Overall accuracy estimation for geographic object-based image classification

Julien Radoux*, **Patrick Bogaert**, and **Pierre Defourny**
Earth and Life Institute-Environmental sciences
Université catholique de Louvain
Louvain-la-Neuve, Belgium
*Julien.Radoux@uclouvain.be

Abstract—**Geographic object-based image analysis is a processing method where groups of spatially adjacent pixels are classified as elementary units. This approach raises concerns about the design of subsequent validation strategies. Indeed, classical point-based sampling strategies based on the spatial distribution of sample points (using systematic, probabilistic or stratified probabilistic sampling) do not rely on the same concept of objects. New methods explicitly built on the concept of objects used for the classification step are thus needed. An original object-based sampling strategy is therefore proposed and compared with other approaches used in the literature for the thematic accuracy assessment of object-based classifications. The new sampling scheme and sample analysis are founded on a sound theoretical framework based on few working hypotheses. The performance of the sampling strategies is quantified using object-based classifications results simulated for a Quickbird imagery. The bias and the variance of the overall accuracy estimates were used as indicators of the methods benefits. The main advantage of the object-based overall accuracy predictor is its performance: for a given confidence interval, it requires less sampling units than the other methods. In many cases, this can help to noticeably reduce the sampling effort. The use of objects-based sampling units leads to practical and conceptual issues, which are sometimes, but not always, similar to those of point-based accuracy assessment. These issues (mixed entities, spatial correlation, effect of the geolocation errors, sample representativity...) are discussed with regard to the representation of environmental variables together with the limitations of the proposed method.**

Keywords: overall accuracy; object; spatial region

I. INTRODUCTION

Land cover maps are of paramount importance in various applications such as land monitoring, land use planning, hydrological modeling or natural resource management. In the last decade, geographic object-based image analysis (GEOBIA) has become an true alternative to pixel-based image analysis.

GEOBIA is an image processing method where groups of spatially adjacent pixels are classified as if they were behaving as a whole unit. This approach raises concerns about the way subsequent validation studies must be conducted. Indeed, classical point-based sampling strategies based on the spatial distribution of sample points (using systematic, probabilistic or stratified probabilistic sampling) do not rely on the same concept of objects and may prove to be less appropriate than methods explicitly built on the concept of objects used for the classification step.

Congalton and Green (2009) suggested that geographic database produced by GEOBIA would preferably be assessed by replacing the standard sampling units by objects. However, there is presently no state-of-the-art method to analyse the results of such object-based sampling. In a review of 20 recent papers on Geographic Object-Based Image Analysis (where the quality assessment protocol was described), point-based sampling was still used in the majority of the cases (60%) while the object-based sampling was analyzed in two different ways : giving the same weigh to the samples or weighing the samples based on their area.

In this study, we will i) demonstrate that the results of object-based sampling may evaluate two different sets of quality indices (map quality and classifier accuracy) and ii) propose a new method that reduces the uncertainty on the estimates of the map accuracy. It is worth noting that different indices exist to assess map accuracy, but that the focus of our study is on the estimation of the overall accuracy (OA), that is the proportion of correctly classified items.

II. METHOD

A. *Overview of the existing methods*

The main characteristics of the three methods found in a review of 20 recent GEOBIA papers are listed in table (1). For the sake of simplicity, we will refer to "points" in the case of point-based sampling and to "objects" in the case of object-based sampling. In practice, "points" can be pixels or clusters of pixels on the image, while "objects" are either produced by image segmentation algorithms (image-regions or image-objects) or by visual photo-interpretation (vector polygons). In terms of sampling result analysis, the important difference between "points" and "objects" is the variable size of the latter.

TABLE I. OVERALL ACCURACY FROM REVIEWED PAPERS

Method	Characteristics	
	Sampling unit	*Equation*
Point-based (PB)	Fixed area	$\Sigma C_i / n$
Simple object-based (SOB)	Variable area	$\Sigma C_i / n$
Weighed object-based (WOB)	Variable area	$\Sigma C_i S_i / \Sigma S_i$

where C_i is a binary variable that indicates if the sampling unit is correctly classified (0 if not correct), n is the number of sampling units and S_i is the area of the object.

It can be shown that these methods do not refer to the same set of indices: PB and WOB both estimate the OA of the map while SOB estimates the OA of the classifier. On average, these two quantities are equal, but this is not necessarily the case for a specific map, especially when the variability of the objects area is high.

B. New estimator of the overall accuracy

In a probabilistic framework, let us consider S_i and C_i as random variables, with the following hypotheses:

- the C_i's are mutually independent and identically distributed with a Bernouilli distribution characterized by a same parameter p, the probability for an object to be correctly classified. The expectations and variances are thus $E[C_i]=p$ and $Var[C_i]=p(1-p)$. It is worth noting that the sampling units are randomly selected from an object list. The sample is therefore not affected by spatial autocorrelation;
- the C_i's and S_i's are mutually independent, i.e. the misclassification does not depend on the object size.
- the S_i's are identically distributed; they obey a distribution with expectations $E[S_i]=\mu$ and variances $Var[S_i]=\sigma^2$ where μ and σ^2 are known (or at least reasonably well estimated).
- each case to be classified belongs fully to one of the classes in an exhaustively defined set of discrete and mutually exclusive classes, that is C_i value can be unambiguously defined for any given object. This hypothesis, often implicitly assumed for pixels (Foody, 2002) is more contentious in GEOBIA and is discussed later.

Let us focus on the OA of the map, denoted here as OA for convenience, defined the percentage of a map (in surface) which is correctly classified (Stehman and Czaplewski, 1998) given by (1):

$$OA = \sum_{i=1}^{N} C_i S_i / \sum_{i=1}^{N} S_i \tag{1}$$

where ΣS_i is the total surface of the map, which is considered as a constant in this study because objects are usually delimited on a fixed extent (for instance, a satellite image or an administrative boundary).

Based on theoretical developments (not shown here), a new predictor of the OA was derived, which minimizes the variance of the prediction under the above-mentioned hypotheses for a sampling without replacement in the object list.

$$\hat{OA} = \left\{ \sum_{i=1}^{n} C_i S_i + \sum_{i=1}^{n} \frac{C_i}{n} \sum_{i=n+1}^{N} S_i \right\} / \sum_{i=1}^{N} S_i \tag{2}$$

Equation (2) accounts for the objects area in the first term and for the classifier accuracy in the second term. It is worth noting the two extreme cases with regards to n and N: When n is close to N, the second term is neglected and the variance of the OA predictor tends to zero. When $n \ll N$, the best predictor of the OA is the classifier accuracy estimate.

C. Synthetic maps for empirical assessment

It can be shown that the parameters that affect the variance of the proposed overall accuracy predictor are the number of objects in the map (N), the coefficient of variation of the distribution of the area of the sampling units (cv), the probability to correctly classify an object (po) and the number of sampling unit (n). In practice, N and cv depend on the segmentation results and p on the classifier of the objects. The number of sampling n is chosen by the user depending on his requirements in terms of precision and his constraints (cost, availability, time, etc). The behavior of the reviewed and the proposed methods are tested here with a set of 300 simulated maps.

A Quickbird image of a fragmented rural landscape in Belgium was segmented using the multiresolution segmentation implemented in the Definiens software (Baatz and Schäpe, 2000) to derive a realistic distribution of the area of image-segments (figure (1)). Segmentation results with a total number of objects of 1 000 and 10 000 was then simulated by randomly selecting the size of these objects from the whole distribution of sizes.

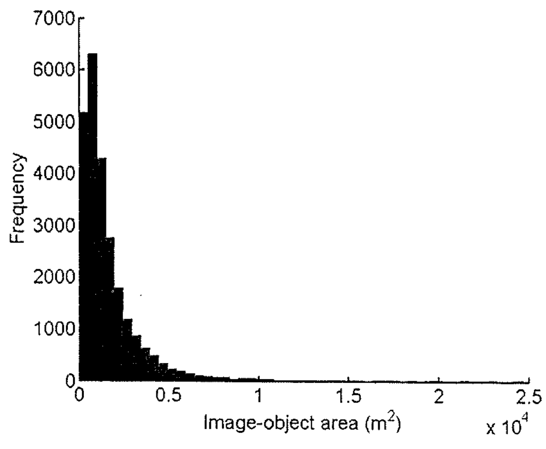

Figure 1. Frequency of objects area in the test zone

In order to test the good match between the theoretical values and the simulated results, the simulated objects were randomly labeled as "correct" or "incorrect" based on given object-based classification accuracy and the true OA was compared with its estimated values. Finally, each simulated map was sampled using n sampling units (points or objects), with n ranging from 1 to N. This step was repeated 100 times in order to derive the bias and the standard deviation of the methods.

III. RESULTS

Figure (2) corroborate the theoretical results and further show that the proposed method is at least as efficient as the best of the other methods in any cases. Indeed, the curve of the proposed method always lies below the curves of the other methods. This is particularly visible for extreme values : i) For large n values, object-based sampling outperforms the pixel-based sampling because it is performed without replacement and is thus exhaustive when n reaches N. The two other object-based formula yielded significantly larger RMSE than the proposed method. ii) For small sample sizes, the proposed method is more efficient than the area-based weighting because of its smaller variance.

The difference between p and OA is highlighted here as using the classifier accuracy with objects when $cv > 0$ leads to a biased estimate of the OA. This is rendered by a large RMSE even with exhaustive sampling. On the other hand, this bias was not significant for the other methods.

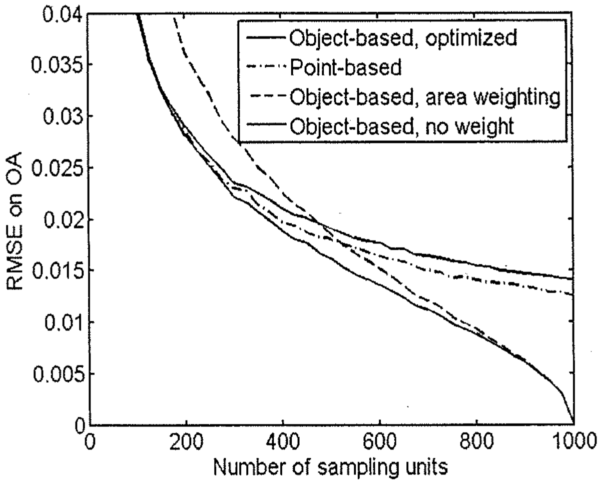

Figure 2. RMSE of the overall accuracy predictors

IV. DISCUSSION

The most obvious difference between object and point-based sampling is the variable area of the sampling units. As a major consequence, the classification accuracy (p) of the objects is no longer equivalent to the overall accuracy (OA) of a map produced with this classifier, to the contrary of point-based samples. With the same sample, the value of p can be estimated by dividing the number of correctly classified objects by the size of the sample, and the value of OA using the predictor proposed in our study. Similarly, object-based sampling could yield two sets of producer accuracy, user accuracy and confusion matrix, i) one with the number of objects in each case to evaluate the performance of the classifier and ii) one with the sum of the area of the objects to evaluate the proportion of the map that is correctly classified. However, it was shown in this study that the area weighed formula was not optimal and that it could be replaced by a new predictor (equation (2)), also unbiased but with a smaller variance.

A. *Analysis robustness issues*

The proposed overall accuracy estimation was derived from sound statistical bases with few assumptions. Its robustness was assessed based on realistic simulated maps to define its scope of application. A similar experiment to the one presented in this study proved that the identical distribution of the C_i was not required, i.e. the proposed method is still relevant when the classification accuracy of an image-segment depends on its class.

On the other hand, the assumption that the C_i are independently distributed has not been tested. However, it is more likely to be the case in an object-based sampling than in a point based sampling because i) image-segments, which are groups of adjacent pixel with the same labels, are sampled without replacement in object-based accuracy assessment but could be sampled several times with point-based simple random sampling; ii) spatial patterns of mixed image-segments are less likely to occur than spatial patterns of mixed pixels, which are most of the time located along the boundaries between two land covers (Foody, 2002).

However, the proposed object-based method loses part of its efficiency when p is correlated with the object size. In practice, this can be the case for two different reasons. First, the choice and the tuning of the segmentation algorithm can lead to a difference of reliability that is function of the size, for instance because larger image-segments have more reliable characteristics or, on the opposite, because larger image-segment are more likely to be under-segmented. Second, the size and the class of an object can be correlated, e.g. built up areas are smaller than other objects, so that the size effect could be indirectly induced by the land cover classes. It is therefore strongly recommended to test for the correlation between C_i and S_i. Further work is necessary to remove the bias induced by this correlation.

B. *Response design issues*

At this point, we assumed that each image-segment could be unambiguously validated (C_i *in* $\{0,1\}$) by comparing its label with a reference. Geographic object-based image analysis allows the producer to build a legend that accounts for homogeneous image-segment labeling (Tiede *et al.*, 2008). However, this requires an object-oriented response design based on object-oriented classification systems accounting for aggregation and association, for instance the LCCS (Di Gregorio and Jansen, 2000). In this framework, it is a particularity of GEOBIA to better describe aggregation than isolated pixels (Lang, 2008). Furthermore, minimum mapping units can be formally accounted for with object sampling units by using class associations. On the other hand, the internal consistency of image-segment is an objective of image segmentation, but under-segmentation creates artificial associations. As a consequence for validation, the legend should include a comprehensive set of mutually exclusive classes that can handle artificial associations, as proposed by the LCCS, but this reduce the map utility. An alternative solution could be to use a fuzzy evaluation similar to the approach proposed by Foody (1996) for coarse resolution pixels.

Anyway, the main practical issue behind object-based validation is the use of an image-segment as a sampling unit.

In practice, the reference dataset is primarily extracted by ground survey or photo-interpretation.

- Photo-interpretation can gain in consistency when object-boundaries are overlaid on the fine resolution images due to the scale information provided to the photo-interpreter. This is particularly useful in case of association or composition or when the shape plays a role in the interpretation. In a small experiment with ten photo-interpreters, the consistency of the object-based interpretation was larger than for point-based validation.
- For ground surveys, it is important to keep in mind that the objects must be validated as a whole in order to be consistent with the legend. A single location near the object center could be enough if the image-segments are coherent, but more work is necessary in case the segmentation fails to isolate the spatial regions. In any cases, the validation of very large objects (more than a few percents of the map) should be avoided. It is indeed difficult to establish a reference for theses objects and they are the most likely to consist in artificial associations. Furthermore, these huge objects strongly contribute to the variance of the object size distribution and therefore generate large *cv* value. If the large image-segment are produced by a multiscale segmentation algorithm, lower level image-segments should be used for the validation instead these large image-segments. Otherwise, an arbitrary splitting of the largest objects, for instance based on a regular grid, is recommended.

C. *Geolocation isues*

Object-based thematic accuracy assessment only accounts for the accuracy of the class labels assigned to the image-derived segments and not how well the spatial characteristics of objects are represented (Stow *et al.*, 2008). This is both an advantage and an inconvenience of the object-based validation to put aside the spatial accuracy, which is a main issue in point-based accuracy assessment (Strahler *et al.*, 2006). Indeed, the sampling location inside each object can be arbitrarily chosen because objects are homogeneously labeled (George, 1986). As image-segments are generally composed of a large number of pixels, sampling locations away from their boundaries helps then to avoid geolocation errors. On the other hand, the assessment of delineation and misregistration errors is a complex task (Radoux and Defourny, 2007) that is most of the time neglected.

V. CONCLUSION

Object-based validation can be used for thematic accuracy assessment of geographic object-based image analysis. An new object-based estimator of the overall accuracy was proposed in this study, which requires less sampling units than the point-based and the other object-based methods for a given precision. However, this study is only a first step toward quantitative object-based thematic accuracy assessment. Further work is necessary to extend the results of this work in order to improve the estimate of the other quality indices and to develop advanced solutions such as object-based stratified sampling or fuzzy object-based response design.

ACKNOWLEDGEMENTS

The authors thank E. Beriaux, V. Bombaerts, S. Bontemps, G. Duveiller, C. Ernst, S. Klimowski, J. Lhoir, V. Obsomer, and A. Verhegghen for their participation in the photo-interpretation experiment.

REFERENCES

Baatz, M. and Schäpe, A. (2000). Multiresolution Segmentation – an optimization approach for high quality multi-scale image segmentation. In: *J. Strobl, T. Blaschke and G. Griesebner, eds. Angewandte Geographische Informationsverarbeitung XII* (pp.12-23). Heidelberg: Wichmann-Verlag.

Congalton, R.G. and Green (2009). *Assessing the accuracy of remotely sensed data: Principles and practices*, Second edition. Boca Raton: CRC Press.

Di Gregorio, A. and 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.

Foody, G. (1996) Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*. 17(7), 1317-1340.

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

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

Lang, S. (2008) Object-based image analysis for remote sensing applications: modeling reality - dealing with complexity. In: *Object-based image analysis*. (pp. 3-28). Berlin: Springer.

Radoux, J. and 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.

Stehman, S. and Czaplewski, R., (1998). Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*. 64, 331-344.

Stow, D., Hamada, Y., Coulter, L. and Anguelova, Z. (2008) Monitoring shrubland habitat changes through object-based change identification with airborne multispectral imagery. *Remote Sensing of Environment*. 112 (3), 1051-1061.

Strahler *et al*, (2006) Global land cover validation: recommendations for evaluation and accuracy assessment of global land cover maps. *GOFC-GOLD Report*. 25, 1-52.