Remote Sensing of Environment 142(2014)pp 9-19 DOI 10.1016/j.rse.2013.10.030

Accounting for the area of polygon sampling units for the prediction of primary accuracy assessment indices

Julien Radoux [a] and Patrick Bogaert

Earth and Life Institute
Université catholique de Louvain
Croix du Sud, L7.05.16, B-1348 Belgium
[a] *Corresponding author : julien.radoux@uclouvain.be, Fax : +32(0)10 47 88 98*

Abstract

GEographic Object-Based Image Analysis (GEOBIA) has become a popular alternative for land cover and land use classification. In this case, polygons can be selected as sampling units to match the conceptual model of the map. However, little attention has been paid to the use of polygons for the validation of those maps. In this paper, we quantitatively assess the prediction of the primary thematic accuracy indices when the sampling unit is a polygon. The variable size of the sample polygons is a major concern for the prediction of the accuracy indices. Indeed, the classification accuracy, in addition to being class-dependent, depends on the polygon area. A practical solution supported by a theoretical framework that is conditional to the sample dataset is proposed in this study. This new predictor takes advantage of the known classification results for an improved efficiency. Empirical results based on synthetic maps show that the new predictor outperforms alternative methods for overall accuracy. The RMSE of the area weighted predictor was achieved with 50% less sample polygons thanks to our new predictor.

1. Introduction

Land cover/land use maps are of paramount importance in various applications such as land monitoring, land use planning, hydrological modelling or natural resource management. Consequently, map users need reliable quality information about those products for using them in an appropriate way. Previous works on accuracy assessment have designed standard quality indices and methods which are now widely accepted by the remote sensing community. The core of the accuracy assessment typically relies on a confusion matrix based on a validation sample, which matches the mapped land cover to some reference information (Congalton, 1991). The confusion matrix is often accompanied by global indices such as the overall, the user's and the producer's accuracy indices (Foody, 2002; Congalton, 1991; Stehman, 1997), which provide a useful summary of the map's quality. According to Liu et al. (2007), these are the three primary thematic accuracy indices.

Standard accuracy assessment methods rely also on the definition of a sampling unit used in the response design. Congalton & Green (2009) identified 3 types of sampling units : points, pixel clusters and polygons. A universally best spatial unit does not exist, so it is critical to recognize how the choice of a sampling unit affects the accuracy assessment process (Stehman & Wickham, 2011). This choice also depends on the conceptual model of the map, i.e. spatial object, field or spatial regions according to definitions of Bian (2007):

- A spatial object is used as a conceptual model for spatially discrete information. The spatial extent of these objects is limited in space and their boundaries are defined by a set of rules. At least one categorical variable, the object type, is associated with those objects after a chosen typology. In a response design, spatial objects are most of the time unambiguously validated, either by photo-interpretation

or from the field, because they can be embraced by the validation crew. Furthermore, their integrity is often assessed as a whole, using reference polygons and resulting in class-specific metrics where the geometric component plays a major role (Persello & Bruzzone, 2010). In this case, Zhan et al. (2005) therefore concluded that polygon-based sampling units provide additional information compared with point-based approaches.

- A field consists in a spatially continuous quantitative variable that can be measured in any point of space. A typical example is the elevation above a reference surface, which is an important variable in, e.g., hydrological modelling. Although it is possible, and in some cases recommended (Lambin, 1999), to describe the land cover using continuous fields, classification is more popular (Huang et al., 2002), especially for large scale mapping. In any case, the validation of spatial fields primarily relies on point-based sampling (Hansen et al., 2002) because polygons would introduce abrupt changes in the fields values.

- A spatial region represents a mass of individuals that can be conceptualised both as a continuous field and as discrete spatial objects, which is often the case of land cover. This duality is also found at the level of the logical model: they can be discretised as vector polygons with consensual (fuzzy) boundaries or represented as a grid with the proportion of each individual and no defined boundaries. Spatial regions are delimited with an arbitrary boundary that is difficult to define with a set of rules (e.g. ecotones), so that their position is often uncertain and the sources of geometric errors are diverse (Radoux & Defourny, 2007). Concerning the labels, the use of unambiguous classification systems, such as the UN Land Cover Classification System (Di Gregorio & Jansen, 2000), is recommended in order to avoid overlapping class definitions.

When spatial objects or spatial regions are identified on a map as polygons, Congalton & Green (2009) recommend the use of sample polygons to assess the thematic accuracy. GEographic Object Based Image Analysis (GEOBIA) is a typical case where the resulting map is partitioned in a set of polygons. GEOBIA is increasingly used to process remote sensing data (Blaschke, 2010) and has been successfully applied in image classification and change detection (Radoux & Defourny, 2010; Bontemps et al., 2008). A rationale of this approach is that the interpretation of a group of spatially adjacent pixels with similar properties is closer to human interpretation of spatial regions than independent pixel interpretation. Intrinsically, the polygons used in GEOBIA are thus considered as homogeneous in terms of land cover (Hay & Castilla, 2008). Those polygons are built based on an image (so called image-segments or image-objects) or obtained from an ancillary data source.

Various methods were developed to evaluate image segmentation goodness based on supervised and unsupervised indices (Clinton et al., 2010; Neubert et al., 2008; Zhang et al., 2008). These indices are most of the time related to the four criteria proposed by Haralick & Shapiro (1985): i) regions should be homogeneous with respect to some characteristics, ii) adjacent regions should exhibit marked differences with respect to these characteristics, iii) region interiors should be free of holes, and iv) boundaries should be spatially accurate and precise. In the frame of GEOBIA, the two first criteria are directly related to over- and under-segmentation concerns for an image, respectively when an image-segment is only a part of a spatial region or a spatial object, and when more than one spatial object or region are included in the same image-segment (Carleer et al., 2005). After classification, over-segmentation and holes are potentially removed while under-segmentation may lead to artificial class associations that often reduce the semantic map quality.

The segmentation goodness indices, primarily based on the topological and geometric matching between corresponding spatial object and image-object, as well as the distance between their centroids, are also applicable for single class classification (object extraction) (Leckie et al., 2003; Ragia & Winter, 2000; Whiteside et al., 2011). However, these methods require reference polygons that are not always available and have not been quantitatively tested for multi-class map validation. On the other hand, various studies focused on the development of new frameworks for assessing the accuracy of GEOBIA LULC products (Hagen, 2003; Castilla et al., 2012; Marinho et al., 2012; Whiteside et al., 2012; Hernando et al., 2012). Those studies go further than the thematic accuracy to include the spatial component in an integrated

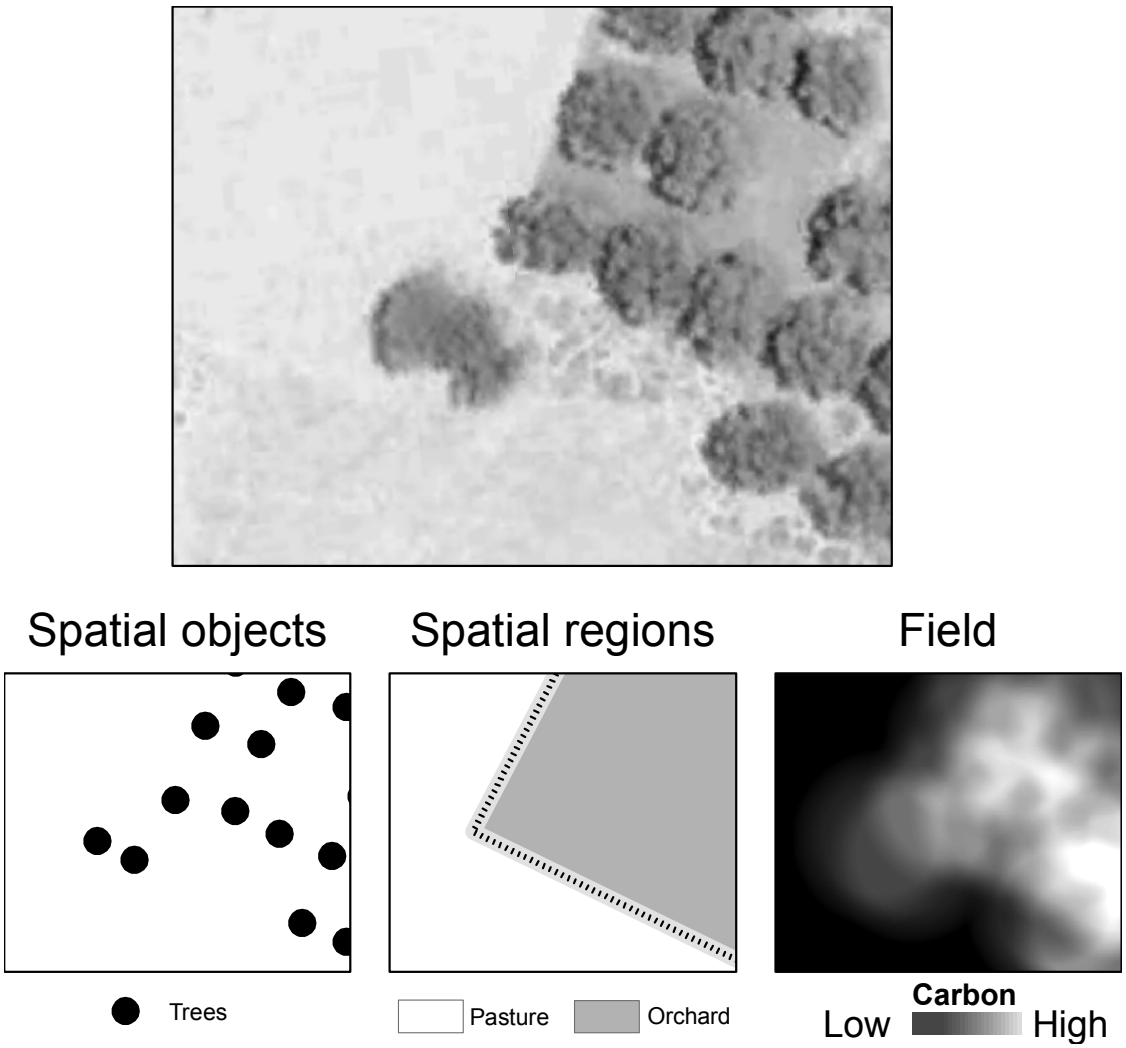


Figure 1: Representations of the same site using three different conceptual models.

index. Again, the methods to derive the proposed quality indices from a sample of the map were not quantitatively assessed.

Sample polygons have been used for a long time, in studies where the only available reference information was the photo-interpretation of higher resolution remote sensing data (George, 1986; Warren et al., 1990). However, the efficiency of the sampling was not a concern in those papers. According to Stehman & Wickham (2011), even the more recent studies paid little attention to the quantitative effects of the varying area of sample polygons on the derived accuracy indices, despite the need to account for polygon size in some ways. There is thus a need to account for the specificity of sample polygons where the findings of the standard point-based accuracy assessment methods are not applicable.

Recently, Radoux et al. (2011) proposed a predictor of the overall accuracy in an object-based image classification framework, showing that the use of sample polygons could help to increase the efficiency of the quality assessment of GEOBIA products compared with point-based sampling under some strong hypotheses. In the same framework and by working along the same line, that is a statistical prediction approach as defined by Valliant et al. (2000), the aim of this paper is to derive efficient predictors of the three primary thematic accuracy indices. These accuracy indices are thus treated as the realized values of

random variables. More specifically, the predictors for these three indices must remain efficient when the count-based accuracy is correlated with the area of the polygon.

The primary thematic map accuracy indices depict the proportion of the area of a particular map that is correctly classified at a given moment in time. The value of these indices may vary from one map to another due to the classification methodology and the landscape structure, hence the use of a predictor to target these map-dependent quantities. The thematic map accuracy, that is the focus of this study, must be distinguished from the proportion of correctly classified polygons, which is called classification accuracy in this paper.

This manuscript starts with a detailed analysis of the relationship between the area of the polygons and the classification accuracy for a GEOBIA case study (Section 2). It then describes the theoretical framework to derive the predictors of the primary map accuracy indices from spatial units of variable size (Section 3). A pragmatic implementation of the proposed predictors is then tested with synthetic maps simulated based on the parameters extracted from the case study (Section 4). Section 5 includes a quantitative comparison of the proposed framework with other methods. The advantages and drawbacks of the proposed method are finally discussed in Section 6.

2. Case study analysis

The analysis of a GEOBIA model presented in this section illustrates the various relationships between the actual and the predicted class of a polygon with respect to its area. The case study is a land use classification of a Quickbird image in the South of France. The classification was performed using a multiresolution segmentation algorithm and a combination of machine learning and statistical classifiers. The scale factor used for the segmentation in the eCognition software was low (over-segmentation) in order to minimize the number of under-segmented polygons, but a multiscale aggregation of the image-segments was used to build more meaningful land use classes.

This map is a single realization of an underlying model (i.e. it is a specific GEographic Object-Based Image Analysis on a single image). However, the analysis of the comprehensive validation results allows us to observe the driving functionals related to polygon area. The focus of this preliminary analysis is thus on the classification accuracy.

The resulting map was validated by photo-interpretation combined with the Corine Land Cover base map and a non-exhaustive reference database from field inventory for the degraded forests. The labelling and the response design used the same set of rules. These rules are non ambiguous at the spatial scale of the polygons, i.e.:

- A polygon belongs to the "urban" class if it contains more than 25 percent of buildings or infrastructures. The low percentage of man made structures set to consider a polygon as urban is due to the importance of suburbs in the mapped region. These suburbs are indeed composed of villas with large gardens.
- The "agriculture" class is defined by more than 25 percent of annual or perennial crops and less than 25 percent buildings or infrastructures. Annual crop parcel may have a bare soil land cover, depending on the vegetation cycle. Perennial crops include groves and vineyards.
- The "natural vegetation" class includes the polygons covered with more than 75 percent natural ligneous vegetation, including trees and shrubs.
- The "degraded natural vegetation" class is characterized by more than 75 percent of natural vegetation or bare soil cover, but less than 75 percent of ligneous vegetation. This class includes fire-breaks and post-fire vegetation regrowth.
- The "water" polygons must include at least 75 percent of water (fresh or salty).

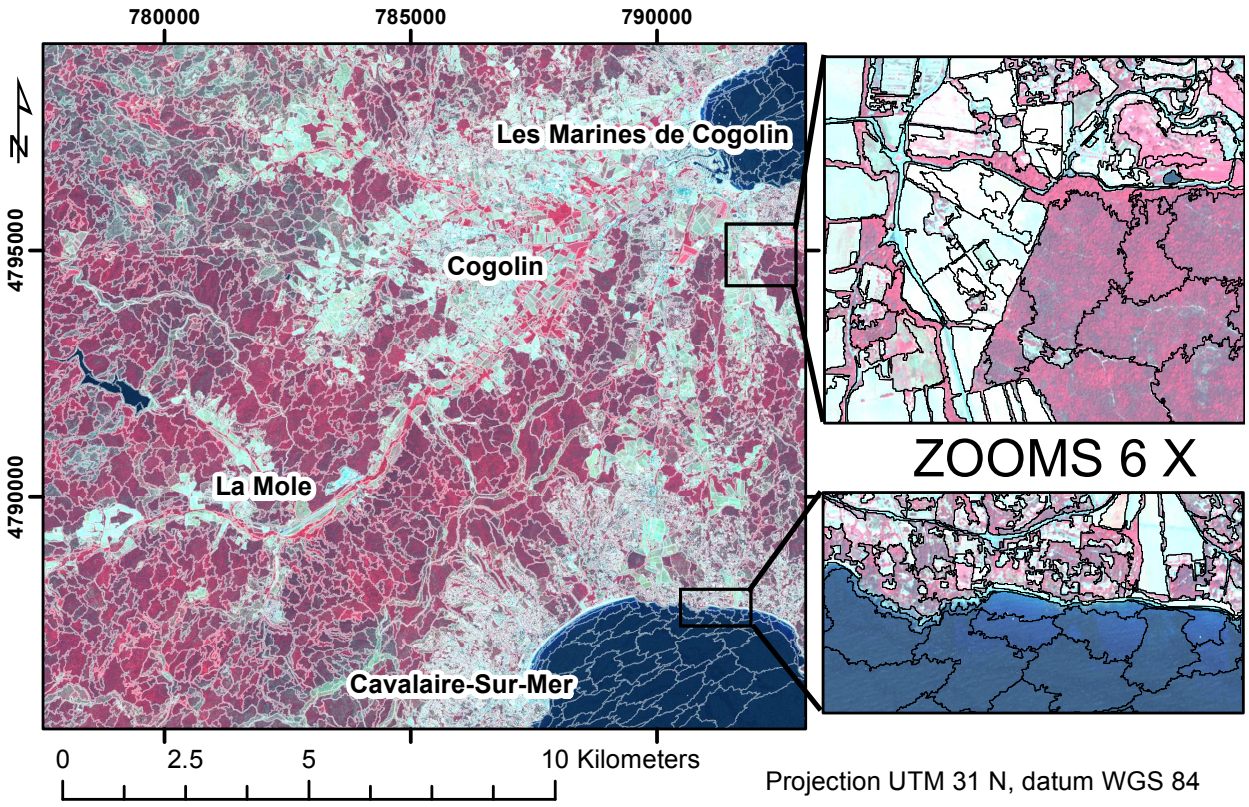


Figure 2: Subset of the study area illustrating the diversity of sizes in the segmented image.

A qualitative analysis of the results of the segmentation highlights a diversity of polygon sizes due to the landscape structure and the segmentation algorithm (figure 2). For instance, there was a set of small polygons in the urban areas and very large ones in the sea. On one hand, this is caused by the use of a spectral heterogeneity threshold in the segmentation algorithm. In the study area, cities are indeed composed of buildings, roads, swimming pools and vegetation that can be individually identified at the spatial resolution of the Quickbird image and hence contribute to a large variance at the image-segment level. On the other hand, annual crop fields and water bodies have a similar homogeneity, but their extent in the landscape is very different. The sea consists in a single large surface while the crop fields are relatively small.

To sum up, the size distribution is driven by the landscape structure and the spectral homogeneity of the spatial regions, which are both linked with the land use class. The corollary is that the probability of occurrence for each class depends on the polygon's area (figure 3). Independently of the class, the majority of the polygons belongs to the same category of size, but the largest polygons are up to 10 000 times larger than the smallest ones. The size distribution is therefore characterized by a large coefficient of variation (1.96) that reflects a strong asymmetry in the corresponding distribution.

Radoux & Defourny (2008) observed an improvement of the inter-class separability when the average polygon area increases. Those improvements were explained by the fact that the intra-class variance is reduced by the use of the mean spectral value of each polygon. The reduction of the confidence interval on the estimated spectral mean can also be observed at the level of individual polygons. The classification accuracy based on maximum likelihood classifiers is therefore expected to be larger when the inter-class separability increases. On the other hand, large polygons are more likely to be under-segmented and hence less representative of their class. Those polygons could in turn be misclassified because their mean spectral values do not reflect their actual content.

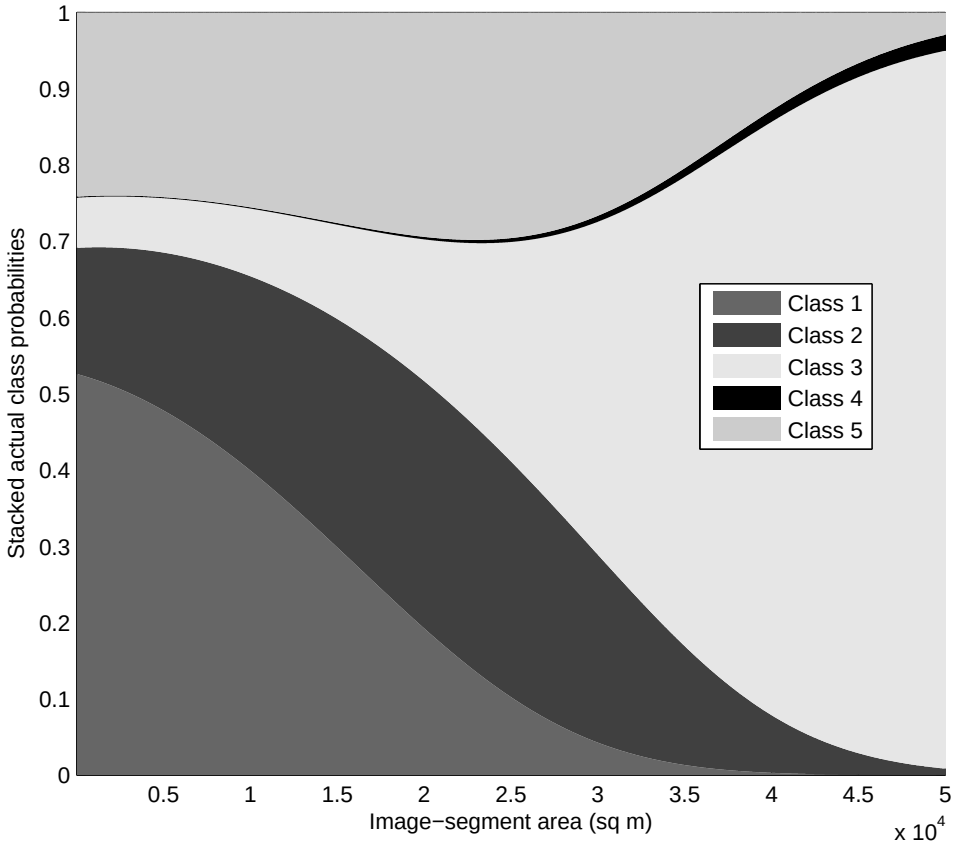


Figure 3: Probability of each actual class with respect to the area of the polygon.

Table 1: Count-based user and overall accuracies in percent, depending on the area of the polygons. The percentages of each class in terms of number and of area of the polygons is also provided.

Quartile	1	2	3	4	overall	count %	area %
Urban	86	70	67	90	77	30	13
Agriculture	47	64	65	72	62	21	14
Forest	50	69	77	87	82	25	49
Water	90	100	100	100	99	2	9
Degraded forest	42	49	50	64	50	22	15

The analysis of the accuracy assessment results presented in Table 1 shows that the per-class classification accuracy (estimated based on 500 photo-interpreted polygons per class) tends to increase when the polygon size increases. However, the relationship between the polygon area and the per-class classification accuracy is not always monotonic. Water, for instance, is only misclassified for small polygons (shallow water) and degraded forests are on average better classified when the polygon area increases. On the other hand, the urban class is not necessarily better classified with larger areas. This is caused by the large heterogeneity of medium to large urban polygons, where the proportion of impervious surface can range from 30 to 80 percent, while small urban objects are generally homogeneous. As a remark, similar trends in the per class and per size classification accuracy were also observed (but not shown here) with other classifiers (artificial

neural network, decision trees and k-nearest neighbours). This suggests that the effect of the polygon's area on the classification accuracy should always be considered in GEOBIA accuracy assessment.

3. Theoretical framework

3.1. Notations and formalization

Let us consider a set of N disjoint polygons that compose the totality of a map, where S_i (with $i = 1, \ldots, N$) is the size (area) of the i^{th} polygon. We will assume that each polygon belongs to a single actual class j among a set of k exhaustive and mutually exclusive possible classes. In practice, this requires a classification system that correctly handles geometric differences and polygon heterogeneity. In an object-based classification context, each polygon will also be classified into one and only one of these k classes. Let us denote $\Omega = \{1, \ldots, k\}$ as the set of these classes. As each polygon belongs to a single class, let us define a_i as the actual class associated with the i^{th} polygon. In addition, α_{ij} is defined as the membership of the i^{th} polygon to the class j. As we assumed that the membership is unambiguously defined, α_{ij} is equal to 1 when the actual class a_i of the i^{th} polygon belongs to the set j and is equal to 0 otherwise. This is formalized by

$$\alpha_{ij} \equiv \delta_{(a_i=j)} = \begin{cases} 1 & \text{if } \; a_i = j \\ 0 & \text{if } \; a_i \neq j \end{cases} \qquad \forall i = 1, \ldots, N \tag{1}$$

where $\delta_{(.)}$ is the Kronecker delta operator. Similarly, according to the results of the classification, we can define b_i as the predicted class of this polygon, as well as the corresponding variable $\beta_{ij} \in \{0, 1\}$, with

$$\beta_{ij} \equiv \delta_{(b_i=j)} = \begin{cases} 1 & \text{if } \; b_i = j \\ 0 & \text{if } \; b_i \neq j \end{cases} \qquad \forall i = 1, \ldots, N \tag{2}$$

i.e. β_{ij} is equal to 1 when the predicted class b_i of the i^{th} polygon belongs to class j and is equal to 0 otherwise. Clearly, the i^{th} polygon is correctly classified if and only if $a_i = b_i$, that is $\sum_{j=1}^{k} \alpha_{ij}\beta_{ij} = 1$.

Using the previous notations, the producer's, user's and overall accuracies, as defined by Congalton & Green (2009), can be expressed and predicted in the same consistent framework. Indeed, as the set of all classes consists in a partition of Ω, then

$$\pi_\Omega = \frac{\sum_{i=1}^{N} \sum_{j=1}^{k} \alpha_{ij}\beta_{ij}S_i}{\sum_{i=1}^{N} S_i} \tag{3}$$

is the overall accuracy, that is the proportion of the map area that is correctly classified. Similarly, the producer's accuracy ($\pi_{p,j}$) and user's accuracy ($\pi_{u,j}$) for any arbitrarily chosen class j are written as

$$\pi_{p,j} = \frac{\sum_{i=1}^{N} \alpha_{ij}\beta_{ij}S_i}{\sum_{i=1}^{N} \alpha_{ij}S_i} \qquad \pi_{u,j} = \frac{\sum_{i=1}^{N} \alpha_{ij}\beta_{ij}S_i}{\sum_{i=1}^{N} \beta_{ij}S_i} \qquad \forall j \in \Omega \tag{4}$$

3.2. Prediction of the overall accuracy

The actual class a_i needs to be determined, for a subset of the map and according to the response design, in order to compute the accuracy indices. Therefore, let us consider an equal probability random sample of n polygons (with $n \leq N$) drawn without replacement from the set of N polygons in order to build the reference dataset. Splitting accordingly summation over these polygons in the numerator leads now, e.g. for the overall accuracy, to the expression

$$\pi_\Omega = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \alpha_{ij}\beta_{ij}S_i + \sum_{i=n+1}^{N} \sum_{j=1}^{k} \alpha_{ij}\beta_{ij}S_i}{\sum_{i=1}^{N} S_i} \tag{5}$$

For any given map, all S_i's and β_{ij} are observable $\forall i = 1, \ldots, N$, but the α_{ij}'s are only observable for the sampled polygons ($\forall i = 1, \ldots, n$). In the proposed framework, α_{ij} is thus a random variable for non sampled

polygon, hence the second term of the numerator needs to be estimated by replacing the unknown values with their expectation.

$$\widehat{\pi}_\Omega = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ij}\beta_{ij}S_i + \sum_{i=n+1}^{N}\sum_{j=1}^{k}\widehat{E}[\alpha_{ij}\beta_{ij}]S_i}{\sum_{i=1}^{N}S_i} \tag{6}$$

This problem was already addressed in Radoux et al. (2011) where the correct classification was defined as C_i, i.e. a Bernouilli distributed random variable equal to 1 when the polygon was correctly classified and to zero otherwise. A predictor for π_Ω was then obtained by defining $C_i = \sum_{j=1}^{k}\alpha_{ij}\beta_{ij}$ and replacing it with its expectation $E[C_i] = \sum_{j=1}^{k}E[\alpha_{ij}\beta_{ij}]$. In this case, the predictor relied on the knowledge of the expected value of C_i, where C_i was assumed to be identically distributed. Based on the sampled polygons, an estimate of the expectation was given by $\widehat{E}[C_i] = \sum_{i=1}^{n}C_i/n = \widehat{p}$, where the probability p was considered as constant.

However, as shown in Section 2, some dependence is expected to occur with respect to the class and to the size, so that the previous result is suboptimal. Various models may indeed be considered related to per class accuracy and the possible relationship between polygon size and accuracy. Furthermore, for any given classified polygon i, one and only one of the β_{ij}'s is non null and equal to 1 (when $j = b_i$). We thus have in particular

$$\sum_{j=1}^{k}E[\alpha_{ij}\beta_{ij}] = \sum_{j=1}^{k}E[\alpha_{ij}|b_i = j]\beta_{ij} = E[\alpha_{ij}|b_i = j] := p_{b_i} \qquad \forall i \qquad \text{with } b_i \in \Omega \tag{7}$$

where the conditional probability of belonging to each class is not *a priori* a constant value. Accordingly, the overall accuracy predictor is now given by

$$\widehat{\pi}_\Omega = \frac{\sum_{i=1}^{n}S_i\sum_{j=1}^{k}\alpha_{ij}\beta_{ij} + \sum_{i=n+1}^{N}\widehat{p_{b_i}}S_i}{\sum_{i=1}^{N}S_i} \tag{8}$$

where the class to pick up for the i^{th} polygon in eq. (8) is the attributed class b_i for this polygon during the classification. The theoretical values for $p_1, \ldots, p_k$ are *a priori* unknown, but they can be estimated from the set of sampled polygons as the observed relative frequencies of the various classes conditionally on the label assigned by the classification process.

$$\widehat{p_j} = \sum_{i=1}^{n}\alpha_{ij}\beta_{ij}\Big/\sum_{i=1}^{n}\beta_{ij} \qquad \forall j \in \Omega \tag{9}$$

Clearly, as predicted and actual classes are supposed to show a good correspondence rate if the classification process provides meaningful results, doing so is expected to reduce the uncertainty about the knowledge of the actual class (the better the classification, the higher the probability of belonging to the same class, where a probability value of 1 would be equivalent to the situation where the actual class is certain).

These first results need to be extended by working along two lines. The first one is the extension of the methodology in order to obtain distinct predictors for the producer's and the user's accuracy, in addition to the overall accuracy. The second one will focus on alleviating some of the assumptions that were set so far for the p_j's, because the case study highlighted that the p_j values depend on the size of the polygons. These aspects will be treated in separate sections for the sake of clarity.

3.3. Prediction of user's accuracy

For each class j, the user's accuracy is defined as the ratio between the correctly classified area for the class and the total area classified as belonging to this class (eq. 4). Similarly to eq. 5, the summation of the numerator can be split among the set of sampled polygons ($i = 1, \ldots, n$) and the remaining ones ($i = n+1, \ldots, N$). After replacing the unknown value by their expectation and using again $E[\alpha_{ij}|b_i = j] \equiv p_j$, the predictor is given by

$$\widehat{\pi}_{u,j} = \frac{\sum_{i=1}^{n} \alpha_{ij}\beta_{ij}S_i + \sum_{i=n+1}^{N} \widehat{p_j}\beta_{ij}S_i}{\sum_{i=1}^{N} \beta_{ij}S_i} \qquad \forall j \in \Omega \tag{10}$$

As already stated, the predicted classes (i.e., the b_i's) are known for all polygons, and so are the β_{ij}'s that do not need to be estimated. Basically, the prediction problem is thus solely relying on the estimation of the various p_j's.

3.4. Prediction of the producer's accuracy

For each class j, the producer's accuracy is defined as the ratio between the correctly classified area for this class and the total area that actually belongs to this class (eq. 4). Using the previous notations and splitting the summation as before, the producer's accuracy for class j is then given by

$$\pi_{p,j} = \frac{\sum_{i=1}^{n} \alpha_{ij}\beta_{ij}S_i + \sum_{i=n+1}^{N} \alpha_{ij}\beta_{ij}S_i}{\sum_{i=1}^{n} \alpha_{ij}S_i + \sum_{i=n+1}^{N} \alpha_{ij}S_i} \tag{11}$$

Replacing the unknown α_{ij}'s with their corresponding expectations in order to obtain a predictor of $\pi_{p,j}$ gives the result

$$\widehat{\pi}_{p,j} = \frac{\sum_{i=1}^{n} \alpha_{ij}\beta_{ij}S_i + \sum_{i=n+1}^{N} \widehat{p_j}\beta_{ij}S_i}{\sum_{i=1}^{n} \alpha_{ij}S_i + \sum_{i=n+1}^{N} \widehat{E}[\alpha_{ij}S_i]} \tag{12}$$

In opposition to the predictors for the overall and the user's accuracy, one part of the denominator also needs to be estimated. Obviously, the actual class is indeed unknown for non sampled polygons. Like for p_j, $E[\alpha_{ij}]$ could be estimated as a frequency. This estimation can be performed, for each actual class j, conditionally on each predicted class j' in order to take advantage of the knowledge of $\beta_{ij'}$ for all non sampled polygons.

$$\widehat{E}[\alpha_{ij}S_i] = \sum_{j'=1}^{k} \widehat{E}[\alpha_{ij}|b_i = j']\beta_{ij'}S_i \tag{13}$$

3.5. Functional dependence on polygons' area

As is, the values for the various $E[\alpha_{ij}|b_i = j']$'s could be directly estimated from a simple confusion matrix where predicted classes are crossed with actual classes based on the set of sampled polygons. However, it was shown in Table 1 that the classification accuracy could also depend on the area of the polygons. In order to account for the size effect, let us assume that, for each class j, there exists a marginal probability distribution function (*pdf*) of polygon area $f_j(s)$. Using Bayes theorem, it then comes that

$$E[\alpha_{ij}|b_i = j', S_i = s] = \frac{f_j(s)E[\alpha_{ij}|b_i = j']}{f(s)} \qquad \forall j \in \Omega \tag{14}$$

where $f(s) = \sum_{j=1}^{k} f_j(s)E[\alpha_{ij}|b_i = j']$ is the marginal *pdf* for the area (i.e. irrespective of the class).

In practice, the $f_j(s)$'s can be estimated from the sample (e.g., using a kernel density estimator approach or a logistic regression) for each class conditionally on the known labels. These results, substituted into eqs. 8, 10 and 12, provide improved predictors of the accuracy indices. Indeed, knowing the polygon's area provides additional information about the true class when the area distributions differ among classes, as it was shown to be the case in figure 3.

4. Empirical quality assessment method

The characteristics of the predictors are derived in this study using a Monte Carlo procedure that makes use of synthetic maps and simulated samples. The overall process is represented in figure 4. For each synthetic map, the measured accuracy indices are compared with their predicted values based on a set of samples. The statistical efficiency of the predictor is then evaluated using the absolute root mean square error (RMSE) on the accuracy values. This unique value encompasses the bias and the variance of the predictors.

The two goals of this empirical quality assessment are to verify (i) the accuracy (being centred on the target parameter to be estimated) and (ii) the efficiency (achieving a high precision with a limited number of sampling units) of the proposed predictors as defined in Stehman (2001). The overall, the user's and the producer's accuracies are computed within this controlled experiment.

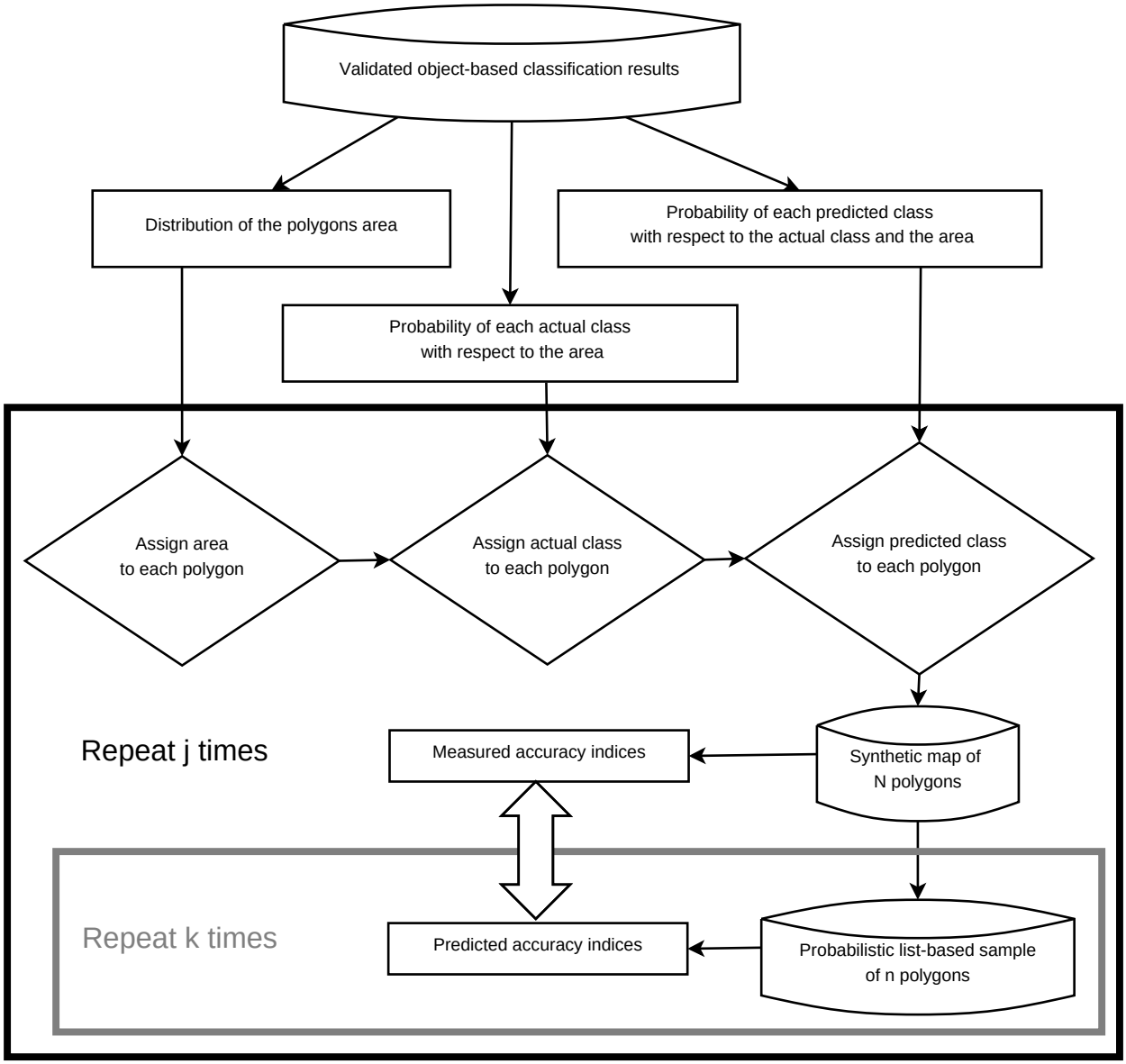


Figure 4: Schema of the synthetic case study. Values for j and k are set to 200 in this study.

4.1. Synthetic maps

The Monte Carlo procedure is based on the simulation of synthetic maps having characteristics derived from the case study (Figure 3 and table 1). This method is similar to the procedure proposed by Radoux et al. (2011), where a binary value for correct/incorrect classification was drawn for each polygon. In the present study, the values of the predicted class and of the actual class are drawn separately instead of the correct/incorrect classification result. This was necessary in order (i) to incorporate the information about β_{ij} in equations 8, 10 and 12, and (ii) to compute and predict the user's and producer's accuracies.

The total number of polygons per map, N, is fixed to 5 000. Though specifying the number of polygons lead to maps with slightly different total areas, the absolute area of the maps does not affect the quality assessment results. It can indeed be shown that the value of the primary accuracy indices is not sensitive to a rescaling of the map.

The area, the actual class and the predicted class are assigned for each polygon in three successive steps :

1. The area of the polygon is randomly drawn from the cumulated probability distribution function of the polygons area distribution.
2. The area of the polygon is used to derive the probability of each actual class. The actual class can then be randomly selected.
3. Knowing the area and the actual class, the functions $f_j(s)$ are selected to find the probability of each predicted class. Like in the previous steps, the predicted class is randomly chosen according to this distribution. This step mimics the classification process.

Two distinct sets of 200 maps are used in the quantitative assessment of the predictors of accuracy indices. These sets differ by the type of size dependence used to produce the synthetic maps. The size of each polygon is assigned using the same rules for each set. However, the dependence between the classification accuracy and polygon area varies:

- The first set, called Independent/Independent (II), is the most simple. Both actual and predicted classes are assigned independently of the polygon size. The classification accuracy is however different for each class because class-specific $\widehat{p_j}$'s are used. This set of maps is used for the fair comparison with Radoux et al. (2011), which assumed the independence between the size and the classes.

- For the second type of sets, called Dependent/Dependent (DD), the actual classes are first selected based on area-related probabilities (figure 3). Different p_j's are then set depending on the land use/land cover classes and the area-based classes. The area-based classes consist in deciles using interpolated values based on table 1. By design, all the previously identified effects of the size on the classification accuracy are thus simulated. These synthetic maps are used to test the predictors under the most realistic simulations of GEOBIA results according to the case study analysis.

Figure 5 illustrates the variability of the second set of synthetic maps. The average coefficient of variation of the polygon size is 1.8 and the average overall accuracy is 0.817. This map accuracy is markedly different from the average classification accuracy (0.697). The different simulations are well distributed around these two values with a range of 0.2 for the coefficient of variation and 0.03 for the overall accuracy.

4.2. Analysis

As shown in section 3, the proposed predictor relies on the estimation of the p_j values. Statistical analysis provides several tools to estimate such functional, including parametric and non parametric approaches. In this study, the underlying model is known : the classification accuracy is either constant for each class or related to the size categories defined by deciles. Predictors with area-independent $\widehat{p_j}$'s and with p_j's estimated with four (i.e. using quartiles) size categories were thoroughly tested. These class-dependent models are referred to as CDQ4 and CD, respectively. In addition, CDQ-like predictors with logistic regression or with 3 and 10 quantiles have been tested in order to give a hint about the robustness of the method with respect to matching between the selected internal model and the actual underlying model.

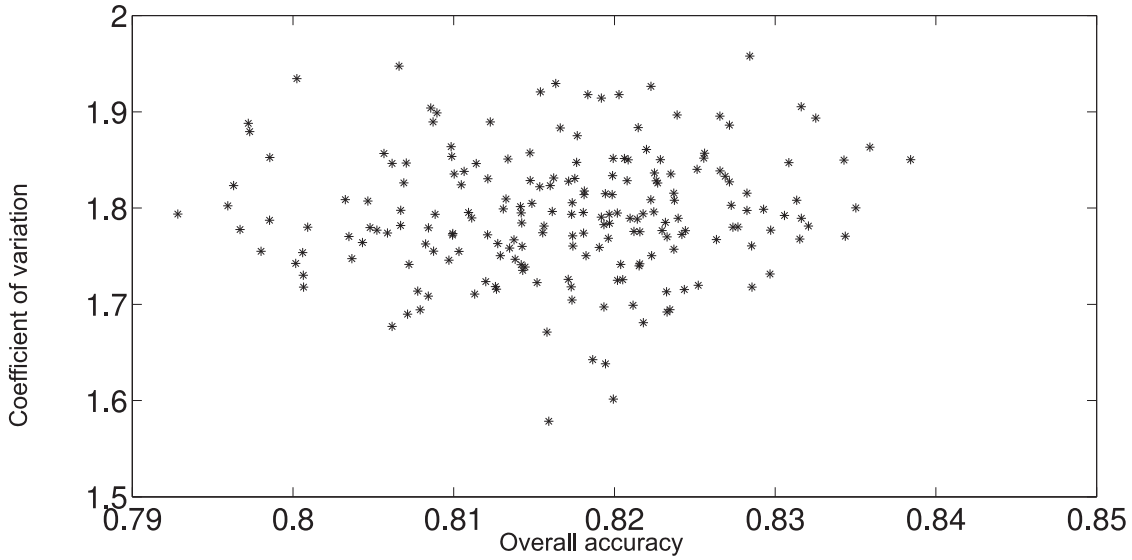


Figure 5: Diversity of the synthetic maps with respect to their overall accuracy and coefficient of variation.

In addition to the method proposed in section 3, an area-weighted (AW) predictor and a class-independent (CI) predictor (Radoux et al., 2011) are also tested. The equations implemented for the comparative study are listed in table 2. An updated version of the class-independent predictor, using quartiles to estimate the values of p and referred to as CIQ, is also tested for the sake of comparison with CDQ.

Table 2: Alternative accuracy assessment predictors using sample polygons, namely Area-Weighted (AW)(e.g. in Desclée et al. (2006)) and class-independent (CI) (Radoux et al., 2011). n is the number of selected sample polygons, C_i is equal to 1 or 0 if the sample polygon is correctly classified or not, respectively, and S_i is the area of the polygon.

Predictor	Equation
AW	$\sum_{i=1}^{n} S_i C_i \Big/ \sum_{i=1}^{n} S_i$
CI	$\left(\sum_{i=1}^{n} C_i S_i + \frac{1}{n}\sum_{i=1}^{n} C_i \sum_{i=n+1}^{N} S_i\right) \Big/ \sum_{i=1}^{N} S_i$

In the case of a quantile-based estimation of the classification accuracy for the prediction of the overall accuracy, $E[\alpha_{ij}|b_i = j', S_i = s]$ reverts to $E[\alpha_{ij}|b_i = j', S_i \in q]$ where q is the area class defined by the quantiles. Equation 14 then simplifies to

$$\widehat{E}[\alpha_{ij}|b_i = j', S_i \in q] = \frac{\sum_{m=1}^{n_q} \alpha_{mj'}\beta_{mj'}}{\sum_{m=1}^{n_q} \beta_{mj'}} \qquad \forall j', q \tag{15}$$

where m is the index of the polygon belonging to the area class q and n_q is the number of sampled polygons in this area class. Equation 15 is then substituted in equations 8, 12 and 10 for the prediction of the accuracy indices. For instance, equation 8 becomes equation 16 when quartiles are used to define the boundaries of the area classes.

$$\widehat{\pi}_{\Omega} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} \alpha_{ij}\beta_{ij} S_i + \sum_{j=1}^{k}\sum_{q=1}^{4}\left[\frac{\sum_{m=1}^{n_q} \alpha_{mj}\beta_{mj}}{\sum_{m=1}^{n_q} \beta_{mj}} \sum_{m=n_q+1}^{N_q} \beta_{mj} S_m\right]}{\sum_{i=1}^{N} S_i} \tag{16}$$

4.3. Sampling design

Equal probability sampling is used in all cases. Inside each area-based class, n_q polygons are drawn without replacement, where n_q is the total sample size divided by the the number of size categories. All polygons are thus selected with the same probability and each area-based class is equally represented.

In this study, the sampling is repeated 200 times per synthetic map. Because our class statistics are based on the case study, there are 5 land use classes. Obviously, this method may need to be adjusted in case of larger number of classes in order to avoid empty bins, but this is outside the scope of this study. For the sake of simplicity, empty bins are here filled with the average classification accuracies per class, which may introduce a small bias.

5. Results

The first set of maps shows the performances of the different predictors under the scope of validity of all the compared predictors (figure 6, top). On one hand, it is observed that the larger the amount of class-related information that is used for a predictor, the better its efficiency : the AW predictor is the least efficient, followed with the CI predictor and eventually with the CD predictor. On the other hand, the estimation of the parameters of the area-related relationships increased the variance of the predictors, also because of the different sampling design. As a results, the two size-independent predictors had a smaller standard deviation than the corresponding predictor with quartile-based size effect modelling. However, this difference is relatively small and overfitting the size effect did not introduce a bias in the quantile-based predictors. Overall, CDQ still remains $\approx 20\%$ more efficient than AW in the range of 500 to 1 000 sample polygons.

If there is a size effect on the classification accuracy, as for the second set of synthetic maps (DD), the area-independent predictors (CI and CD) need to be discarded due to the contribution of large absolute biases ($> 10\%$) to their RMSE (fig 6 bottom). The predicted map accuracy indeed tends toward the classification accuracy when i) the size effect is neglected and ii) $n << N$. As shown in figure 7 with 200 maps and samples of 600 polygons, the CDQ predictor still slightly underestimates the overall accuracy by 0.6 percent on average. The absolute value of this bias decreases linearly with the number of samples until it reaches 0 at $n = N$; it increases in the other direction (-0.8 with 200 samples for DD and CQ4). On the other hand, the AW predictor is statistically unbiased for any sample size. But, despite AW being more accurate, the CDQ predictor is in this case $\approx 30\%$ more efficient than the AW predictor because its lower variance compensates its (small) bias.

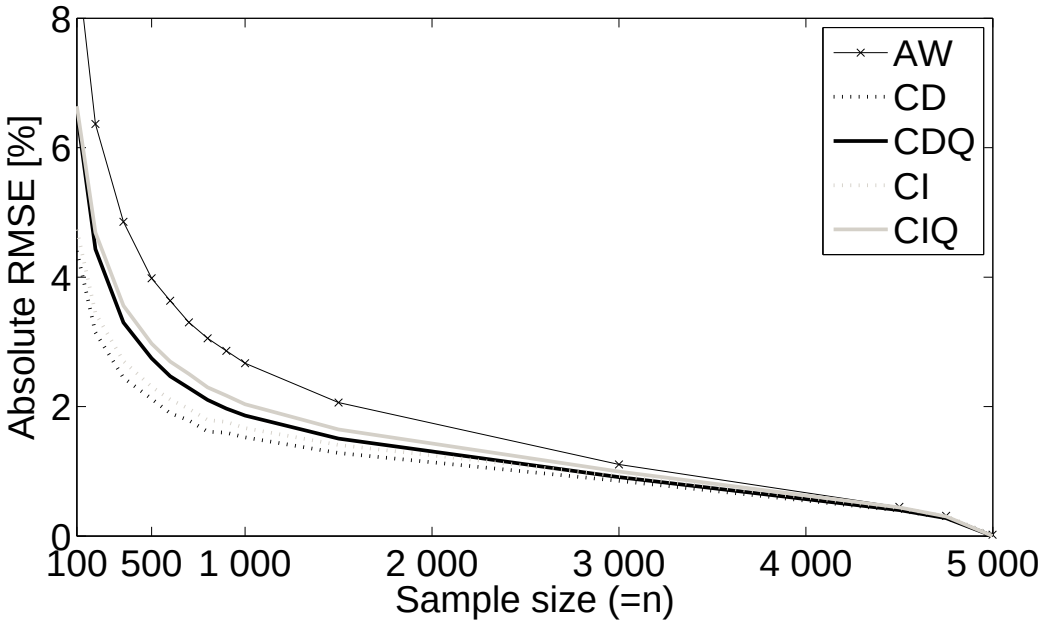
In practice, map producers aim at achieving a given level of uncertainty on their accuracy indices with as few sampling units as possible in order to reduce the validation costs. When we compare the number of polygons needed to achieve the same RMSE, it appears that approximately twice as many sample polygons are needed for AW compared with CDQ with both sets of maps. In particular, the RMSE achieved with 500 sample polygons using CDQ would need 1 000 sample polygons when using the AW predictor for the second set of maps, and 935 sample polygons with the first set of maps.

The RMSE of the predictors of the user's and the producer's accuracies are markedly larger than for the overall accuracy (Table 3). This is obviously due to the dispersion of the sample polygons between the different classes; the actual classification accuracies (which come from table 1) also contribute to those differences. Most of the time, it can be seen that the CDQ predictor of the user's accuracy is more efficient than the AW predictor thanks to the additional information provided by the knowledge of the class label. However, a poor relationship between α_{ij} and β_{ij} reduces the available knowledge useful for our predictor.

For the case of the producer's accuracy, the proposed method does not benefit from the additional information that is knowing the predicted values for all polygons. The results of the producer's accuracy CDQ predictor are however better than those of user's accuracy predictor in the II set, and systematically outperforms the AW predictor in this case. This is no longer the case when there is an effect of the area of the polygons on the classification accuracy, due to the uncertainty when fitting the $f_j(s)$'s. In this case, the AW predictor is most of the time (4 out of 5 classes) better than the CDQ predictor.

Table 5 shows the main statistics of the CDQ-like predictors when different models are used internally to estimate the p_j's. These models differ by the number of quantiles used. The best results are achieved using

First set of maps (II)



Second set of maps (DD)

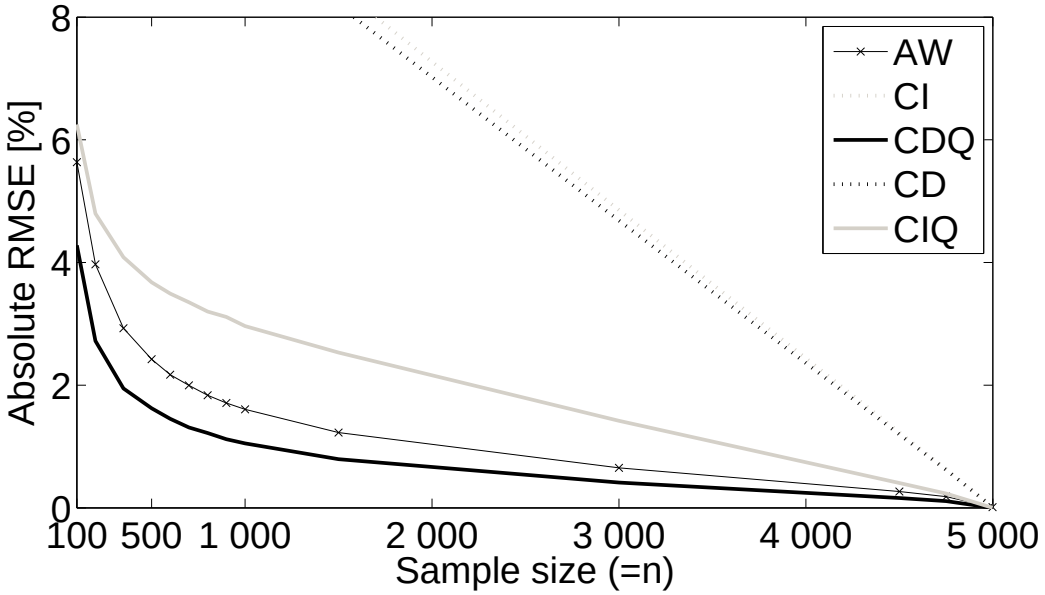


Figure 6: Comparison of the absolute RMSE for the overall accuracy prediction ($N = 5\,000$) with sample size ranging from 100 to 5 000 based on AW(Area-weighted), CD (class-dependent), CDQ (CD with area quartiles), CI (class-independent) and CIQ (CI with area quartiles) predictors.

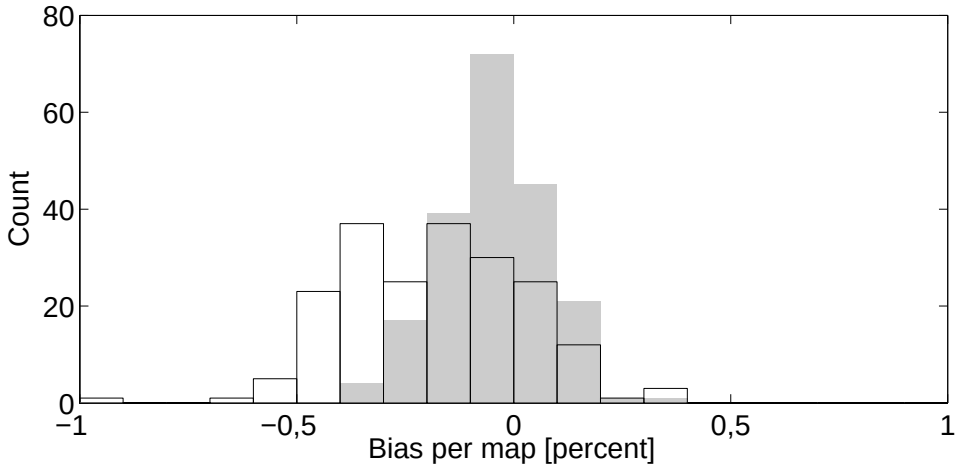


Figure 7: Histogram of the bias of the overall accuracy for the second set of synthetic maps (DD) using the area-weighted (grey faces) and the class-dependent (transparent faces) predictors.

Table 3: RMSE and bias (both in %) on the user's accuracy for 200(left) and 600(right) sample polygons in maps of 5 000 polygons and 5 classes. The synthetic maps come from the DD set.

	RMSE		Bias			RMSE		Bias	
Class	AW	CDQ	AW	CDQ	Class	AW	CDQ	AW	CDQ
1	7.5	6.9	0.2	0.3	1	4.2	3.9	0.0	0.3
2	7.9	8.2	-0.1	-0.3	2	4.3	4.3	-0.1	-0.3
3	4.0	5.5	-0.3	-3.3	3	2.2	3.9	-0.3	-3.0
4	11.3	10.7	-1.7	0.9	4	0.7	0.5	-0.1	-0.0
5	10.4	11.7	0.7	5.4	5	5.6	7.3	0.1	4.8

Table 4: RMSE and bias (both in %) on the producer's accuracy for 200(left) and 600(right) sample polygons in maps of 5 000 polygons and 5 classes. The synthetic maps come from the DD set.

	RMSE		Bias			RMSE		Bias	
Class	AW	CDQ	AW	CDQ	Class	AW	CDQ	AW	CDQ
1	7.3	7.4	0.2	-1.5	1	4.0	4.4	0.0	-1.9
2	8.6	10.5	-0.2	-5.4	2	4.7	7.5	-0.1	-5.7
3	5.7	7.0	-0.1	-4.0	3	3.2	5.0	-0.0	-3.8
4	0.3	0.1	-0.0	0.0	4	0.0	0.0	-0.0	-0.0
5	9.2	10.1	-0.0	3.5	5	5.0	6.3	-0.1	3.8

deciles as it matches the model used to create the synthetic maps. CDQ however remains more efficient than AW with as few as 4 quantiles, while it becomes noticeably affected by its bias with 3 quantiles. On the other hand, CDQ10 remained as efficient with a set of synthetic maps based on quartiles, so that overfitting does not seem to be an issue. In addition, a logit model has been tested. The results of this model are comparable with the decile-based model. However, fitting the logit model requires a minimum number of points per class. Failure to fit the logit model rarely (0.3%) happens with $n = 600$, but it is frequent (62%) when $n = 200$.

Table 5: Comparison of the predictors of the overall accuracy with $N = 5\,000$ and $n = 600$ under the DD4 synthetic map dataset.

Model	Bias	Standard deviation	RMSE(in %)
CD 3 quantiles	-1.6	1.6	2.3
CD 4 quantiles	-0.6	1.5	1.7
CD 10 quantiles	-0.0	1.2	1.2
CD Logit	0.0	1.2	1.2
AW	-0.0	2.2	2.2

6. Discussion

This paper provides a methodological framework for the thematic accuracy assessment of a map based on sample polygons. The results show that the relationship between the classification accuracy and the area of the polygons has to be taken into account in order to efficiently predict the primary thematic accuracy indices. Besides the statistical concerns that are due to their variable area, the initial choice of polygons as sampling units should also be discussed with regard to the delineation errors. These two aspects are presented in separate subsections.

6.1. Efficiency and area dependence

Evidence of the relationship between the polygon's area and its classification accuracy were quantitatively highlighted in a case study. Overall, there is a trend to better classify the largest polygons. However, this is not true for all classes. While it seems plausible to encounter some relationships between the classification accuracy and the size of polygons with most GEOBIA results, the type of these relationships could not be generalised based on our results.

Accounting for the size of the polygons is necessary for an exact prediction of the thematic map accuracy indices. The proposed theoretical framework allows to account for any $f_j(s)$ model of the various size effects, so that it is in theory possible to find a model that fits reality. However, the variety of size effects and the relatively small sample at hand make the estimation of the $f_j(s)$ parameters challenging. In practice, the simple $f_j(s)$ model based on quantiles emerged as a safe and efficient choice because it does not make any assumption on the shape of the distribution and reduces the risks of empty bins, compared with more elaborate quantile-based models. On a case-by-case basis, the selection of alternative parametric (e.g. logit) or nonparametric (e.g. kernel smoothing) models could improve the efficiency of the proposed predictor. For instance, further evidence of mostly monotonic relationships between the size and the classification accuracy would be in favour of a first degree logit model. The analysis of a large number of GEOBIA results will therefore provide better insight on the best eligible models. However, the results have shown that different model could yield equivalent results in terms of bias and standard deviation. This does not prove the robustness of our method in general and under any possible circumstances, but it is an indication that elaborate selection of models could be of minor benefit.

For an optimal use of the proposed method, the shape of the $f_j(s)$ functions should be statistically analysed. For instance, the number of bins can be selected using state-of-the-art statistical tools (see e.g. He & Meeden (1997)). On the other hand, some testing can be conducted in order to check the hypothesis that the classification accuracies are statistically significantly different among classes of area and land cover/land use. Krishnamoorthy et al. (2004) developed an exact method for testing the equality of several binomial

proportions to a specified standard. This method can be applied to the classification accuracy values with the overall classification accuracy per class as standard value. In case of equality (which was not the case in table 1), $f_j(s)$ would be constant. Similarly, the distributions of the classification accuracy with respect to the size can be compared 2 by 2. In case of equality, two classes can then be grouped in order to consolidate the estimation.

In any case, the AW predictor is a robust alternative that should not be discarded because it can be more efficient for user's and producer's accuracy, especially when the estimation of $f_j(s)$ is subject to large uncertainties (e.g. for small sample size and classification accuracy close to 50%). A better handling of the empty bins through elaborate model or *ad hoc* sampling designs are likely to improve the efficiency of CDQ for the class-based estimates. On the other hand, AW estimation is less efficient for the overall accuracy alone, despite the fact that it is unbiased. At this point, the results remain inconclusive but suggest using CDQ for overall and user's accuracies, and AW for the producer's accuracies. This can be done based on the same sample.

When the sampling unit has a variable size, the classification accuracy and the thematic accuracy of the map take different values (Radoux et al., 2011). While the focus of this paper was the prediction of the thematic accuracy of the map, the classification accuracy provides meaningful information about the classifier performance. This information is therefore useful to compare different classification algorithms and could be derived from the same sample than the map accuracy indices. Furthermore, the relationship between the polygon area and classification accuracy could also be used to create confidence maps.

The analytical formula of the predictors variance is complex because of the presence of estimated value in the numerator and in the denominator. Furthermore, an exact solution is not available without some assumptions on the shape of the $f_j(s)$. To date, there is indeed too few information about classification accuracy functions based on real case studies in order to generalize the use of a particular model. A preliminary sampling analysis is therefore needed on a case by case basis, which would markedly increase the costs of the validation process. However, the variance of the CI predictor (equation 13 in Radoux et al. (2011)) can be used for an approximation of the standard deviation to make sure that the sample size is large enough for the requirements of the validation plan. This theoretical variance primarily depends on $p(1-p)/n$ and $(N-n)$, but it also increases when the coefficient of variation of the polygons' area is large (which does not seem to be the case with CDQ). Anyway, the estimation of the standard deviation should be further developed in order to optimise the number of sample polygons in the validation plan or to derive confidence interval on the predicted values. In the meantime, it is possible to estimate the confidence interval *a posteriori* based on the sample alone using, e.g., bootstrap (Efron & Tibshirani, 1986).

Empirical results in our study showed that the RMSE of the CDQ predictor was smaller than for the CI predictor or for a standard point-based accuracy assessment, despite the large coefficient of variation (c_v = 1.96) of the realistic maps. For instance, the RMSE achieved with 500 sample polygons would have needed 584 sample points. In the field, however, the validation effort could be larger for polygons than for points, depending on the response design. A case by case cost analysis including both travel cost and validation cost is therefore necessary to compare the cost effectiveness of the two approaches.

6.2. Effect of delineation errors on polygon-based validation

GEOBIA usually assigns a unique label to each polygon based on crisp classifiers. The proposed method addressed the case of a response design that fits to the categorical legend. However, object-based classification relies on the delineation of polygons (from ancillary data and/or image segmentation) that could include several land cover or land use classes (e.g. due to under-segmentation). This issue exists with point-based validation, but at a different scale. The state-of-the-art solutions, which are also applicable to sample polygons, could be divided in two categories depending on the type of legend.

- If the legend is composed of pure end members, a "hard" response design can be derived using a set of decision rules including absolute or relative thresholds. For instance, Desclée et al. (2006) considered a polygon as changed if the area of change was larger than a minimum mapping unit (in that case, 0.5 ha) and Zhan et al. (2005) considered that an object is thematically correct if its match with a reference object is greater than 50%. However, Strahler et al. (2006) suggest that "hard" validation

should be replaced with "soft" validation when the percentage of mixed pixels is "too large". The "hard" validation indeed implies a trade-off between omission and commission errors depending on the selected thresholds. For dichotomic classification, the Pareto boundary can be used to analyse this trade-off (Boschetti et al., 2004) and soft validation was shown to outperform the "hard" validation protocol (Pepe et al., 2010). For a "soft" validation with multiple classes, fuzzy set theory is generally used to translate the goodness-of-fit between the reference and the classification results in case of mixed pixels (Laba et al., 2002; Muller et al., 1998; Stehman et al., 2007) or classes that are not mutually exclusive (Woodcock & Gopal, 2000). In the case of GEOBIA, soft validation as such has not been quantitatively addressed, but specific indices have been proposed. For instance, the segmentation accuracy (Liu & Xia, 2010) provides the upper bound of the overall accuracy when all polygons are correctly classified according to the majority class. The theoretical framework proposed in our study could also be adjusted to "soft" validation by defining $\alpha \in [0,1]$ instead of $\alpha \in \{0,1\}$. However, a low segmentation accuracy should warn the end user that GEOBIA was not appropriate for the data model. In such a case, the use of pixel-based validation should be preferred instead of complex soft validation with sample polygons.

- On the other hand, it is always possible to build a legend composed of a set of non overlapping classes, and the semantic content of a polygon can be higher than a pixel. For instance, the LCCS (UN Land Cover Classification System or Land Cover Meta Language (Di Gregorio & Jansen, 2000; Di Gregorio, 2005)) uses basic classifiers and their spatial relationships for standardized and consistent class definitions (Bajracharya et al., 2010). There is now a general agreement that LCCS provides a valuable common land cover language for building land cover classes (Herold et al., 2008). The use of LCCS combined with classifier-based response design is thus a key for the thematic validation of GEOBIA results as it allows the map producers to distinguish thematic mixed classes and spatial mixed classes (see Di Gregorio (2005)). The thematic precision and the formalization of the legend are then additional indices of the map quality. A high thematic accuracy with a poor thematic precision may indeed not fit the purpose of the map.

The geometric quality of the product is another important aspect of the map validation, which can interact with the thematic quality assessment, especially in fragmented landscapes (Smith et al., 2003). Therefore, Couturier et al. (2009) proposed a fuzzy framework to handle thematic and positional accuracies together. However, these approaches make more difficult the comparison between maps because the error sources are not isolated. In our paper, we focused on the thematic map accuracy in order to provide tools for the statistical comparison of different classification results. Geolocation quality is therefore externalised, assuming that it can be evaluated by other means and considering that geolocation errors have less impact on the thematic classification results with polygons than with smaller sampling units (Stehman & Wickham, 2011; Strahler et al., 2006). In any case, assessing the geometric quality requires additional efforts. For instance, Schopfer et al. (2008) used manually delineated reference polygon to derive topological and geometric information about image-objects, which is summarised in Object Fate Analysis matrix (Hernando et al., 2012), while Radoux & Defourny (2007) assessed the precision and the accuracy of boundary delineation based on regularly distributed points along polygon boundaries.

7. Conclusion

There is no universal best choice of the sampling unit for the accuracy assessment of a map, but polygons should prevail for GEOBIA results based on a sound legend composed of non overlapping classes (e.g. LCCS-based legend). In this case, information on the size distribution and the classification results can be used to improve the efficiency of the prediction of the primary accuracy indices in a consistent framework.

Further work is necessary to extend the polygon-based validation framework to the high standards of point-based validation. Specific GEOBIA issues include the need for i) sample-based estimation of topological consistency and geolocation precision and ii) a better standardisation of polygon-based response designs. In addition, the issues of i) deriving the analytical expression of the variance of the proposed predictors and ii)

testing robustness of polygon-based validation with respect to different sampling designs should be further investigated.

Acknowledgement

This research was funded by the Belgian Science Policy Office (ORFEO-ASSIMIV project). The authors would like to thank the 5 anonymous reviewers for their valuable comments that helped to improve the quality of this paper. Their detailed comments were particularly appreciated.

References

Bajracharya, B., Uddin, K., Chettri, N., Shrestha, B., & Siddiqui, S. (2010). Understanding land cover change using a harmonized classification system in the Himalaya, *Mountain Research and Development*, *30* (2), 143–156

Bian, L. (2007). Object-oriented representation of environmental phenomena: Is everything best represented as an object?, *Annals of the Association of American Geographers*, *97* (2), 267–281

Blaschke, T. (2010). Object based image analysis for remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing*, *65* (1), 2–16

Bontemps, S., Bogaert, P., Titeux, N., & Defourny, P. (2008). An object-based change detection method accounting for temporal dependencies in time series with medium to coarse spatial resolution., *Remote Sensing of Environment*, *112*, 3181–3191

Boschetti, L., Flasse, S., & Brivio, P. (2004). Analysis of the conflict between omission and commission in low spatial resolution dichotomic thematic products: The Pareto Boundary, *Remote Sensing of Environment*, *91* (3-4), 280–292

Carleer, A., Debeir, O., & Wolff, E. (2005). Assessment of very high resolution satellite image segmentations., *Photogrammetric Engineering and Remote Sensing*, *71* (11), 1284–1294

Castilla, G., Hernando, A., Zhang, C., Mazundar, D., & McDermid, G. (2012). An integrated framework for assessing the accuracy of GEOBIA landcover products., in *Proceedings of the 4th GEOBIA conference, Rio de Janeiro - Brazil*, (572–575)

Clinton, N., Holt, A., Scarborough, J., Yan, L., & Gong, P. (2010). Accuracy Assessment Measures for Object-based Image Segmentation Goodness, *Photogrammetric Engineering and Remote Sensing*, *76* (3), 289–299

Congalton, R. (1991). A review of assessing the accuracy of classification of remotely sensed data., *Remote Sensing of Environment*, *37* (1), 35–46

Congalton, R. & Green, K. (2009). *Assessing the accuracy of remotely sensed data: Principles and practices*, Taylor and Francis, second edition edn.

Couturier, S., Mas, J.-F., Cuevas, G., Benitez, J., Vega-Guzman, A., & Coria-Tapia, V. (2009). An accuracy index with positional and thematic fuzzy bounds for Land-use/Land-cover Maps, *Photogrammetric Engineering and Remote Sensing*, *75* (7), 789–805

Desclée, B., Bogaert, P., & Defourny, P. (2006). Forest change detection by statistical object-based method., *Remote Sensing of Environment*, *102* (1-2), 1–11

Di Gregorio, A. (2005). Land Cover Classification System : Classification concepts and user manual for Software version 2, *FAO Environment and Natural Resources Service*

Di Gregorio, A. & Jansen, L. (2000). Land Cover Classification System (LCCS): Classification concepts and user manual., *GCP/RAF/287/ITA Africover-East Africa Project and Soil Resources, Management and Conservation Service, Food and Agriculture Organization*

Efron, B. & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, *Statistical Science 1* (1), 54–75

Foody, G. (2002). Status of land cover classification accuracy assessment., *Remote Sensing of Environment*, *80*, 185–201

George, T. (1986). Aerial verification of polygonal ressource maps: a low-cost approach to accuracy assessment., *Photogrammetric engineering and remote sensing*, *52* (6), 839–848

Hagen, A. (2002). Fuzzy set approach to assessing similarity of categorical maps., *International Journal of Geographical Information Science* , *17*, 235–249

Hansen, M. C., DeFries, R. S., Townshend, J. R. G., Marufu, L., & Sohlberg, R. (2002). Development of a MODIS tree cover validation data set for Western Province, Zambia., *Remote Sensing of Environment*, *83*, 320–335

Haralick, R. & Shapiro, L. (1985). Image segmentation techniques, *Computer Vision and Image Processing*, *29* (1), 100–132

Hay, G. & Castilla, G. (2008). Geographic Object Based Image Analysis : a new name for a new discipline., in Blaschke, T., Lang, S., & Hay, G., eds., *Object-based image analysis*, (91–111), Springer Berlin

Hernando, A., Tiede, D., Albrecht, F., & Lang, S. (2012). Spatial and thematic assessment of object-based forest stand delineation using an OFA-matrix, *International Journal of Applied Earth Observation and Geoinformation*, *19*, 214–225

He, K. and Meeden, G. (1997). Selecting the number of bins in a histogram : a decision theoretic approach., *Journal of statistical planning and inference.*, *61*, 59–69

Herold, M., Mayaux, P., Woodcock, C., Baccini, A., & Schmullius, C. (2008). Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets, *Remote Sensing of Environment*, *112* (5), 2538 – 2556, ¡ce:title¿Earth Observations for Terrestrial Biodiversity and Ecosystems Special Issue¡/ce:title¿

Huang, C., Davis, L., & Townshend, J. (2002). An assessment of support vector machines for land cover classification, *International Journal of Remote Sensing*, *23* (4), 725–749

Krishnamoorthy, K. and Thomson, J. and Cai, Y. (2004). An exact method of testing equality of several binomial proportions to a specified standard, *Computational Statistics & Data Analysis*, *45* (4), 697–707

Laba, M., Gregory, S., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J., & DeGloria, S. (2002). Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover map, *Remote Sensing of Environment*, *81* (2-3), 443–455

Lambin, E. (1999). Monitoring forest degradation in tropical regions by remote sensing: Some methodological issues, *Global Ecology and Biogeography*, *8* (3-4), 191–198

Leckie, D., Gougeon, F., Walsworth, N., & Paradine, D. (2003). Stand delineation and composition estimation using semi-automated individual tree crown analysis, *Remote sensing of environment*, *85* (3), 355–369

Liu, C., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measure of thematic classification accuracy, *Remote Sensing of Environment*, *107*, 606 – 616

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

Marinho, E., Fasbender, D., & De Kok, R. (2012). Spatial assessment of categorical maps: a proposed framework., in *Proceedings of the 4th GEOBIA conference, Rio de Janeiro - Brazil*, (602–607)

Muller, S., Walker, D., Nelson, F., Auerbach, N., Bockheim, J., Guyer, S., & Sherba, D. (1998). Accuracy assessment of a land-cover map of the Kuparuk River Basin, Alaska: Considerations for remote regions, *Photogrammetric Engineering and Remote Sensing*, *64* (6), 619–628

Neubert, M., Herold, H., & Meine, G. (2008). Assessing image segmentation quality: concepts, methods and application., in Blaschke, T., Lang, S., & Hay, G., eds., *Object-based image analysis*, (769–784), Springer Berlin

Pepe, M., Boschetti, L., Brivio, P., & Rampini, A. (2010). Comparing the performance of fuzzy and crisp classifiers on remotely sensed images: A case of snow classification, *International Journal of Remote Sensing*, *31* (23), 6189–6203

Persello, C. & Bruzzone, L. (2010). A novel protocol for accuracy assessment in classification of very high resolution images, *IEEE Transactions on Geoscience and Remote Sensing*, *48* (3, Part 1), 1232–1244

Radoux, J., Bogaert, P., Fasbender, D., & Defourny, P. (2011). Thematic accuracy assessment of geographic object-based image classification, *International Journal of Geographical Information Science*, *25* (6), 895–911

Radoux, J. & Defourny, P. (2007). A quantitative assessment of boundaries in automated forest stand delineation using very high resolution imagery., *Remote Sensing of Environment*, *110*, 468–475

Radoux, J. & Defourny, P. (2008). Quality assessment of segmentation devoted to object-based classification., in Blaschke, T., Lang, S., & Hay, G., eds., *Object-based image analysis*, (257–272), Verlag Berlin Heidelberg: Springer, Berlin

Radoux, J. & Defourny, P. (2010). Automated image-to-map discrepancy detection using iterative trimming, *Photogrammetric Engineering and Remote Sensing*, *76* (2, Sp. Iss. SI), 173–181

Ragia, L. & Winter, S. (2000). Contributions to a quality description of areal objects in spatial data sets, *ISPRS Journal of Photogrammetry and Remote Sensing*, *55* (3), 201–213

Schopfer, E., Lang, S., & Albrecht, F. (2008). Object fate analysis - A virtual overlay method for the categorisation of object transition and object-based accuracy assessment, in Blaschke, T., Lang, S., & Hay, G., eds., *Object-based image analysis*, (785–801), Springer Berlin

Smith, J., Stehman, S., Wickham, J., & Yang, L. (2003). Effects of landscape characteristics on land-cover class accuracy., *Remote Sensing of Environment*, *84*, 342–349

Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy., *Remote Sensing of Environment*, *62*, 77–89

Stehman, S. & Wickham, J. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment, *Remote Sensing of Environment*, *115* (12), 3044–3055

Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy assessment, *Photogrammetric Engineering and Remote Sensing*, *67* (6), 727–734

Stehman, S. V., Arora, M. K., Kasetkasem, T., & Varshney, P. K. (2007). Estimation of fuzzy error matrix accuracy measures under stratified random sampling, *Photogrammetric Engineering and Remote Sensing*, *73* (2), 165–173

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., Mayaux, P., Morisette, J. T., Stehman, S. V., & Woodcock, C. E. (2006). Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps, *GOFC-GOLD Report*, *25*, 1–52

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). Finite Population Sampling and Inference: A Prediction Approach.,*John Wiley & Sons, NY.*

Warren, S., Johnson, M., Goran, W., & Victor, E. (1990). An automated, objective procedure for selecting selecting representative field sample size., *Photogrammetric engineering and remote sensing*, *56* (3), 333–336

Whiteside, T., Boggs, G., & Maler, S. (2011). Extraction of tree crowns from high resolution imagery over Eucalypt dominant tropical savannas, *Photogrammetric Engineering and Remote Sensing*, *77* (8), 813–824

Whiteside, T., Maier, S., & Boggs, G. (2012). Site-specific area-based validation of classified products., in *Proceedings of the 4th GEOBIA conference, Rio de Janeiro - Brazil*, (153–157)

Woodcock, C. & Gopal, S. (2000). Fuzzy set theory and thematic maps: accuracy assessment and area estimation, *International Journal of Geographical Information Science*, *14* (2), 153–172

Zhan, Q., Molenaar, M., Tempfli, K., & Shi, W. (2005). Quality assessment for geo-spatial objects derived from remotely sensed data, *International Journal of Remote Sensing*, *26* (14), 2953–2974

Zhang, H., Fritts, J., & Goldman, S. (2008). Image segmentation evaluation: A survey of unsupervised methods, *Computer Vision and Image Understanding*, *110* (2), 260 – 280

List of Tables

1 Count-based user and overall accuracies in percent, depending on the area of the polygons. The percentages of each class in terms of number and of area of the polygons is also provided. 6

2 Alternative accuracy assessment predictors using sample polygons, namely Area-Weighted (AW)(e.g. in Desclée et al. (2006)) and class-independent (CI) (Radoux et al., 2011). n is the number of selected sample polygons, C_i is equal to 1 or 0 if the sample polygon is correctly classified or not, respectively, and S_i is the area of the polygon. 12

3 RMSE and bias (both in %) on the user's accuracy for 200(left) and 600(right) sample polygons in maps of 5 000 polygons and 5 classes. The synthetic maps come from the DD set. . . 15

4 RMSE and bias (both in %) on the producer's accuracy for 200(left) and 600(right) sample polygons in maps of 5 000 polygons and 5 classes. The synthetic maps come from the DD set. 15

5 Comparison of the predictors of the overall accuracy with N = 5 000 and n = 600 under the DD4 synthetic map dataset. 16

List of Figures

1 Representations of the same site using three different conceptual models. 3

2 Subset of the study area illustrating the diversity of sizes in the segmented image. 5

3 Probability of each actual class with respect to the area of the polygon. 6

4 Schema of the synthetic case study. Values for j and k are set to 200 in this study. 10

5 Diversity of the synthetic maps with respect to their overall accuracy and coefficient of variation. 12

6 Comparison of the absolute RMSE for the overall accuracy prediction (N = 5 000) with sample size ranging from 100 to 5 000 based on AW(Area-weighted), CD (class-dependent), CDQ (CD with area quartiles), CI (class-independent) and CIQ (CI with area quartiles) predictors. . . 14

7 Histogram of the bias of the overall accuracy for the second set of synthetic maps (DD) using the area-weighted (grey faces) and the class-dependent (transparent faces) predictors. 15