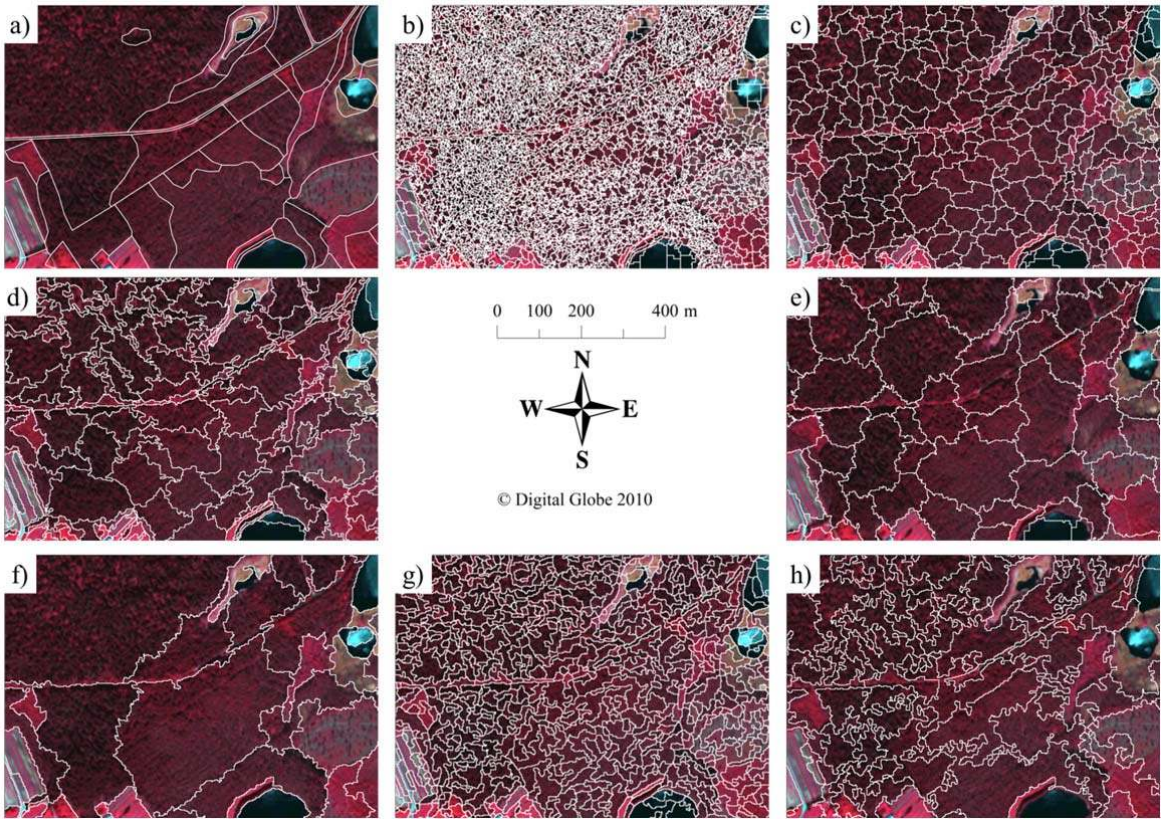

Figure 2. Reference polygons and a visually chosen set of different segmentations drawn on a WV-2 false colour image (red: band 7/NIR1, green: band 5/red, blue: band 3/green) in background. a) reference polygons, b) FNEA_b_s5_c.75, c) FNEA_c_s10_c.5, d) FNEA_a_s20_c.75, e) FNEA_b_s25_c.5, f) FNEA_c_s50_c.75, g) IDRISI_d_s30_m9v1, and h) IDRISI_b_s70_m9v1. Images are from the southern part of the sub-area 2.