Unsupervised Quantification of Under- and Over-Segmentation for Object Based Remote Sensing Image Analysis

Andrés Troya-Galvis, Pierre Gançarski, Nicolas Passat, Laure Berti-Équille

***Abstract*—Object Based Image Analysis (OBIA) has been widely adopted as a common paradigm to deal with very high resolution remote sensing images. Nevertheless, OBIA methods strongly depend on the results of image segmentation. Many segmentation quality metrics have been proposed. Supervised metrics give accurate quality estimation but require a ground-truth segmentation as reference. Unsupervised metrics only make use of intrinsic image and segment properties; yet most of them strongly depend on the application and do not deal well with the variability of objects in remote sensing images. Furthermore, the few metrics developed in a remote sensing context mainly focus on global evaluation. In this article we propose a novel unsupervised metric which evaluates local quality (per segment) by analysing segment neighbourhood, thus quantifying under- and over-segmentation given a certain homogeneity criterion. Additionally, we propose two variants of this metric, for estimating global quality of remote sensing image segmentation by the aggregation of local quality scores. Finally, we analyse the behaviour of the proposed metrics and validate their applicability for finding segmentation results having good trade-off between both kinds of errors.**

***Index Terms*—Image segmentation, Quality control, Object oriented methods, Image region analysis.**

I. Introduction

Automatic interpretation of remote sensing images is a challenging but mandatory task, since manual processing tends to become infeasible as images size and acquisition frequency are rapidly increasing. The purpose is to determine the nature of the objects being represented in the image. A possible solution is to analyse the image pixelwise and apply machine learning methods [1] in order to infer the thematic class of pixels in function of their radiometric values. The resulting labelled image can be further used by geographers in a wide range of applications such as urban planning [2], deforestation tracking [3], or disaster risk management [4]. However, pixel-oriented methods have reached their limits with the development of High and Very High Spatial Resolution (VHSR) remote sensing images [5]. In fact, at VHSR, each pixel represents a region ranging from 0.5m to 2m, which means that the complexity and the variety of identifiable objects increase considerably. Thus, a single pixel generally does not represent a single geographic object but rather a part of it. Moreover, at this level of detail, different thematic classes may share similar spectral signatures, or a complex object may contain pixels with different spectral properties. Object-Based Image Analysis (OBIA) [5] approaches try to overcome these difficulties by grouping pixels into higher level objects, called regions or segments. These regions allow the computation of more informative features such as shape or texture, which can be used to better identify and describe structures of interest in the studied area.

Generally, image segmentation methods are employed as a first step in OBIA in order to construct the segments which are further used to perform analysis, e.g., object extraction or classification [6]. There exists a wide spectrum of segmentation approaches, and thousands of *ad hoc* variants devoted to specific applications. We refer the reader to [7] for a complete survey on this topic. In the context of remote sensing imaging, the most popular approaches are mainly those relying on region-based and spectral homogeneity paradigms: mean-shift [8], region-growing [9], split-and-merge [10], watershed [11], or hierarchical strategies [12].

By contrast with the huge literature on image segmentation methods, fewer efforts have been devoted to segmentation quality evaluation. A perfect segmentation should provide a partition of the image that induces a one-to-one mapping between each segment in the image and each object in the studied area. From this definition, segmentation errors can be characterized as over-segmentation which happens whenever many segments map to a single object (i.e., the corresponding segments are too small); and under-segmentation which happens when a single segment maps to many objects (i.e., the corresponding segments are too large). Mismatching between segments and objects may lead to erroneous or irrelevant computed features. In such cases, machine learning methods may fail in predicting the segments class. Thus, image segmentation is a critical step, as every error is propagated throughout the whole process and may end up with unexpected or misleading results [3], [13].

It is then essential to be able to quantify image segmentation quality in order to limit, or at least, to be aware of the number of errors being propagated through later stages of the process. However, image segmentation is an ill-posed problem and the instantiation of segmentation algorithms is often dependent on the applicative context, and so are the existing quality metrics.

Zhang et al. classify segmentation evaluation methods given different criteria [14]. On the one hand, subjective methods

Andrés Troya-Galvis and Pierre Gançarski are affiliated to the ICube Laboratory UMR CNRS 7357 - University of Strasbourg, France (e-mail: troyagalvis@unistra.fr, gancarski@unistra.fr).

Nicolas Passat is affiliated to the Université de Reims Champagne-Ardenne, CReSTIC, Reims, France (e-mail: nicolas.passat@univ-reims.fr).

Laure Berti-Équille is affiliated to the Qatar Computing Research Institute (QCRI) (e-mail: lberti@qf.org.qa)

basically consist on visual examination of segmentation results made by humans. Although this kind of evaluation implies a long and tedious task, it is the only way to ensure that the results actually correspond to user's expectations. In remote sensing, such an approach is globally intractable, since images (in particular HSR and VHSR) are huge. Indeed, segmentation in this context does not consist of computing a partial partition – focusing on specific structures – but most often a global partition. Thus, the image is divided into a huge set of segments that cannot be manually processed by a human expert. On the other hand, objective methods aim at quantifying the segmentation accuracy, based on specific criteria associated to metrics. They can be divided into two types of methods: 1) system level evaluation methods, which evaluate segmentation impact in the final performance of a whole system or application; and 2) direct evaluation methods. These latter deal either with algorithmic efficiency (time and space cost), behaviour or segmentation results. Objective direct evaluation approaches are typically divided into supervised and unsupervised methods.

Supervised evaluation methods basically rely on distance measures between segmentations. They compare results with one or more ground-truth segmentations, usually obtained manually [15]. A complete study on supervised segmentation evaluation methods is out of the scope of this article; the interested reader may refer to [16] and [17] for such studies. We focus our work solely on unsupervised evaluation.

Unsupervised evaluation methods rely merely on intrinsic properties which can be computed directly from the segments; they have been less explored than supervised metrics. Indeed, the definition of unsupervised metrics represents a real challenge, as they should somehow model the human notion of quality, which is not a trivial task. In fact, segmentation quality is a subjective concept which may vary from one application to another, and even from one individual to another. In summary, subjective methods rely on visual inspection, they are necessary to make sure that results actually correspond to user's expectations. Objective methods such as supervised and unsupervised metrics provide a numerical score quantifying segmentation quality, thus, they allow the automatic finding of the best segmentations or the best parameter set for a given segmentation algorithm.

In general, unsupervised metrics assume that each segment should be homogeneous given certain criteria. They usually try to maximize intra-segment homogeneity while minimizing inter-segment similarity [3]. Thus, most of existing metrics consist basically in combining two terms: one characterizing over-segmentation and the second characterizing under-segmentation [14]. One can find methods relying on homogeneity criteria such as region contrast [18], pixels entropy [19], [20], texture [21], or the mean and variance of pixel values [22]. Others are based on measurements in particular color spaces, thus trying to model the human visual system by taking into account perceptual color difference [23]. For a detailed study of unsupervised quality metrics, the reader is referred to [22].

In remote sensing image segmentation, the purpose is to extract a wide variety of objects having different sizes and shapes or – more generally – to define a partition of the whole image into regions of homogeneous and relevant semantics. Classical unsupervised metrics rely mainly on the number of segments or consider the contrast among (few) regions [14]. Thus, they are weakly suitable for remote sensing image segmentation, as the number of object occurrences for a given class may vary from a few tens to many hundreds.

A few metrics have been proposed within the context of remote sensing image analysis. Zhang et al. in [24], proposed a metric considering the size of the image, the number of segments, as well as segment and band mean values. Corcoran et al. in [17], proposed a metric which takes into account the spatial domain; they argue that human vision performs segmentation in the spatial domain by seeking contrast across object boundaries. Johnson [25], proposed a metric which characterizes intra-segment homogeneity as a weighted sum of feature variances, and Moran's index to characterize inter-segment heterogeneity.

The values provided by these metrics are not bounded and may vary considerably from an image to another, making it complex to fully understand their meaning. Moreover, they mainly assess quality in a global manner, i.e., by combining intra- and inter-segment similarity scores which are computed globally for the whole image. Thus the local information carried by individual segments is lost. Reasonably, segmentation evaluation metrics for OBIA applications should remain within the object-based paradigm. Then, the quality of a segmentation should be regarded as a combination of its composing segments quality.

To address these challenges, this article presents the following contributions: 1) a novel metric which evaluates the quality of each segment individually as a function of its spatial neighbourhood and a given homogeneity criterion; 2) two variants for quantifying global quality as an aggregation of local quality scores, by considering both over- and under-segmentation errors; and 3) experiments showing that the proposed approach is more robust when dealing with VHSR images, compared to two state-of-the-art unsupervised evaluation metrics. Furthermore, the compelling feature of our approach is to enable the refinement of post-processing by keeping track of local quality estimates, thus enhancing the succeeding classification process.

The rest of this article is as follows. In Section II, the application context of our work is presented, the choice for unsupervised evaluation as well as the experimental validation methodology are justified. In Section III, the proposed unsupervised quality metric for local and global evaluation of remote sensing image segmentation is presented. In Section IV, we asses the behaviour of this metric and validate its applicability in the context of remote sensing. Section V presents our research perspectives.

II. Context

The VHSR dataset employed in our work is a Pleiades pansharpened image[1] with 60cm spatial resolution and four spectral bands (R, G, B, NIR). It represents Strasbourg's urban

[1] VHSR datasets were provided by the LIVE laboratory UMR CNRS 7263.

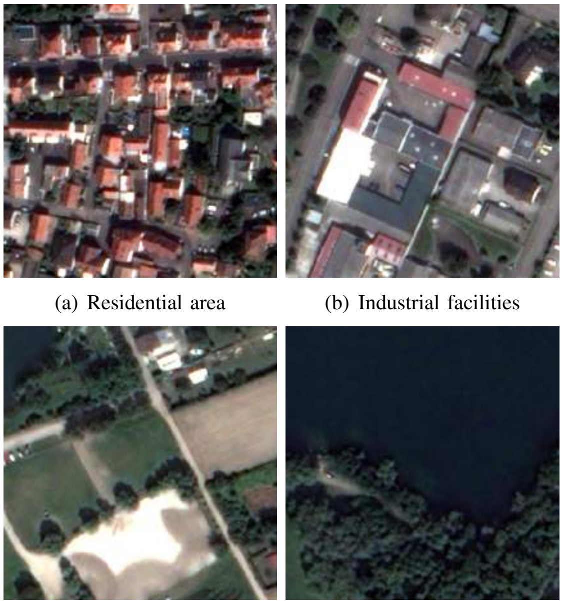

(a) Residential area
(b) Industrial facilities
(c) Agricultural parcels
(d) Forest and water zone

Fig. 1. Extracts from different zones of the Strasbourg image used in our study.

community as well as its surroundings, and is made up of approximately 10000×10000 pixels, which is a considerably huge amount of data to be processed. Figure 1 shows different extracts of the image containing objects of very different nature, size, and shape, such as residential buildings, industrial buildings, different types of vegetation and water bodies.

As stated above, supervised evaluation methods require a ground-truth segmentation to be used as a reference. However, in remote sensing, it is often impossible to obtain such a complete ground-truth. Indeed, manual segmentation feasibility is limited by the huge dimensionality of images. Moreover, vectorial data from topographic databases are taken *in situ*. Thus, due to the satellite position during image acquisition, the further geometric corrections, and outdated topographic data, they usually do not match the objects in the image, even by considering subsequent alignment procedures; this phenomenon can be verified in Figure 2. Moreover, if we want to implement fully automated VHSR OBIA methods it is essential to have robust and accurate unsupervised quality metrics. Indeed, the need of ground-truth or human validation at intermediate steps represents a bottleneck which motivates our approach with unsupervised evaluation.

Evaluation metrics described in Section I focus on global evaluation in order to find an optimal parameter set for a given segmentation algorithm [26], [27]. Nevertheless, to our knowledge there is no segmentation algorithm for which a given parameter set allows the correct segmentation of every class of objects. Indeed, the ideal segment sizes may vary considerably depending on the semantic level of objects of interest. Furthermore, at a given semantic level, objects may still substantially vary in size and shape. Thus, unsupervised metrics should be easily adapted in order to enable the selection of the best segmentations at different scales. As stated in [17], good unsupervised metrics should work with features which model correctly the human visual system, and should take into consideration spatiality. Our working hypothesis is that, local (i.e., per segment) evaluation should also be considered in OBIA applications, as it may allow a better understanding while carrying out further analysis, thus enhancing decision-making processes.

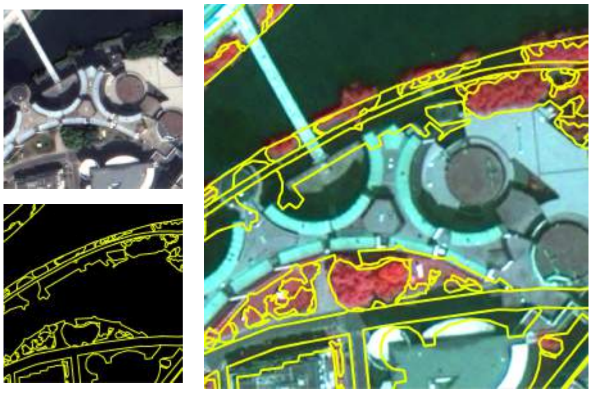


Fig. 2. Misalignment between vectorial topographic data and the actual image. Left: The original image as well as the corresponding vectorial data. Right: False color image with the superimposed vectorial data.

III. Under- & Over-Segmentation Aware (UOA) Metric

As stated above, the assumptions of OBIA approaches should lead us to rely on metrics that assess segmentation quality by accurately considering each segment, i.e., in a local way. In this section, we present our main contribution, namely an *unsupervised local* metric that allows the quantification of under- and over-segmentation for each segment. We also propose two dual ways of aggregating the local quality results into a global quality measure. Since our metric integrates – as meta-parameter – an homogeneity index, it is highly adaptable, and can be used in particular either to find segmentations at different scales, or to assess the relevance of such indexes in specific applications.

A. Local Evaluation

Let $S = \{R_i \mid 0 \leq i < M\}$ be a partition of an image space composed of M segments, R_i. Let $\mathcal{N}(R_i)$ be the set of segments in the neighbourhood of R_i. Let $H(R_i)$ be a function returning a score in $[0,1]$ characterizing the homogeneity of the segment R_i, 0 meaning complete homogeneity and 1 meaning complete heterogeneity. Finally, let δ be a threshold on the values of $H(R_i)$. We define the local evaluation function ϕ_δ as follows:

$$\phi_\delta(R_i) = \begin{cases} -1 & \textbf{if } H(R_i) > \delta \\ 1 & \textbf{if } H(R_i) \leq \delta \textbf{ and} \\ & \exists R_j \in \mathcal{N}(R_i) \mid H(R_i \bigcup R_j) \leq \delta \\ 0 & \textbf{otherwise} \end{cases} \quad (1)$$

Intuitively, we consider that a segment R_i is of good quality if it is well separated from its neighbours at a given scale; this scale is determined by the homogeneity threshold δ. Indeed, the larger a segment, the more it is likely to be heterogeneous (i.e., higher H value). Thus, as δ increases, larger segments

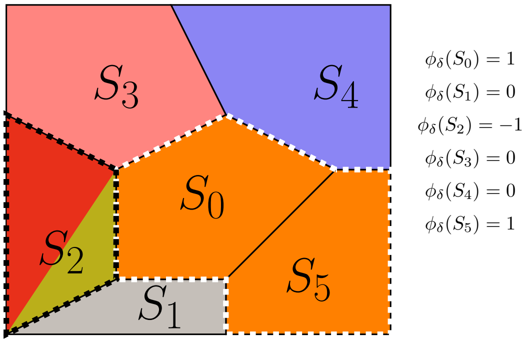


Fig. 3. Toy-example illustrating the local evaluation approach. The white dashed line shows an example of over-segmentation. The black dashed line shows an example of under-segmentation.

have more chances to be considered as under-segmentation, and smaller segments to be considered as over-segmentation. There are then three possible cases:

- R_i is too heterogeneous (i.e., $H(R_i) > \delta$), it can be seen as an under-segmented region of the image. In such case, it is penalized with a negative weight.
- R_i is homogeneous enough (i.e., $H(R_i) \leq \delta$), but there is at least one neighbouring segment such that their union is also homogeneous (i.e., $\exists R_j \in \mathcal{N}(R_i) \mid H(R_i \bigcup R_j) \leq \delta$), it can be seen as an over-segmented region of the image. In such case, it is penalized with a positive weight.
- R_i is homogeneous and well isolated from its neighbouring segments. No penalization is then applied.

Figure 3 illustrates this local evaluation approach. Consider a trivial homogeneity index H which associates 0 to a completely homogeneous region (i.e., a region composed of a unique color) and 1 otherwise; and let $\delta = 0.5$. The segment S_2 (black dashed contour) is not homogeneous, then $H(S_2) = 1 > \delta$, so it is considered as under-segmented. Segments S_0 and S_5 are both considered as over-segmentation, indeed, we have $H(S_0) = 0 < \delta$, but also $H(S_0 \bigcup S_5) = 0 < \delta$ (the same reasoning holds for S_5). Practically, these two segments may be merged together to be correctly segmented. Finally, segments S_1, S_3, and S_4 are homogeneous and well separated from their neighbours.

B. Global Evaluation

The global quality of a segmentation should depend on the qualities of its segments. We propose and compare hereafter two global variants of our proposed metric based on two ways of aggregating local quality measures.

1) UOA_Σ: The first aggregation function is similar to the one used for classification accuracy assessment in [28]. It is defined as a weighted sum of the local quality scores. It gives a direct estimate on the amount of over- or under-segmentation. However, it may fail if both errors are evenly present in the segmentation, as penalization weights will tend to cancel out. It is defined as follows:

$$UOA_\Sigma = \sum_i \omega(R_i).\phi_\delta(R_i) \tag{2}$$

where ω is a weighting function such that for all R_i we have $\omega(R_i) \geq 0$ and $\sum_i \omega(R_i) = 1$. In our experiments, we chose $\omega(R_i) = \frac{N_i}{N}$ where N_i and N denote the number of pixels in segment R_i and the number of pixels in the whole image, respectively. Each segment is then weighted proportionally to its size in the image.

2) UOA_{L_2}: The second aggregating function we define, first quantifies under- and over-segmentation independently and considers them as a 2-dimensional vector. The final score is then the L_2 norm of this vector. Thus, it aims at minimizing both types of errors simultaneously at the cost of losing the informative behaviour of UOA_Σ. It is defined as follows:

$$\Psi = \sum_{R_i \mid \phi_\delta(R_i) = -1} \omega(R_i)$$

$$\Theta = \sum_{R_i \mid \phi_\delta(R_i) = 1} \omega(R_i)$$

$$UOA_{L_2} = \sqrt{\Psi^2 + \Theta^2} \tag{3}$$

where Ψ and Θ represent the under- and over-segmentation rates respectively.

C. UOA Metric Properties

We now discuss some properties of the proposed metrics:

- Both variants of our metric are bounded, which is a useful property lacking in many existing metrics. Indeed, UOA_Σ varies from -1 (all of the segments are considered as under-segmentation) to 1 (all of the segments are considered as over-segmentation); while UOA_{L_2} varies from 0 (absence of errors) to 1 (complete over and/or under-segmentation).
- UOA_Σ is more informative when reporting very over-segmented (high positive value) or very under-segmented (high negative value) results; this may lead to undesired results if the image is half over-segmented and half under-segmented, as it flattens down these two opposite components. On the other hand, UOA_{L_2} is capable of finding the segmentation with the less errors, but it is not possible to determine the role of over and under-segmentation.
- The choice of the meta-parameter H requires a certain level of expertise. In fact, it is important to employ an homogeneity index which is relevant to the application.
- The choice of δ is crucial as well as it depends on H and has to be adapted to the desired scale.
- The algorithmic complexity of our global metrics is in the best case $\mathcal{O}(M)$ where M is the number of segments in the segmentation if all of the segments are under-segmented; in average it can be approximated by $\mathcal{O}(MK)$ with K the average number of neighbours by segment, as we have to check neighbouring segments when verifying over-segmentation. Though, K is considerably smaller than M and is generally bounded by a small value, so we can approximate the complexity to be linear with respect to M.

IV. Experimental Results

In this section, we analyse the behaviour of the proposed metric, and in particular, its sensitivity with respect to the parameter δ. To this end, we employed different parameter sets in order to generate 100 Mean Shift [29] segmentations of a 1024×1024 and a 3000×4000 pixels extracts, (cf. Figures 8(c) and 9(a)) of the Strasbourg image containing objects from different thematic classes as illustrated in Figure 1.

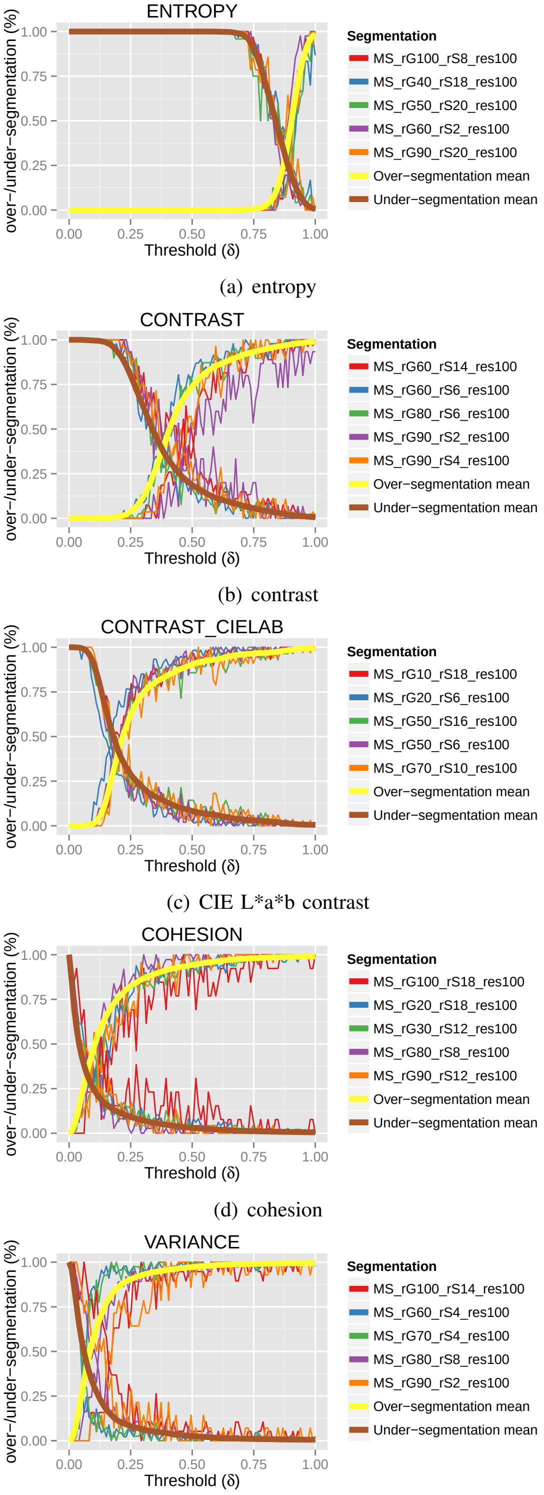
A. *UOA and H Parameter*

The definition of the UOA metric is generic and allows us to choose an arbitrary homogeneity criterion which can vary from one application to another. Indeed, H is a meta parameter of UOA so it is possible to use a large variety of existing homogeneity criteria such as the spectral angle [30], texture homogeneity measures based on gray level co-occurrence matrices [31], or even to define new homogeneity measures if the application requires so. We studied the behaviour of UOA with respect to 6 homogeneity indexes: entropy [19], contrast [18], CIE L*a*b contrast [23], cohesion [17] and variance [25], as well as the mean of these 5 indexes. All of these indexes show a strong correlation between their values and the segment size: the larger the segment, the higher the score.

We computed each homogeneity index H for every segment and we observed the variation of over- and under-segmentation with respect to δ. Figures 4 and 5 show this sensitivity analysis for the 6 homogeneity measures, each subfigure shows measurements for 5 segmentations chosen randomly. We remark that they all behave differently. Indeed, each index is sensitive to δ within different intervals. Entropy is very sensitive in $[0.7, 0.9]$ (Figure 4(a)). Contrast is very sensitive over $[0.2, 0.5]$, then it varies slowly over $]0.5, 0.9]$ (Figure 4(b)). Contrast in the CIE L*a*b color space is very sensitive over $[0.1, 0.3]$ (Figure 4(c)). Cohesion (Figure 4(d)) and variance (Figure 4(e)) are remarkably sensitive over $[0.0, 0.1]$, then they vary smoothly over $]0.1, 1.0]$. Although, the over-segmentation increases slightly slower with the cohesion index than with the variance index. Figure 5 shows the sensitivity of the mean of indexes. We observe that over-segmentation varies over the whole $[0, 1]$ interval, approaching the cumulative distribution function of a normal distribution; under-segmentation also decreases smoothly over the whole range of values. Additionally, Figure 5 shows the ratio of well isolated segments; the optimal value for the well isolation ratio corresponds to the cross-point between over- and under-segmentation, as one would expect. Remark that the best isolation ratio obtained is about 0.6 for a particular segmentation; in average it is around 0.4. This shows that even in the best scenario, there are still either over- or under-segmentation errors. Furthermore, we see that the behaviour of UOA may vary in function of the chosen H; more particularly δ has to be carefully chosen, as its optimal value depends on H.

B. *UOA and δ Parameter*

For the following experiments we chose the mean of indexes as it provides a quite regular behaviour, allowing us to focus



(a) entropy

(b) contrast

(c) CIE L*a*b contrast

(d) cohesion

(e) variance

Fig. 4. Sensitivity analysis of Ψ and Θ over δ for 5 different homogeneity indexes. Segmentation labels correspond to the Mean Shift parameters used: range resolution (rG), spatial resolution (rS) and minimum segment size (res). The mean lines were computed over all of the 100 segmentations.

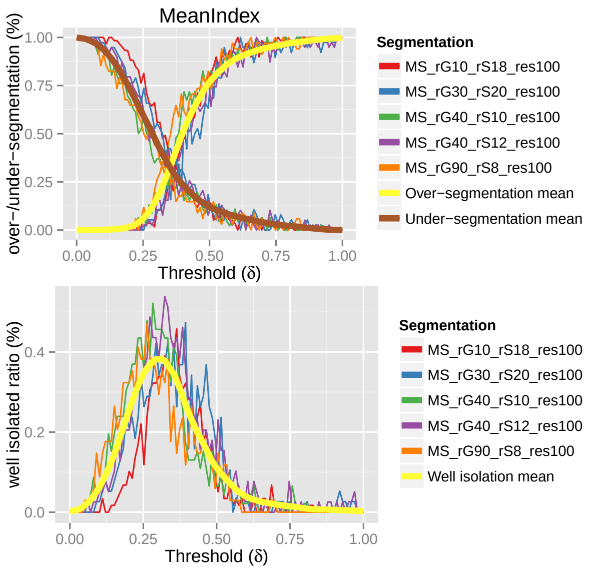


Fig. 5. Sensitivity analysis of Ψ and Θ as well as the well isolation ratio over δ for the mean of the 5 former homogeneity indexes.

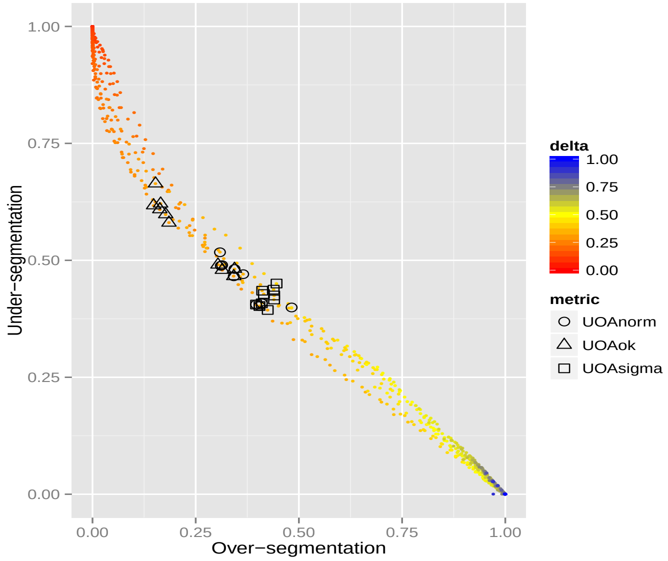


Fig. 6. Under- vs. over-segmentation with δ varying from 0 to 1, for 10 randomly chosen segmentations.

on the evolution of UOA with respect to δ, while avoiding H-based bias. We computed UOA_Σ and UOA_{L_2} for δ, varying from 0 to 1 by 0.01 steps, thus leading to 10 000 observations. We also kept track of under-segmentation (Ψ) and over-segmentation (Θ) rates, as well as the well isolated segments rate, namely:

$$UOA_{ok} = 1 - (\Theta + \Psi) \tag{4}$$

We observed the behaviour of UOA_Σ and UOA_{L_2} and $1 - (\Theta + \Psi)$ in function of δ. Figure 6 shows how under-segmentation varies in function of over-segmentation for 10 different segmentations selected randomly. Each data point in the plot corresponds to a given δ. We can observe a strong correlation between over- and under-segmentation as well as δ. Indeed, as δ increases, the over-segmentation score reported by UOA increases and under-segmentation score decreases. The shaped points in the plot correspond to optimal δ values found by UOA_{L_2} (UOAnorm), UOA_{ok} (UOAok), and UOA_Σ (UOAsigma); actual values for these 3 points are displayed in Table I. Recall that optimal values should minimize $|UOA_\Sigma|$, UOA_{L_2} and maximize UOA_{ok}. We can observe that UOA_Σ finds optimal δ values between 0.30 and 0.42; in average the under and over-segmentation rates are about 0.42. UOA_{L_2} aims at minimizing both components; it finds optimal δ values between 0.30 and 0.43; we observe that UOA_Σ is not always optimal when UOA_{L_2} is. These metrics find a good trade-off between the two types of errors. However, UOA_{L_2} keeps well isolated segment rates higher than UOA_Σ. This can be explained by the fact that UOA_Σ flattens some information as it averages indifferently both under- and over-segmentation. UOA_{ok} finds optimal δ values between 0.28 and 0.35. While the rate of well isolated segments is higher than for the other metrics (0.20 compared to 0.15 and 0.17), the under-segmentation rate is quite larger (around 0.56). Globally, we observe that our metrics are consistent; they succeed in finding an optimal δ which constitutes a good compromise between both kinds of errors.

TABLE I
BEST δ FOUND FOR 10 RANDOMLY CHOSEN SEGMENTATIONS

metric	aggregate	δ	UOA_Σ	Θ	Ψ	UOA_{L_2}	UOA_{ok}
UOA_Σ	Min.	0.300	**-0.023**	0.396	0.393	0.567	0.103
	Median	0.365	0.001	0.419	0.420	0.597	0.155
	Mean	0.360	0.002	0.422	0.420	0.596	0.157
	Max.	0.420	0.030	0.445	0.450	0.633	0.198
UOA_{L_2}	Min.	0.300	-0.208	0.308	0.399	**0.567**	0.118
	Median	0.345	-0.114	0.354	0.468	0.580	0.186
	Mean	0.348	-0.085	0.367	0.452	0.587	0.179
	Max.	0.430	0.083	0.482	0.517	0.626	0.200
UOA_{ok}	Min.	0.280	-0.511	0.148	0.465	0.573	0.174
	Median	0.310	-0.407	0.181	0.589	0.616	0.210
	Mean	0.314	-0.331	0.229	0.560	0.614	0.209
	Max.	0.350	-0.123	0.344	0.664	0.681	**0.234**

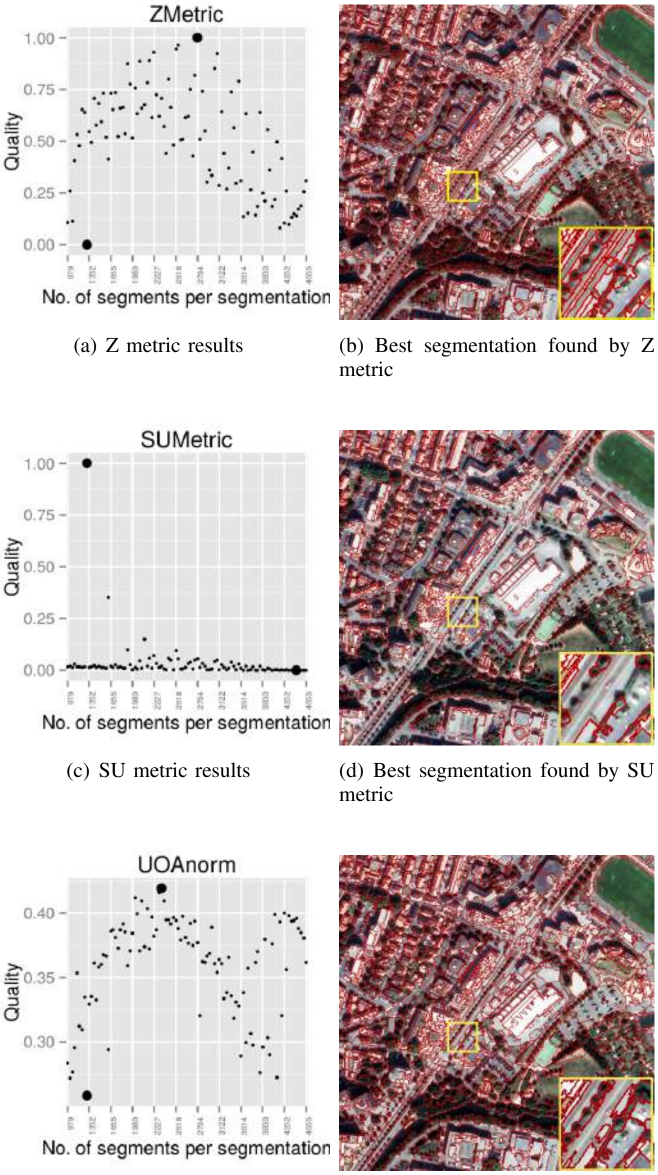
C. Validation

In order to validate the usability of the proposed metrics for remote sensing image analysis, we proceed in two steps. First, we compare UOA_Σ and UOA_{L_2} to two of the unsupervised metrics presented in Section I: the Z metric [24] and the SU metric [17]. Then, we show how the local evaluation approach could be used in further refinement procedures.

1) Comparative study: For the computation of UOA_Σ and UOA_{L_2} in this experiment, we fixed $\delta = 0.37$ as it was one of the optimal threshold values found in the previous experiment. Then, we evaluated and ranked the 100 Mean Shift segmentations.

In Figures 7(a), 7(c), 7(e) and 8(a), the x axis represents the 100 segmentations sorted by their number of segments, ranging from 979 to 4 555 segments. The y axis represents actual metric values. Remark that both the Z and the SU metrics are not bounded, so they were normalized between 0 and 1. Moreover, these data were adapted so that 0 represents low quality and 1 represents high quality.

Based on the results of this experiment, we make the following observations:



(a) Z metric results

(b) Best segmentation found by Z metric

(c) SU metric results

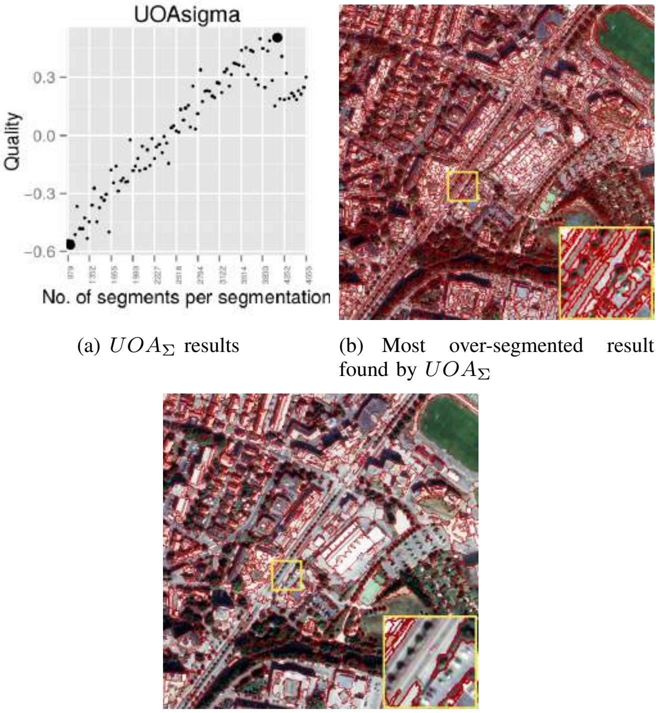
(d) Best segmentation found by SU metric

(e) UOA_{L_2} results

(f) Best segmentation found by UOA_{L_2}

Fig. 7. Best segmentation results found by Z, SU and UOA_{L_2} metrics. The scores of each metric are shown in function of the number of segments, the higher the score the better the quality. Thus, the thick points correspond to best and worst quality scores found. The images in the right correspond to the best segmentation found according to each metric.

- The SU metric presents some surprising outliers (Figure 7(c)); this may be explained by the fact that the feature set used for the computation of the metric is not explicitly reported in the original article. We used only radiometrical values for this purpose. A visual inspection of the best result found shows that it is indeed a case of under-segmentation. We conclude that the metric is perhaps not appropriated for radiometrical features alone in a remote sensing context.
- UOA_{L_2} and the Z metric find optimal segmentations which are quantitatively very close to each other: 2 291 and 2 781 segments respectively (Figures 7(e) and 7(a)). Nevertheless, the Z metric seems to be less consistent than the UOA_{L_2}, as it can give very different metric values to very close segmentations (e.g., 0.45 for a segmentation containing 2 694 segments and 0.80 for a segmentation with 2 736 segments).
- A visual inspection of the results reveals that both segmentations are quite good, but the segmentation found by UOA_{L_2} has slightly smoother boundaries (e.g., the stadium at the top right corner, or the road in the highlighted zone). Indeed, irregular boundaries are often composed by many little and irrelevant segments, UOA_{L_2} then results in a low quality score as these segments highly increase the over-segmentation rate.



(a) UOA_{Σ} results

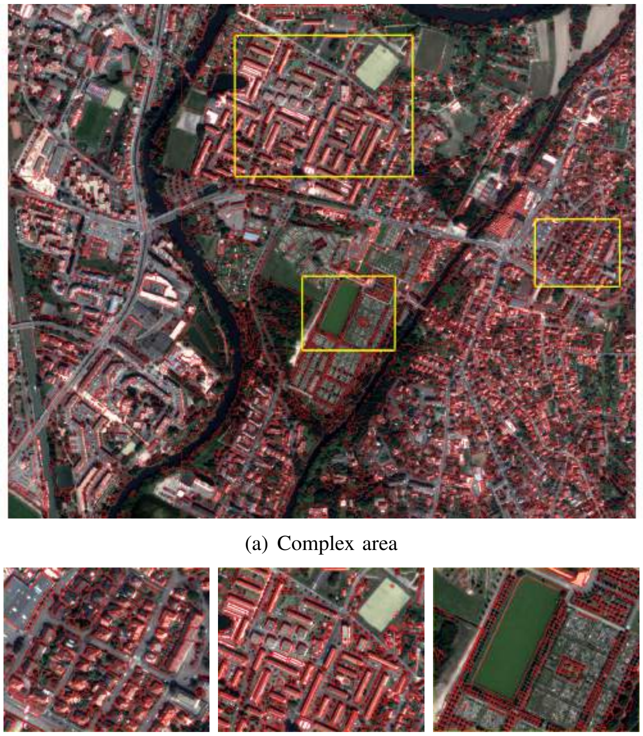
(b) Most over-segmented result found by UOA_{Σ}

(c) Most under-segmented result found by UOA_{Σ}

Fig. 8. The scores of UOA_{Σ} are shown in function of the number of segments, a negative score means a mostly under-segmented result, and a positive score means a mostly over-segmented result (see Eq. 2). Thus, the highlighted points correspond to the most under- and over-segmented result found by UOA_{Σ}.

Figure 8(a), shows the data resulting from the evaluation using UOA_{Σ}. Note that for this plot we kept the original metric values. In fact, one of the main advantages of UOA_{Σ} is its ability to report very over-segmented (high positive values) and under-segmented (high negative values) results. The plot shows how it increases linearly as the number of segments increases. Figures 8(b) and 8(c) show the boundary representation of the most over-segmented and the most-under-segmented result. A visual inspection lets us corroborate that UOA_{Σ} correctly identifies both kinds of erroneous segmentations.

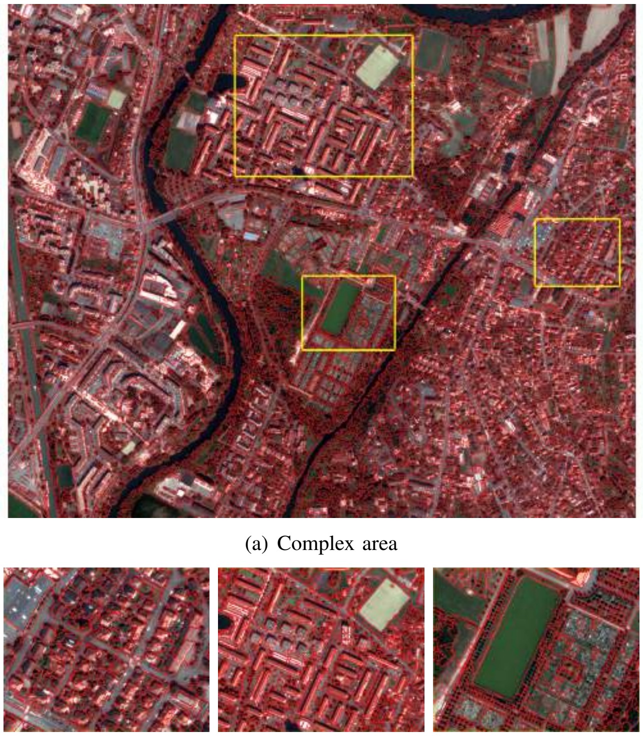
For further validation of our approach, we repeated the experiment with a 4000 × 3000 image extract, using the



(a) Complex area

(b) Urban area (c) Industrial area (d) Vegetation area

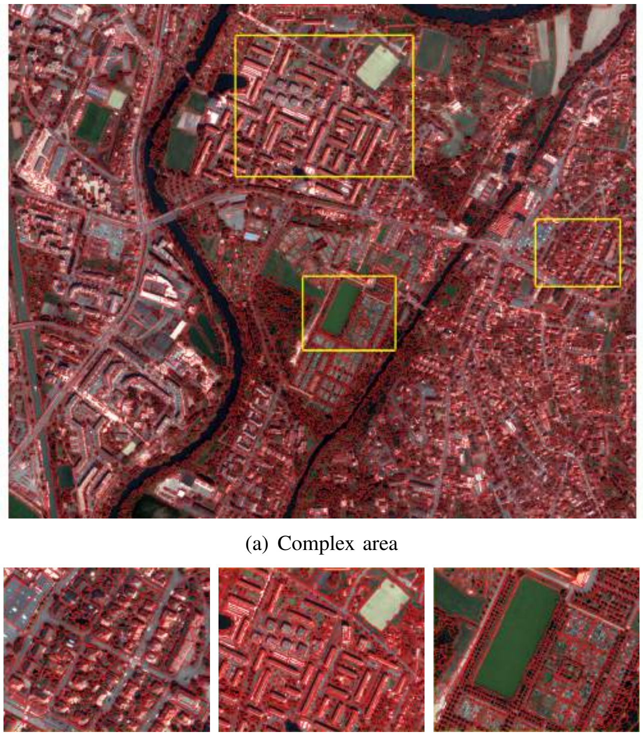
Fig. 9. SU Metric best result on a 3000 × 4000 extract containing multiple and complex objects. The highlighted zones correspond to an urban residential area (b), industrial facilities (c) as well as some vegetation (d).



(a) Complex area

(b) Urban area (c) Industrial area (d) Vegetation area

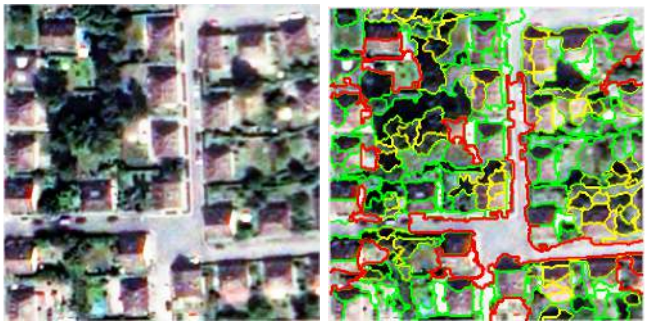
Fig. 10. Z Metric best result on a 3000 × 4000 extract containing multiple and complex objects. The highlighted zones correspond to an urban residential area (b), industrial facilities (c) as well as some vegetation (d).



(a) Complex area

(b) Urban area (c) Industrial area (d) Vegetation area

Fig. 11. UOA_{L_2} best result on a 3000×4000 extract (a) containing multiple and complex objects. The highlighted zones correspond to an urban residential area (b), industrial facilities (c) as well as some vegetation (d).



(a) Extract of a residential area (b) Local quality evaluation

Fig. 12. Visualisation of local quality estimates. Yellow segments correspond to over-segmentation, red segments correspond to under-segmentation and green segments correspond to well isolated segments.

same parameter set. The visual results are shown in figures 9, 10 and 11. The results are consistent with the previous experiment. The best segmentation found by the SU metric is globally under-segmented, even if some small structures such as buildings are well segmented. The result obtained by using the Z metric is good, but there are still some under-segmented regions. The best segmentation found by our metric UOA_{L_2}, is globally good, both small structures (trees, buildings, etc.) and large structures (rivers, roads, grass fields, etc.) are well segmented; there is no remarkable presence of under-segmentation, but there are some over-segmented regions. Although, these errors represent a lesser problem as they are correctly detected by the local quality scores of our approach and they can be easily corrected as we suggest in the following sub-section.

2) Local quality: Now, we show how the local quality information provided for every segment may be used in later process in order to refine segmentation results or to help guiding the classification process by the injection of this supplementary knowledge. Let us observe Figure 12; the image is a 256 × 256 pixels extract containing mainly residential buildings. Figure 12(b) shows an optimal segmentation of this extract found using UOA_{L_2}. If we focus on the houses, we can see that some of them are over-segmented, and this is correctly indicated by the local evaluation function ϕ_δ (yellow segments), these segments could be merged together in order to obtain a more accurate segmentation of the houses for example. Similarly, some segments were correctly identified as under-segmented (red segments) which could be improved by a shrinking operation, for instance.

V. Conclusion

Supervised metrics provide efficient solutions to quantify similarity between segmentation results, but they require a ground-truth segmentation, which is mostly impossible to obtain in remote sensing applications. Unsupervised metrics rely on intrinsic properties computed directly from the resulting segments. Most of them rely only on global measurements; we argued that segmentation quality should be an aggregation of the composing segments local quality. Based on this assertion, we presented a novel approach allowing to estimate local quality for each segment (i.e., indicating whether the segment is over-segmented, under-segmented or well-isolated from its neighbourhood). We then defined two aggregation functions which can be used to combine the local quality estimates into a global quality score. The first estimate, UOA_Σ, is informative when results are rather very under-segmented or very over-segmented, but it can lead to undesired results when both errors occur simultaneously. The second estimate, UOA_{L_2}, considers independently over- and under-segmentation in a 2-dimensional space, aiming to minimize both types of error at the same time, thus reporting global errors due to over- and under-segmentation. We compared our metrics to the Z and the SU metrics which are, to our knowledge, the state of the art in unsupervised evaluation of remote sensing image segmentation. The results show that UOA_{L_2} outperforms the Z and SU metrics, based on close visual inspection. Furthermore, both UOA_Σ and UOA_{L_2} are bounded and show more consistent results than the compared metrics (i.e., similar segmentations have similar quality scores).

Our experiments show that the proposed metrics can be used to filter out *bad* quality segmentations, and to find segmentations having a good trade-off between over- and under-segmented regions at a given scale. Nevertheless, a single segmentation scale is not enough to isolate correctly every object in the image. Yet, the homogeneity threshold δ has an obvious influence on *good* segments size. Thus, it seems possible to learn from examples the right threshold for a given class of objects. Moreover, the local evaluation function we proposed in this article measures the quality of each segment in a hard manner (i.e., a segment is either over-segmented, under-segmented or well isolated). In our future work, we will enhance our approach by more accurately quantifying local quality, for example, by considering the distance between the homogeneity $H(R_i)$ of a given segment and the homogeneity threshold δ. These two directions constitute our future research perspectives.

Acknowledgements

The research leading to these results has received funding from the French *Agence Nationale de la Recherche* (Grant Agreement ANR-12-MONU-0001).

References

[1] M. İlsever and C. Ünsalan, *Two-Dimensional Change Detection Methods: Remote Sensing Applications*. SpringerBriefs in Computer Science, Springer, 2012.

[2] H. M. Pham, Y. Yamaguchi, and T. Q. Bui, "A case study on the relation between city planning and urban growth using remote sensing and spatial metrics," *Landscape Urban Plan*, vol. 100, pp. 223–230, 2011.

[3] A. Räsänen, A. Rusanen, M. Kuitunen, and A. Lensu, "What makes segmentation good? a case study in boreal forest habitat mapping," *Int. J. Remote Sens.*, vol. 34, pp. 8603–8627, 2013.

[4] C. V. Westen, "Remote sensing and gis for natural hazards assessment and disaster risk management," in *Treatise on Geomorphology*, pp. 259–298, Academic Press, 2013.

[5] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J Photogramm*, vol. 65, pp. 2–16, 2010.

[6] T. Esch, M. Thiel, M. Bock, A. Roth, and S. Dech, "Improvement of image segmentation accuracy based on multiscale optimization procedure," *IEEE Geosci. Remote. S.*, vol. 5, pp. 463–467, July 2008.

[7] H. Cheng, X. Jiang, Y. Sun, and J. Wang, "Color image segmentation: Advances and prospects," *Pattern Recogn*, vol. 34, pp. 2259–2281, 2001.

[8] J. Michel, D. Youssefi, and M. Grizonnet, "Stable mean-shift algorithm and its application to the segmentation of arbitrarily large remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, pp. 952–964, 2015.

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

[10] Z. Wang, J. R. Jensen, and J. Im, "An automatic region-based image segmentation algorithm for remote sensing applications," *Environ Modell Softw*, vol. 25, pp. 1149–1165, 2010.

[11] S. Derivaux, G. Forestier, C. Wemmert, and S. Lefèvre, "Supervised image segmentation using watershed transform, fuzzy classification and evolutionary computation," *Pattern Recogn Lett*, vol. 31, pp. 2364–2374, 2010.

[12] B. Peng, L. Zhang, and D. Zhang, "A survey of graph theoretical approaches to image segmentation," *Pattern Recogn*, vol. 46, pp. 1020–1038, 2013.

[13] Q. Zhan, M. Molenaar, K. Tempfli, and W. Shi, "Quality assessment for geo-spatial objects derived from remotely sensed data," *Int. J. Remote Sens.*, vol. 26, pp. 2953–2974, July 2005.

[14] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Comput Vis Image Und*, vol. 110, pp. 260–280, 2008.

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV, Proc.*, pp. 416–423, 2001.

[16] D. W. Paglieroni, "Design considerations for image segmentation quality assessment measures," *Pattern Recogn*, vol. 37, pp. 1607–1617, 2004.

[17] P. Corcoran, A. Winstanley, and P. Mooney, "Segmentation performance evaluation for object-based remotely sensed image analysis," *Int. J. Remote Sens.*, vol. 31, pp. 617–645, 2010.

[18] H. Vojodi, A. Fakhari, and A. M. E. Moghadam, "A new evaluation measure for color image segmentation based on genetic programming approach," *Image Vision Comput*, vol. 31, pp. 877–886, 2013.

[19] H. Zhang, J. E. Fritts, and S. A. Goldman, "An entropy-based objective evaluation method for image segmentation," in *Electronic Imaging 2004*, pp. 38–49, 2003.

[20] J. F. Khan and S. M. Bhuiyan, “Weighted entropy for segmentation evaluation,” *Opt. Laser Technol.*, vol. 57, pp. 236–242, 2014.

[21] S. Chabrier, C. Rosenberger, H. Laurent, B. Emile, and P. Marché, *Evaluating the segmentation result of a gray-level image*. 2004.

[22] S. Srubar, “Quality measurement of image segmentation evaluation methods,” in *SITIS, Proc.*, pp. 254–258, 2012.

[23] H.-C. Chen and S.-J. Wang, “The use of visible color difference in the quantitative evaluation of color image segmentation,” in *ICASSP. Proc.*, pp. 593–596, 2004.

[24] X. Zhang, P. Xiao, and X. Feng, “An unsupervised evaluation method for remotely sensed imagery segmentation,” *IEEE Geosci Remote Lett*, vol. 9, pp. 156–160, 2012.

[25] B. Johnson and Z. Xie, “Unsupervised image segmentation evaluation and refinement using a multi-scale approach,” *ISPRS J Photogramm*, vol. 66, pp. 473–483, 2011.

[26] N. S. Anders, A. C. Seijmonsbergen, and W. Bouten, “Segmentation optimization and stratified object-based analysis for semi-automated geomorphological mapping,” *Remote Sens. Environ.*, vol. 115, no. 12, pp. 2976 – 2985, 2011.

[27] L. Drăguţ, O. Csillik, C. Eisank, and D. Tiede, “Automated parameterisation for multi-scale image segmentation on multiple layers,” *ISPRS J. Photogramm. Remote Sens.*, vol. 88, no. 0, pp. 119 – 127, 2014.

[28] C. Persello and L. Bruzzone, “A novel protocol for accuracy assessment in classification of very high resolution images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, pp. 1232–1244, March 2010.

[29] E. Christophe and J. Inglada, “Open source remote sensing: Increasing the usability of cutting-edge algorithms,” *IEEE Geosci. Remote Sens. Newslett.*, vol. 35, pp. 9–15, 2009.

[30] J. Yang, P. Li, and Y. He, “A multi-band approach to unsupervised scale parameter selection for multi-scale image segmentation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 94, no. 0, pp. 13 – 24, 2014.

[31] H. Kekre, S. D. Thepade, T. K. Sarode, and V. Suryawanshi, “Image retrieval using texture features extracted from glcm, lbg and kpe,” *International Journal of Computer Theory and Engineering*, vol. 2, no. 5, pp. 1793–8201, 2010.

Andrés Troya-Galvis obtained the MSc at Université de Bourgogne, France in 2012. He is now a PhD Student at Université de Strasbourg, France. His scientific interests include machine learning, image segmentation, data quality, and collaborative methods for remote sensing image analysis.

Pierre Gançarski is full Professor of Computer Science. He is affiliated to the Laboratory ICUBE (University of Strasbourg). His work focuses on collaborative multistrategy clustering, with application to remote sensing image analysis for urban and biodiversity applications. He proposes new approaches to a better exploitation of Earth Observation integrating multisources data to identify and monitor natural or anthropic environments. He has been the coordinator of different research projects on Collaborative Classification applied to Environmental Data. He has been leader (until 2012) of the DataMining Group of the laboratory ICube. He is also codirector of a national research group in Big Data and Data Sciences (GDR MaDiSC). He is author of over 30 papers in international or national peer-reviewed journals and more than 50 papers in national or international conferences. He has been promoter of 7 PhD students since 2007.

Nicolas Passat obtained the MSc and PhD from Université Strasbourg 1 in 2002 and 2005, and Habilitation from Université de Strasbourg in 2011. He was an assistant professor at Université de Strasbourg, France, between 2006 and 2012. He is now a full professor at Université de Reims Champagne-Ardenne, France. His scientific interests include mathematical morphology, discrete topology, medical imaging and remote sensing.

Dr. Laure Berti-Équille is currently a Senior Scientist at *Qatar Computing Research Institute Director (QCRI)*. Before, she was a Research Director (“Directeur de Recherche”) at *IRD, Institut de Recherche pour le Développement* (2011-2013) and a tenured Associate Professor at the University of Rennes 1 in France (2000-2010). From 2007 to 2009, she was a visiting researcher at AT&T Labs Research (NJ, USA) with a Marie Curie fellowship. Her research interests range from data preprocessing, data cleaning, data integration, truth discovery to anomaly detection and exploratory data analysis.