ISPRS Journal of Photogrammetry and Remote Sensing 102 (2015) 73–84

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Segmentation quality evaluation using region-based precision and recall measures for remote sensing images

Xueliang Zhang, Xuezhi Feng, Pengfeng Xiao *, Guangjun He, Liujun Zhu

Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University, China
Key Laboratory for Satellite Mapping Technology and Applications of State Administration of Surveying, Mapping and Geoinformation of China, Nanjing University, China
Department of Geographic Information Science, Nanjing University, China

ARTICLE INFO

Article history:
Received 27 August 2014
Received in revised form 4 December 2014
Accepted 16 January 2015

Keywords:
High-spatial resolution remote sensing
Image segmentation
Segmentation evaluation
Discrepancy measure
Precision and recall
Geographic object-based image analysis

ABSTRACT

Segmentation of remote sensing images is a critical step in geographic object-based image analysis. Evaluating the performance of segmentation algorithms is essential to identify effective segmentation methods and optimize their parameters. In this study, we propose region-based precision and recall measures and use them to compare two image partitions for the purpose of evaluating segmentation quality. The two measures are calculated based on region overlapping and presented as a point or a curve in a precision–recall space, which can indicate segmentation quality in both geometric and arithmetic respects. Furthermore, the precision and recall measures are combined by using four different methods. We examine and compare the effectiveness of the combined indicators through geometric illustration, in an effort to reveal segmentation quality clearly and capture the trade-off between the two measures. In the experiments, we adopted the multiresolution segmentation (MRS) method for evaluation. The proposed measures are compared with four existing discrepancy measures to further confirm their capabilities. Finally, we suggest using a combination of the region-based precision–recall curve and the *F*-measure for supervised segmentation evaluation.

1. Introduction

Detailed geometric information of geographical objects is typically presented in high-spatial resolution remote sensing images. To retrieve information from these images, geographic object-based image analysis (GEOBIA) has been widely used and represents an evolving paradigm (Blaschke et al., 2014). GEOBIA operates based on segmented objects, which make it less sensitive to spectral variation in high spatial resolution images. Moreover, it can effectively use the spatial information of objects, especially contextual information (Blaschke and Strobl, 2001; Schiewe et al., 2001; Burnett and Blaschke, 2003; Yu et al., 2006; Hay and Castilla, 2008).

Image segmentation is a prerequisite for GEOBIA. Its objective is to partition an image into spatially contiguous and homogeneous regions. In GEOBIA, segmented regions are viewed as image objects, which represent the core of object-based analysis (Benz et al., 2004; Blaschke, 2010). Thus, evaluation of segmentation quality is essential for GEOBIA in selecting effective segmentation approaches and determining optimal parameters (Neubert et al., 2008).

In general, strategies for evaluating segmentation quality include visual analysis, system-level evaluation, and empirical goodness and empirical discrepancy methods. Visual analysis is easy to perform, but it cannot provide quantitative evaluation results and is inevitably affected by subjectivity. In system-level evaluation, segmentation is viewed as a part of the classification system. Segmentation quality is evaluated by classification accuracy, in which a high classification accuracy indicates high segmentation quality (Laliberte and Rango, 2009; Smith, 2010; Gao et al., 2011; Dronova et al., 2012). This strategy proves that segmentation quality directly affects the performance of GEOBIA. However, it does not reflect the many properties intrinsic to a segmentation algorithm that are independent of applications (Unnikrishnan et al., 2007).

The empirical goodness method (also known as unsupervised evaluation method) assesses segmentation quality primarily by calculating indicators of homogeneity within regions or differences between regions (Levine and Nazif, 1985; Stein and De Beurs, 2005; Chabrier et al., 2006; Espindola et al., 2006; Faur et al.,

* Corresponding author at: Department of Geographic Information Science, Nanjing University, China. Tel.: +86 25 89680612.
E-mail address: xiaopf@nju.edu.cn (P. Xiao).

http://dx.doi.org/10.1016/j.isprsjprs.2015.01.009
0924-2716/

2009; Corcoran et al., 2010; Drăguţ et al., 2010; Gao et al., 2011; Johnson and Xie, 2011; Zhang et al., 2012). Recently, this strategy has received much attention (Zhang et al., 2008), primarily because it is objective and does not require the ground truth. However, researchers who employ the strategy have the added difficulty of designing effective indicators and explaining the meaning of absolute indicator values.

The empirical discrepancy method, or called as supervised evaluation method, evaluates segmentation quality by comparing segmentation results with ground truth (Zhang, 1996; Neubert et al., 2008; Clinton et al., 2010). This strategy is apparently effective but it faces the difficulty of generating objective references because of the uncertainty caused by different interpreters. Martin (2003) and Albrecht (2010) examined the uncertainty for natural images and remote sensing images, respectively. The results showed that the references produced by different interpreters vary in their degree of details, but the reference objects are consistent with each other. It is indicated that the evaluation based on a single reference is applicable to a given amount of details.

The empirical discrepancy method can be performed on several selected objects (Neubert et al., 2008; Clinton et al., 2010) or on a partition of an image (Cardoso and Corte-Real, 2005; Carleer et al., 2005). The first case is suitable for optimizing segmentation parameters or evaluating segmentation performance for certain land cover classes. However, ensuring that the selected objects appropriately represent the land cover classes or the image scene is a challenge. The second case is based on the definition of image segmentation in which two partitions are compared (Cardoso and Corte-Real, 2005). However, it requires considerable effort to delineate all the objects in an image as reference objects.

The discrepancy between a segmentation and reference directly reveals the segmentation quality. If the discrepancy is small, the segmentation quality is high. In an ideal case, if no discrepancy exists, the segmented regions are identical to the reference objects. Two types of discrepancies should be evaluated: geometric and arithmetic (Liu et al., 2012). Measuring the discrepancies separately is not difficult. However, measuring them concurrently is a challenge.

Geometric discrepancy is usually measured by boundary matching or region overlapping. Segmentation can simply be treated as a boundary map. Then the discrepancy measures are calculated on an edge-versus-non-edge basis or by prioritizing the edge pixels according to their distance to the reference (Martin et al., 2004; Estrada and Jepson, 2009; Albrecht, 2010). However, in most cases, especially in remote sensing image segmentation, the region overlapping strategy is often employed (Räsänen et al., 2013; Lucieer, 2004; Carleer et al., 2005; Zhan et al., 2005; Möller et al., 2007; Tian and Chen, 2007; Neubert et al., 2008; Weidner, 2008; Clinton et al., 2010; Liu et al., 2012; Marpu et al., 2010; Persello and Bruzzone, 2010; Witharana et al., 2014). This may be not only because the region overlapping reflects the geometric discrepancy but also because the image objects in GEOBIA are closer to the unit of region than to the edge pixels.

Arithmetic discrepancy corresponds to over- and under-segmentation. During over-segmentation, a reference object may be separated into several segments, whereas during under-segmentation, a segmented region may correspond to several reference objects. A direct approach to identifying arithmetic discrepancy is to compare segments with reference objects regarding their total numbers (Carleer et al., 2005; Persello and Bruzzone, 2010; Xiao et al., 2010; Liu et al., 2012).

Certain geometric discrepancy measures can reflect arithmetic discrepancy implicitly. Some measures can tolerate over-segmentation but incur penalties during under-segmentation, such as the measure E proposed by Carleer et al. (2005), asymmetric partition distance (Cardoso and Corte-Real, 2005), and global consistency error (Martin, 2003). In a stricter case, certain measures are intolerant to both over- and under-segmentation, such as the Rand index (Rand, 1971; Hubert and Arabie, 1985), symmetric partition distance (Cardoso and Corte-Real, 2005), and bidirectional consistency error (Martin, 2003). Measures that tolerate under-segmentation and mutual refinement are examined by Cardoso and Corte-Real (2005) and Martin (2003).

Some studies have proposed evaluating both geometric and arithmetic discrepancies by combining several indicators. For example, the indicators of over- and under-segmentation have been combined to evaluate the two discrepancies concurrently (Levine and Nazif, 1982; Yang et al., 1995; Clinton et al., 2010). Persello and Bruzzone (2010) presented a set of indices that characterize five types of geometric errors and then combined them. Hoover et al. (1996) designed six indicators to apply to different segments in order to evaluate segmentation performance. However, these combination strategies must be carefully designed to achieve potential trade-offs of different measures. In addition, when the number of measures is large, combining them is difficult.

This study proposes discrepancy measures that are sensitive to both geometric and arithmetic errors when comparing two partitions. Region-based precision and recall measures are proposed based on region overlapping, which indicate segmentation quality in a precision–recall space. A precision–recall curve is a standard evaluation technique employed in the field of information retrieval (Van Rijsbergen, 1979) and is commonly used for evaluating edge or boundary detection results (Martin et al., 2004; Estrada and Jepson, 2009). Martin (2003) proposed the region precision–recall curve method based on information theory. In this study, we similarly design the precision and recall measures based on region overlapping and extend them for use with remote sensing images. When calculating these measures, we focus on matching direction and corresponding criteria for region overlapping. Furthermore, the precision and recall measures are combined to reveal segmentation quality by catching the trade-offs of the two measures. We reveal the effectiveness of four combination strategies by means of geometric illustration.

The remainder of the paper is organized as followed. Section 2 describes the study area as well as image data and references. Section 3 discusses the evaluation and segmentation methods employed in this study. The experimental results are presented in Section 4 and reveal the effectiveness of our discrepancy measures. Section 5 includes additional discussion. Finally, a conclusion is drawn in Section 6.

2. Study area and data

A QuickBird scene in Hangzhou, China, acquired on March 2, 2008, is used as the image data. The city is located at approximately 30.3°N and 120.2°E and is a Chinese city undergoing rapid economic development. The spatial resolution is 0.6 m after pan-sharpening is performed using the method proposed by Zhang (2002). The radiometric resolution is 16 bits. Four subsets in the QuickBird scene are used as test images to show segmentation evaluation results. The test images are identified as T1, T2, T3, and T4, representing the landscape for a watery area, farmland, and residential and industrial zones, respectively. The image sizes are 538×546, 474×489, 793×623, and 996×550 pixels for test images T1–T4, respectively.

References for the test images were digitized by a specialist in remote sensing. Manual extractions were then reviewed by a second operator to catch any obvious errors. To reduce variation in levels of details, operators were told to outline the boundaries of each geographic object rather than those of homogeneous regions because the definition of homogeneity is more arbitrary and

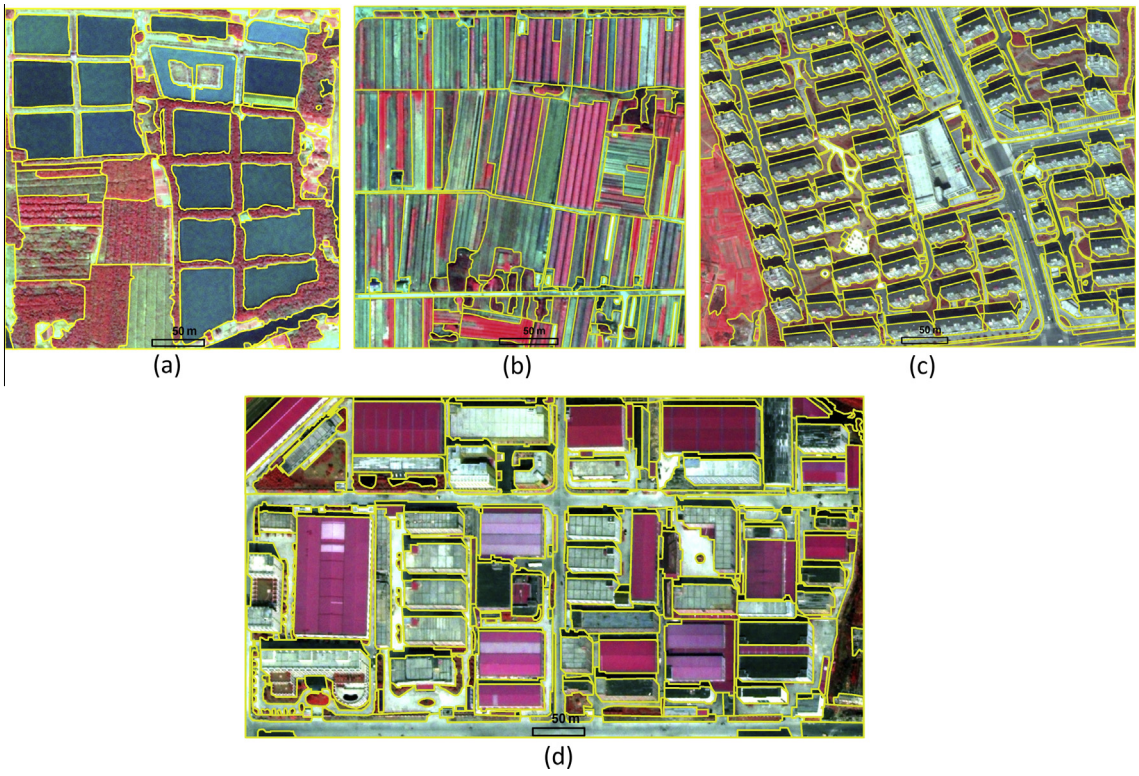

Fig. 1. Reference of test images T1 (a), T2 (b), T3 (c), and T4 (d). The images are shown with a combination of near infrared, red and green bands. Yellow lines represent the boundaries of reference objects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

limiting than that of a geographic object (Martin, 2003). Finally, 107, 156, 554, and 292 reference objects were produced in test images T1–T4, respectively, as shown in Fig. 1.

In this study, the multiresolution segmentation (MRS) method (Baatz and Schäpe, 2000) is used to produce segmentation results for evaluation. The shape and scale parameters of MRS are set as followed. Five shape parameters (0.1, 0.3, 0.5, 0.7, and 0.9) are set for each test image. For Shape 0.5, the scale parameters are set as presented in Table 1. For the other shape parameters, the scale parameters are set differently to achieve similar segmentation scales to those of Shape 0.5. The serial numbers of scales are defined as {1, 2,...,n}, where n is 11 for T1 and T2 and 14 for T3 and T4.

Table 1
Scale parameters for test images T1–T4 given Shape 0.5.

Serial number of scale	Scale parameter	
	T1&T2	T3&T4
1	20	20
2	25	30
3	30	40
4	40	50
5	50	60
6	60	70
7	70	80
8	80	90
9	90	100
10	100	110
11	110	120
12		130
13		140
14		150

3. Methodology

A schematic of calculating precision and recall measures to evaluate segmentation quality is shown in Fig. 2. First, we overlap segmentation result with reference to calculate the precision and recall values by inverse matching direction. The two measures are jointly used to indicate segmentation quality. Furthermore, the two measures can be combined into a single measure to evaluate segmentation quality.

3.1. Region-based precision and recall measures

The precision and recall measures are calculated based on region overlapping. Two aspects related to region overlapping are declared prior to experimentation: the matching direction and the corresponding criteria. The matching direction for the precision measure is defined as a reference-to-segment directional correspondence. In other words, we match the reference objects to the segments, and not vice versa. However, for the recall measure, we reverse this and match segments to reference objects. The precision measure reflects the quality of the segments, whereas the recall measure indicates the manner in which the reference objects are described by the segments. The corresponding criteria help to determine the sensitivity to over- and under-segmentation. Some studies have defined the corresponding regions as those with greater than 50% of pixels located in the target regions of the matching process (Weidner, 2008; Clinton et al., 2010). Other studies have identified corresponding segments as those with a maximal overlapping area (e.g., Lucieer, 2004; Carleer et al., 2005). We adopt this second strategy because it is more sensitive to the over- and under-segmentation than is the first strategy. As shown in Fig. 3, if the first strategy is adopted when matching the three regions $\{O_{11}, O_{12}, O_{13}\}$ to the region O_{21}, all three regions are viewed as corresponding segments, and the overlapping area is equal to O_{21}. However, if the second strategy is adopted, only the region O_{12} is defined as the corresponding segment and the

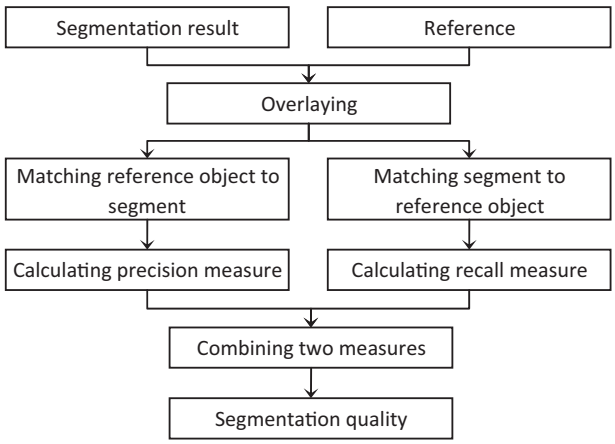

Fig. 2. Schematic of calculating precision and recall measures to evaluate segmentation quality.

overlapping area is equal to O_{12}. The area of O_{12} is smaller than that of O_{21}, incurring a penalty for over-segmentation if O_{21} is a reference object or for under-segmentation if O_{21} is a segmented region.

Given the segmentation result S with n segments $\{S_1, S_2, \ldots, S_n\}$ and the reference R with m objects $\{R_1, R_2, \ldots, R_m\}$, the precision measure is calculated by matching $\{R_i\}$ to each segment S_i and the recall measure by matching $\{S_i\}$ to each reference object R_i.

When calculating the precision measure, the matched reference object (R_{imax}) for each segment S_i is first identified, where R_{imax} has the largest overlapping area with S_i. The precision measure is then defined as:

$$Precision = \sum_{i=1}^{n} |S_i \cap R_{imax}| \Big/ \sum_{i=1}^{n} |S_i|, \quad (1)$$

where $|\cdot|$ denotes the area that is represented by the number of pixels in a region.

Similarly, the matched segment (S_{imax}) for each reference object R_i is searched according to the maximal overlapping area criterion and the recall measure is defined as:

$$Recall = \sum_{i=1}^{m} |R_i \cap S_{imax}| \Big/ \sum_{i=1}^{m} |R_i|. \quad (2)$$

The precision and recall measures are jointly used to indicate segmentation quality. If the precision and recall values of one segmentation result are both higher than a second segmentation result, this indicates high segmentation quality. In an ideal case in which the segmentation is identical to the reference, both the precision and recall measures achieve the largest value of 1. Moreover, the joint use of the two measures can indicate both over- and under-segmentation and is not biased toward over- or under-segmented images. When an image is over-segmented, the precision value is large but the recall value decrease to incur the penalty. In an extreme over-segmentation case in which each pixel is a segment, the precision value is 1 and the recall value is as low as m/N, where m is the number of reference objects and N is the number of pixels in the image. By contrast, when an image is under-segmented, the recall value is high but the precision value decreases. In an extreme under-segmentation case in which the entire image is viewed as a segment, the recall value is 1 and the precision value decreases to $|R_{max}|/N$, where R_{max} represents the largest reference object.

Precision and recall values can be presented in a precision–recall space to indicate segmentation quality visually. The precision–recall space allows us to characterize the performance of segmentation algorithms in a manner that is independent of particular choices regarding input parameters (Estrada and Jepson, 2009). Each point in the two-dimensional space corresponds to a segmentation result. A precision–recall curve can then be plotted based on a set of segmentations produced by setting different values for a single parameter. Thus, a precision–recall curve can show a change in segmentation quality caused by a parameter. For example, a curve can indicate a change from a fine scale to a coarse scale, as in Fig. 6. In addition, we can determine the effectiveness of more than one segmentation parameter by comparing different curves.

3.2. Combination of precision and recall measures

As previously mentioned, if the precision and recall values of one segmentation are both higher or both lower than those of another segmentation, distinguishing between the two segmentations is easy. For example, a point (0.7, 0.7) in the precision–recall space apparently indicates higher segmentation quality than (0.6, 0.6). However, it is common that the two measures of one segmentation do not concurrently have higher or lower values than those of another segmentation. For example, determining the higher segmentation quality achieved between (0.7, 0.7) and (0.8, 0.6) is difficult. To deal with this complexity, the precision and recall measures can be combined into a single measure to capture the trade-off between the two measures. The combined measures can then clearly expose segmentation quality. However, the effectiveness of the combined measures is directly determined by the manner in which they are combined. Accordingly, we compared four combination strategies to reveal their effectiveness.

The first combination strategy is the F-measure (Van Rijsbergen, 1979), which is defined as:

$$F = \frac{1}{\alpha \frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}}, \quad (3)$$

where the weight α is a constant 0.5 in this study. The value of F-measure ranges from 0 to 1.

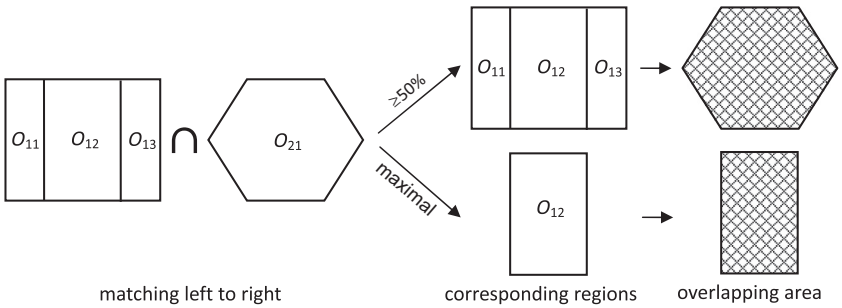

Fig. 3. Sample illustrations of corresponding criteria and its sensitivity to over- and under-segmentation when matching regions $\{O_{11}, O_{12}, O_{13}\}$ to region O_{21}.

The second combination strategy involves the sum of the precision and recall measures, which is similar to the method of combining the over- and under-merging indicators (Levine and Nazif, 1982). The *SUM* value is less than 2.

$$SUM = Precision + Recall. \tag{4}$$

The Euclidean distance strategy is similar to the method combining the over- and under-segmentation measures, as introduced by Clinton et al. (2010). However, we define two distances for the point in the precision–recall space. The first is identified as *ED*, indicating the distance to the point (0, 0) in the space. The second is identified as *ED′*, indicating the distance to (1, 1). Both *ED* and *ED′* values change from 0 to $\sqrt{2}$.

$$ED = \sqrt{Precision^2 + Recall^2}. \tag{5}$$

$$ED' = \sqrt{(1 - Precision)^2 + (1 - Recall)^2}. \tag{6}$$

As discussed in Section 3.1, large precision and recall values indicate high segmentation quality. According to the above definitions of the combined measures, large precision and recall values increase the *F*-measure, *SUM*, and *ED* values but decrease the *ED′* value. Hence, large *F*-measure, *SUM*, and *ED* values and small *ED′* value indicate high segmentation quality. Geometrically, all the four combined measures show that the points located in the upper-right portion of the precision–recall space indicate high segmentation quality. The combined measures determine the manner in which to separate the upper-right and lower-left portions of the precision–recall space. The space is separated by the precision–recall curve, in which all points have the same value of a combined measure, as shown in Fig. 4. We call this an "isocurve". Each value of a combined measure has a corresponding isocurve in the precision–recall space. When comparing two segmentations by means of a combined measure, if two corresponding points in the precision–recall space are located in the same isocurve, the two segmentations have performed identically. Otherwise, if the two points are not in the same isocurve, at least one isocurve on either side of which the two points are located must exist. The point located in the upper-right portion of the isocurve indicates higher segmentation quality than that of the other point.

The definition of a combined measure determines the shape of its isocurve. Different definitions correspond to different isocurve shapes, which results in different forms of separation within the precision–recall space. The isocurve shapes of the four combined measures are shown in Fig. 4. The isocurve shape of *SUM* is a straight line with a slope of 135°. The isocurve shape of *ED* and *ED′* is an arc of a circle centered in (0, 0) and (1, 1), respectively. However, the isocurve shape of the *F*-measure changes along with the values. Therefore, the indications of the four combined measures are inconsistent with one another due to the obtained trade-offs of the precision and recall measures. For example, when comparing the points P1 (0.5, 0.9) and P2 (0.7, 0.7) in the precision–recall space as shown in Fig. 5(a), we determine that the *SUM* measure indicates equivalent segmentation quality because the two points are located in the isocurve that has a value of 1.4. However, in the other three measures the two points are not located in the same isocurve. For the *ED* measure, P1 is located in the upper-right portion, indicating the better segmentation quality than that of P2. By contrast, P2 is located in the upper-right portion according to the isocurve of *ED′* and *F*-measure. In the experiments, the influence of different combination strategies on revealing segmentation quality will be analyzed in detail.

In addition, the isocurve shapes of the combined measures directly relate to the sensitivity of the measures to over- and under-segmentation. An example is shown in Fig. 5(b). Given point (0.9, 0.1) or point (0.1, 0.9) in the precision–recall space, severe over- or under-segmentation is indicated according to the definition in Section 3.1. The given point is located in the same isocurve with points (0.18, 0.18), (0.36, 0.36), (0.5, 0.5), and (0.64, 0.64) for the *F*-measure, *ED′*, *SUM*, and *ED*, respectively. This shows that the *F*-measure incurs the greatest penalty for over- and under-segmentation and has the highest sensitivity. The sensitivity of *ED′* is greater than that of *SUM* and the sensitivity of *ED* is the weakest.

3.3. Discrepancy measures for comparison

In this section, we present four existing discrepancy measures and compare them to the proposed measures. All measures are sensitive to both geometric and arithmetic discrepancies when two partitions are compared.

The first measure is the quality rate (*QR*) proposed by Weidner (2008). However, in our study, we change the corresponding criteria to the maximal cover because of measure sensitivity to arithmetic discrepancy. Furthermore, the matching direction is considered. QR_{sr} is then calculated by matching the segments to the reference objects and QR_{rs} by matching the reference objects to the segments.

$$QR_{sr} = \sum_{i=1}^{m} \frac{|R_i \cap S_{imax}| \cdot |R_i|}{|R_i \cup S_{imax}| \cdot |R|}, \tag{7}$$

$$QR_{rs} = \sum_{i=1}^{n} \frac{|S_i \cap R_{imax}| \cdot |S_i|}{|S_i \cup R_{imax}| \cdot |S|}, \tag{8}$$

where the meaning of the variables are equivalent to that in Eqs. (1) and (2). The largest value of both QR_{sr} and QR_{rs} is 1 when the

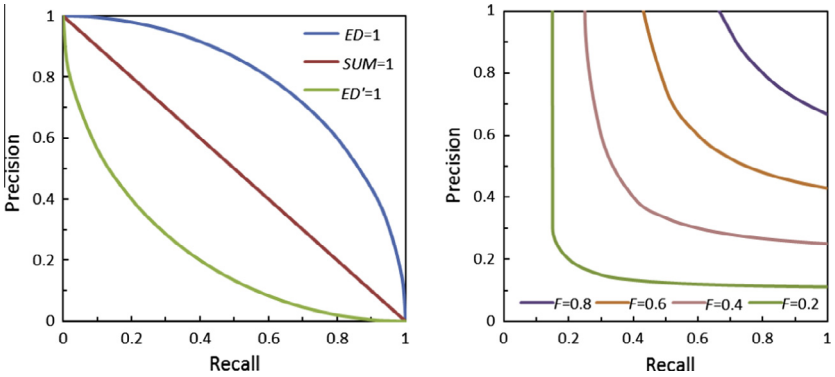

Fig. 4. Sample isocurves of the combined measures *SUM*, *ED* and *ED′* (left) and *F*-measure (right). An isocurve represents the precision–recall curve in which all points have the same value of a combined measure.

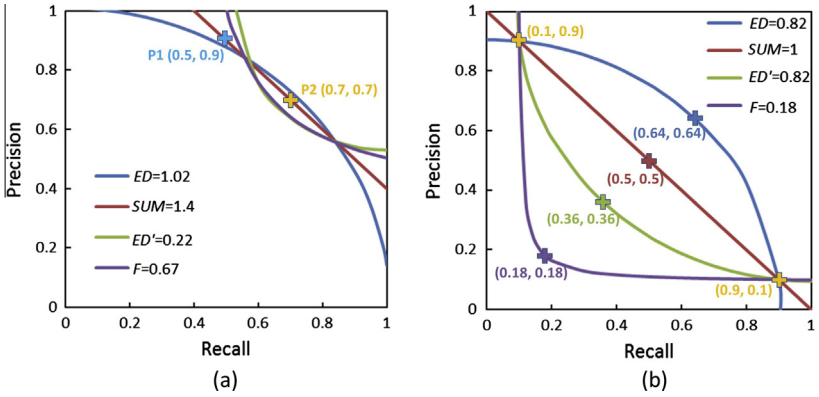

Fig. 5. (a) An example of inconsistent indications of the four combined measures; (b) an example to show the sensitivity of the combined measures to over- and under-segmentation.

reference is identical to the segmentation. Both over- and under-segmentation decrease the values of the two measures according to the definition.

The second measure is the symmetric partition distance (D_{sym}) proposed by Cardoso and Corte-Real (2005). D_{sym} is defined as the minimal number of pixels that must be removed from both the reference (R) and segmentation result (S) so that R and S in the remaining pixels are identical. Then D_{sym} is normalized to [0, 1] by dividing $(N-1)$, where N is the number of pixels in the image. In this study, we change D_{sym} to D_{sym}' according to the formula $D_{sym}' = 1 - D_{sym}$, which can be described as the maximal number of pixels that remain such that R equals S. The maximal D_{sym}' value is 1, indicating the equivalence of R and S without any pixels being deleted.

The third measure is the bidirectional consistency error (BCE) proposed by Martin (2003). First, the local refinement error (LRE) is defined to measure the degree to which R and S agree at a pixel p_i.

$$LRE(R,S,p_i) = \frac{|B(R,p_i) \setminus B(S,p_i)|}{|B(R,p_i)|}, \tag{9}$$

$$LRE(S,R,p_i) = \frac{|B(S,p_i) \setminus B(R,p_i)|}{|B(S,p_i)|}, \tag{10}$$

where $B(A,p)$ is the segment in partition A that contains point p, and $\cdot \setminus \cdot$ denotes the set difference. For example, $B(R,p) \setminus B(S,p)$ indicates the set of pixels $\{p | p \in B(R,p), p \notin B(S,p)\}$. In the over-segmentation case, $LRE(S,R,p_i)$ is large and $LRE(R,S,p_i)$ is small. By contrast, in the under-segmentation case, $LRE(S,R,p_i)$ is small and $LRE(S,R,p_i)$ is large. To incur penalty to both over- and under-segmentation, BCE is defined to adopt the larger local error at each pixel and combine all N pixels within the image.

$$BCE = \frac{1}{N}\sum_{i=1}^{N} \max\{LRE(R,S,p_i),\ LRE(S,R,p_i)\}. \tag{11}$$

We change BCE to the bidirectional consistency accuracy (BCA) based on $BCA = 1 - BCE$ so that the larger BCA value indicates higher segmentation quality. The largest BCA value is 1 when the segmentation and reference are identical.

The fourth measure is the adjusted Rand index (ARI) proposed by Hubert and Arabie (1985), which examines the correspondence between two partitions. ARI is an extension of the Rand index (Rand, 1971), which measures correspondence based on how object pairs are classified in a contingency table. However, ARI evaluates the level of agreement between two partitions based on a comparison of object triples. In other words, it examines the manner in which three distinct objects are delineated by the two partitions. ARI has a maximal value of 1, which means perfect agreement between the reference and segmented result. A large ARI value indicates a high correspondence to the reference.

3.4. Segmentation method under evaluation

Of the various image segmentation methods, the region-based method is particularly suitable and thus widely used for segmentation of remote sensing images (Schiewe et al., 2001; Carleer et al., 2005; Dey et al., 2010). The region-based method can produce spatially contiguous segmented regions having inherent continuous boundaries and these regions can be viewed as image objects directly within GEOBIA. In this study, the MRS method (Baatz and Schäpe, 2000) embedded in the commercial software eCognition Developer 8 is selected and evaluated to show the effectiveness of the proposed discrepancy measures.

The MRS method adopts a bottom-up region growing strategy starting from the pixel level based on local-mutual best merging. The merging criteria consists of the region size, the spectral homogeneity, and the region shape (compactness and smoothness), which indicates the similarity or merging probability of two adjacent regions. The scale parameter is defined as the threshold of the similarity, which determines the upper limit of the merging criteria. If the scale parameter is set at a large value, additional merging iterations are allowed and the segmented result becomes coarse. By contrary, a small scale parameter results in fine-scale segmentation result. Two other parameters for the MRS method exist: shape and compactness, both of which range from 0 to 1. If the shape parameter is set as small, MRS will concentrate primarily on generating segments with spectral homogeneity. By contrast, if the shape parameter is set as large, the segments will likely have a regular shape and neglect the spectral constraint. As the segmentation results are only somewhat sensitive to the compactness parameter (Witharana and Civco, 2014), we set the compactness parameter for this study at a constant 0.5 and focus on optimizing the scale and shape parameters.

4. Experimental results

4.1. Effectiveness of the region-based precision–recall curve

Given a shape parameter for the MRS method, a precision–recall curve is plotted on a set of multiscale segmentations in the precision–recall space, which reflects the change in quality in conjunction with segmentation scale. Furthermore, several precision–recall curves are plotted to show the influence of shape parameters on segmentation quality across scales. The precision–recall curves with different shape parameters are shown in Fig. 6.

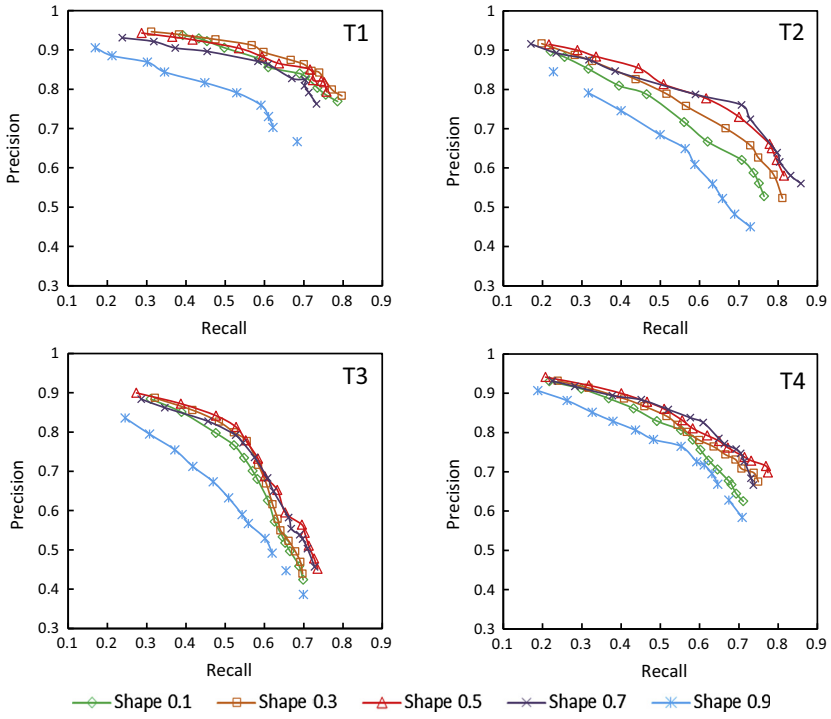

Fig. 6. Precision–recall curves plotted on multiscale segmentations produced by the MRS method. Given a shape parameter, each curve covers segmentations from Scale 1 to 11 for test images T1 and T2 and from Scale 1 to 14 for T3 and T4.

Each curve moves from the upper left to the lower right as the segmentation scale coarsens, which means that the precision value decreases as the recall value increases.

As explained in Section 3.2, a curve located in the upper-right part of the precision–recall space indicates high segmentation quality. This suggests the Shape 0.9 performs the worst for all test images in Fig. 6, because all Shape 0.9 curves are located in the lower-left part of the space. In addition, it is clear that Shape 0.1 is the second worst performer for test images T2–T4 and Shape 0.7 for T1. Then we can know with certainty that Shape 0.3 and 0.5 perform best for T1 and T3, respectively, because all points in the curve are located in the upper-right portion. But for T2 and T4, Shape 0.5 and 0.7 curves alternately perform best in conjunction with the change of segmentation scales, which means judging the best shape parameter is difficult. Moreover, although the precision–recall curve can reflect the change in quality during a coarsening of segmentation scales, it cannot directly reveal the optimal scale parameters. The combination of the precision and recall measures resolves these difficulties, as discussed in the next subsection.

4.2. Effectiveness of different combined measures

The combined measures *F*-measure, *SUM*, *ED*, and *ED′* for each point in Fig. 6 are plotted as curves in Fig. 7 for each test image. As illustrated in Section 3.2, the larger *F*-measure, *SUM* and *ED* values and the lower *ED′* values indicate higher segmentation quality. In general, the *ED* measure shows significant difference about indications from the other three measures. The change in range of *ED* values is small during the change of scale, and the changing direction of *ED* varies more often. Moreover, *ED* is not as sensitive to over-segmentation as are the other indicators. Taking T4 as an example, the poor performance of the *ED* indicator is revealed in Fig. 8. The isocurve shape of *ED* more closely resembles the actual precision–recall curves than do the other three indicators. Therefore, it cannot distinguish segmentation quality as well as the other indicators can.

Based on the combined indicators *F*-measure, *SUM* and *ED′*, we can definitely know the change of segmentation quality when setting different shape and scale parameters for MRS. For example, in Fig. 7, Shape 0.9 always performs the worst. Shape 0.3 and Shape 0.5 perform the best in most scales for T1 and T3, respectively. These indications confirm the results presented in Fig. 5. However, in Fig. 7, Shape 0.7 at Scale 6 and Shape 0.5 at Scale 13 perform the best for T2 and T4, respectively. These facts are not very clear in the precision–recall curves given in Fig. 5. Furthermore, the combined indicators clearly reflect the change of segmentation quality when the scale coarsens. For example, all the three indicators in Fig. 7 show that the segmentation quality tends to improve as the scale parameters for T1 and T4 increase. However, for T2 and T3, the segmentation quality improves in the initial stages, and then, after the highest quality is attained, the segmentation quality diminishes as the scale coarsens.

We present sample segmentations to further show the effectiveness of the combined measures. Given a shape parameter, sample segmentations at different scales of T1, T3, and T4 are shown in Fig. 9. For T1, the segmentation at Scale 1 is apparently over-segmented and the coarse segmentations at both Scale 6 and 10 can discriminate geographic objects well. For T3, the segmentations at Scales 1 and 10 are apparently over- and under-segmented, respectively. For T4, the segmentations at Scales 12, 13, and 14 are presented to reveal a slight difference. The green rectangles point out that the segmentation at Scale 13 can more effectively describe large buildings as single regions than can the segmentation at Scale 12. The yellow[1] rectangles show the under-segmentation at Scale 14. Accordingly in , the combined indicators *F*-measure, *SUM* and *ED′* show that the segmentations at Scale 6 and 10 have similar but significantly higher quality than that at Scale 1 for T1, the segmentation quality at Scale 6 is significantly higher than that

[1] For interpretation of color in Fig. 9, the reader is referred to the web version of this article.

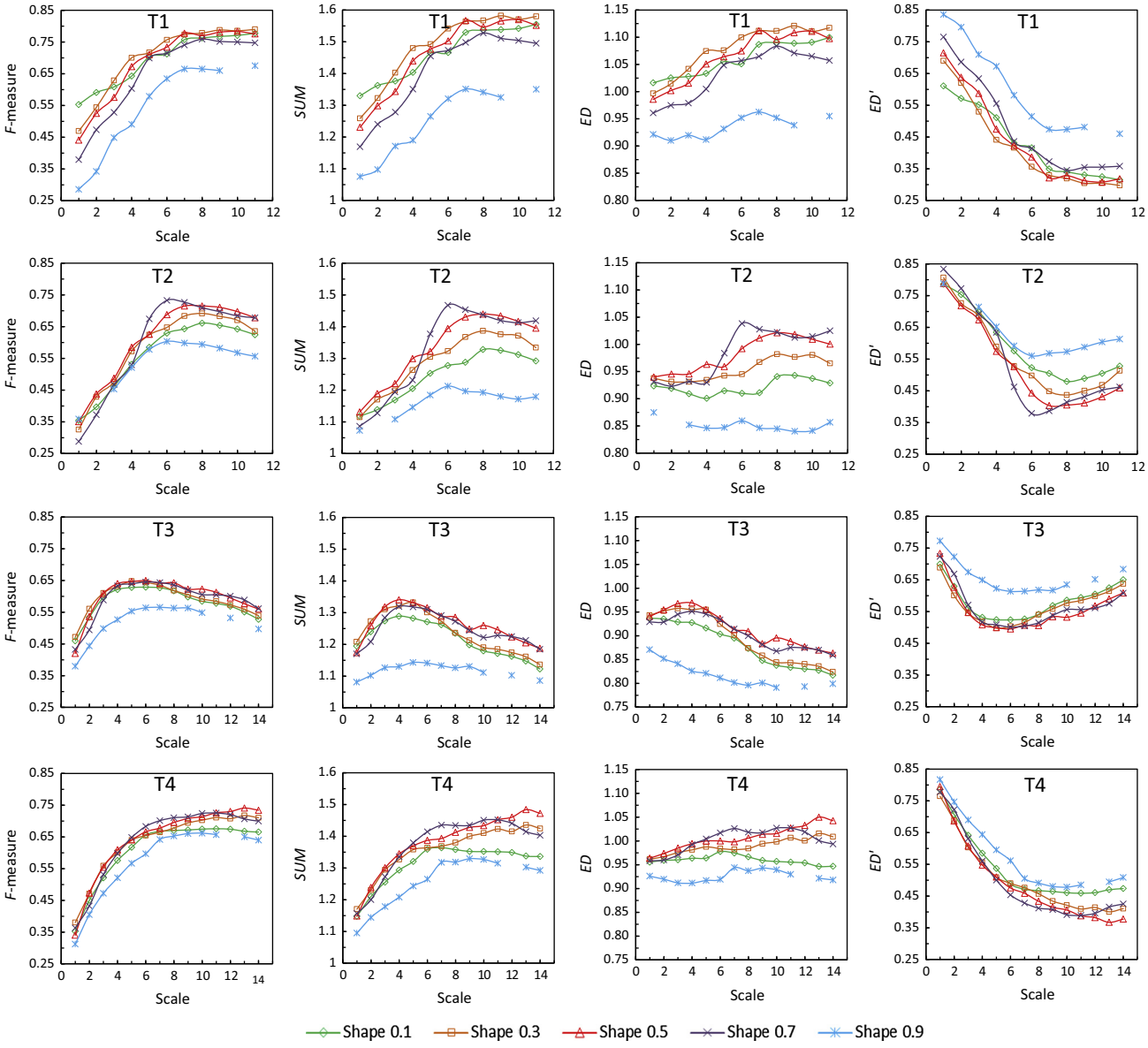

Fig. 7. Plots of the combined measures *F*-measure, *SUM*, *ED* and *ED′* for MRS segmentations with different scale and shape parameters.

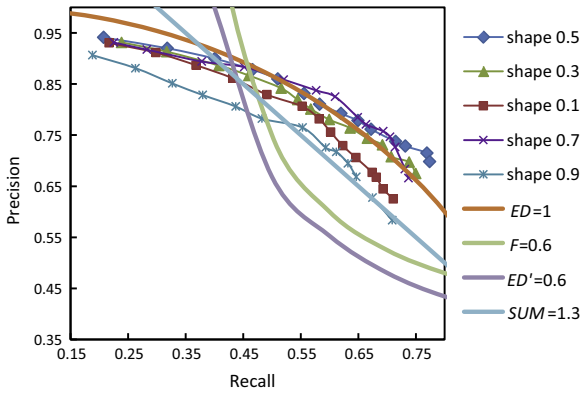

Fig. 8. An example to show the effectiveness of the combined indicators on distinguishing segmentation quality by overlaying the isocurves with the precision–recall curves plotted on MRS segmentations of test image T4.

at Scale 1 and 10 for T3, and the segmentation quality at Scale 13 is slightly higher than that at Scale 12 and 14 for T4.

In Fig. 10, sample MRS segmentations of T1 and T2 are presented to show differences caused by shape parameters. The differences are addressed by the yellow rectangles. For T1, given the scale serial number of 9, it shows that the segmentation with Shape 0.1 can produce accurate boundaries, but it is sensitive to spectral variation near boundaries and can result in over-segmentation in these areas. By contrast, the segmentation with Shape 0.7 does not reveal over-segmentation, but it produces inaccurate boundaries compared to that of Shape 0.3. Given Scale 6 for T2, Fig. 10 shows that the segmentation with Shape 0.7 can separate farmland with less influence of spectral variations near the boundaries, which results in over-segmentation for Shapes 0.5 and 0.3. Accordingly in Fig. 7, segmentations with Shapes 0.1, 0.3, and 0.7 at Scale 9 are ranked from better to worse as Shapes 0.3, 0.1 and 0.7. For T2, segmentation with Shape 0.7 is of higher quality than that with Shapes 0.5 and 0.3 at scale 6.

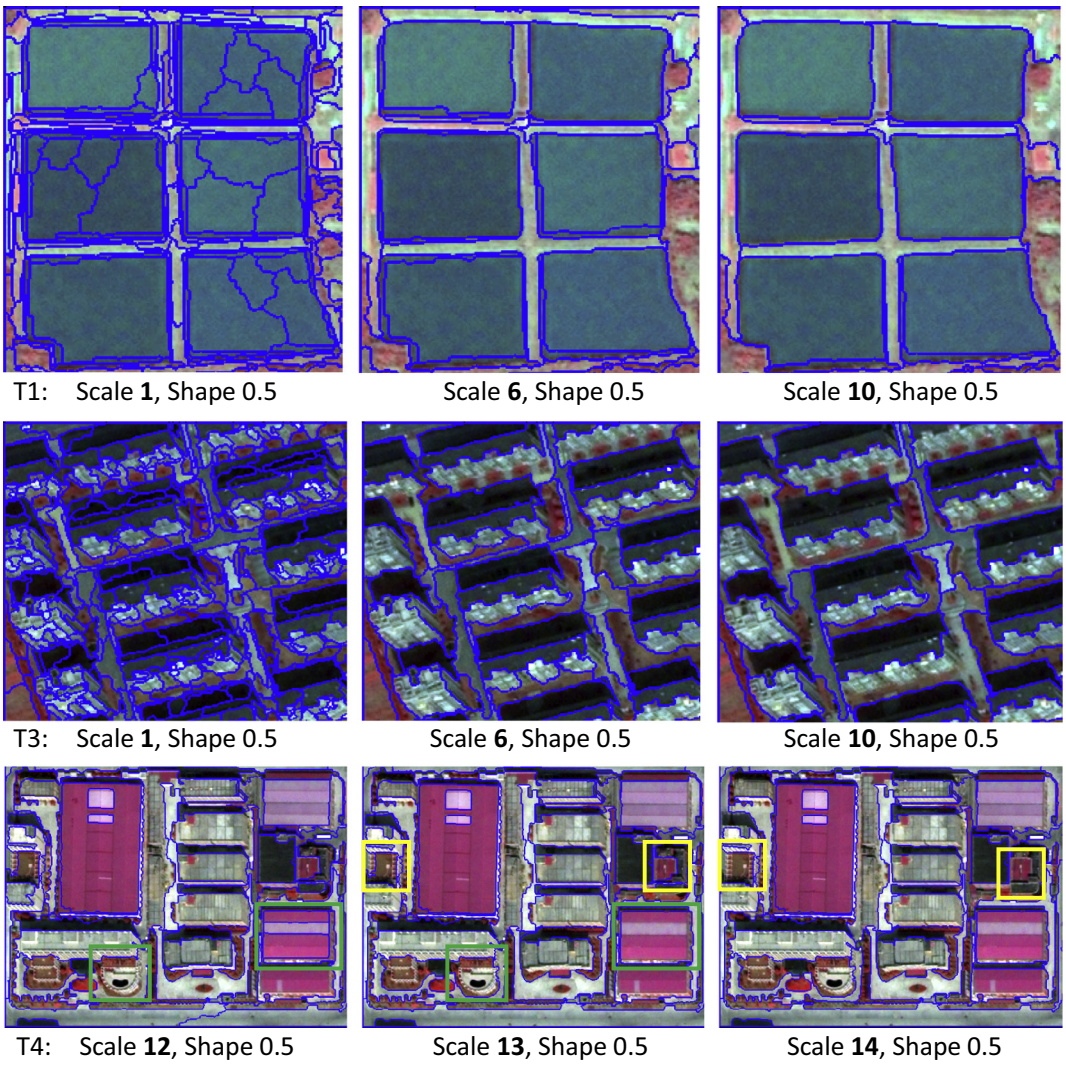

Fig. 9. Subsets of multiscale segmentations produced by the MRS method for test images T1, T3, and T4. The green and yellow rectangles in T4 highlight the differences between Scale 12 and 13 and Scale 13 and 14, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. Comparison with existing discrepancy measures

The effectiveness of the combined indicators is further demonstrated by comparing with the existing discrepancy measures QR_{sr}, QR_{rs}, D_{sym}', *BCA*, and *ARI* presented in Section 3.3. Because the combined indicators *F*-measure, *SUM* and *ED*′ perform similarly in Fig. 7, only the *F*-measure is presented for comparison in this subsection. The shape parameter of MRS is set as 0.5 for all test images and the scale parameters are set the same as those described in Section 2. The evaluation results of the multiscale segmentations are plotted in Fig. 11.

For all test images, the qualitative changes in Fig. 11 indicated by the *F*-measure and the measures employed for comparison are similar when the segmentation scale becomes coarser, except for the unusual changes of the *ARI* measure for T3. All measures show that the optimal scale is 13 for T4. Both the *F*-measure and the employed measures show a sudden decrease at Scale 8 and indicate the optimal segmentation at Scale 7 or Scale 10 for T1. Large variations in segmentation quality occur between the *F*-measure and the employed measures for T2 and T3. However, the variations are limited at Scales 7–9 for T2 and at Scales 4–6 for T3 because the differences in these scales for each measure are slight.

5. Discussion

The precision and recall measures are jointly used to evaluate segmentation quality. Evaluation results are presented in the precision–recall space. They are thus independent of segmentation parameters, which may prove advantageous when comparing different segmentation algorithms or optimizing parameters. When comparing two segmentation results, if one segmentation has higher values regarding both of the two measures than the other, then its quality is higher than that of the other. The proposed measures are sensitive to both over- and under-segmentation. In the case of over-segmentation, the precision value is large and the recall value is small. On the contrary, a large recall value and small precision value indicate under-segmentation.

A precision–recall curve is plotted on a set of multiscale segmentations, which clearly reflects the change from over- to under-segmentation. In general, precision value deceases and recall value increases as the segmentation scale coarsens. Through a comparison of different curves, the influences of shape parameter for the MRS method across multiple scales can be presented, as in Fig. 6. However, the precision–recall curve is not limited in this manner and can be applied to other cases. By combining Figs. 6

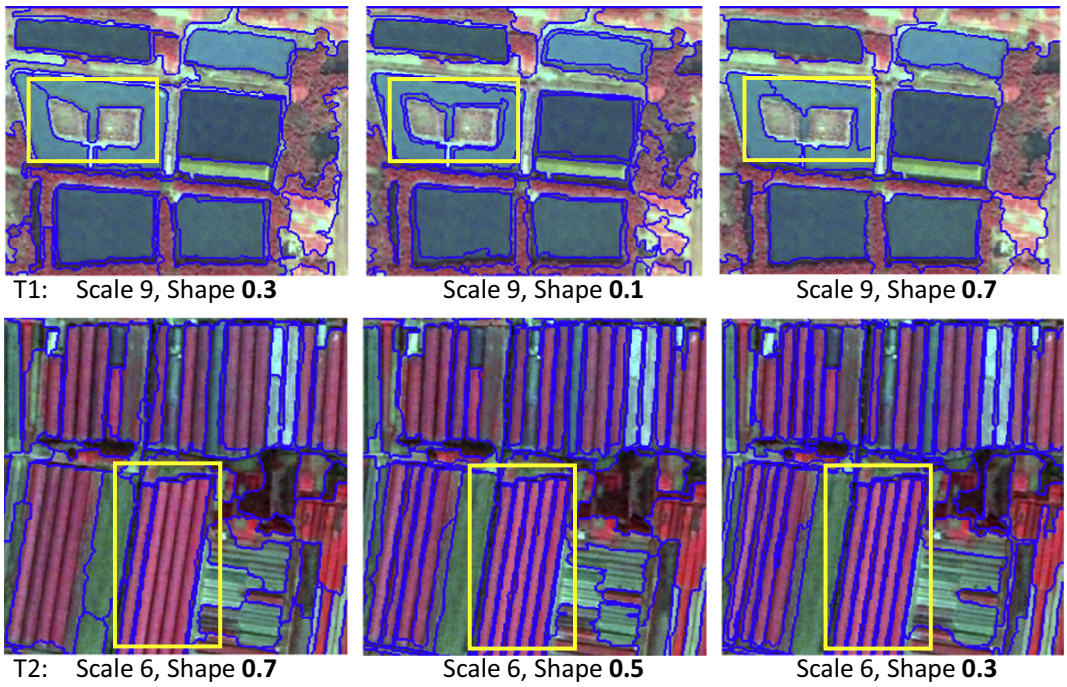

Fig. 10. Subsets of MRS segmentations with different shape parameters for test images T1 and T2. The differences are highlighted by the yellow rectangles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

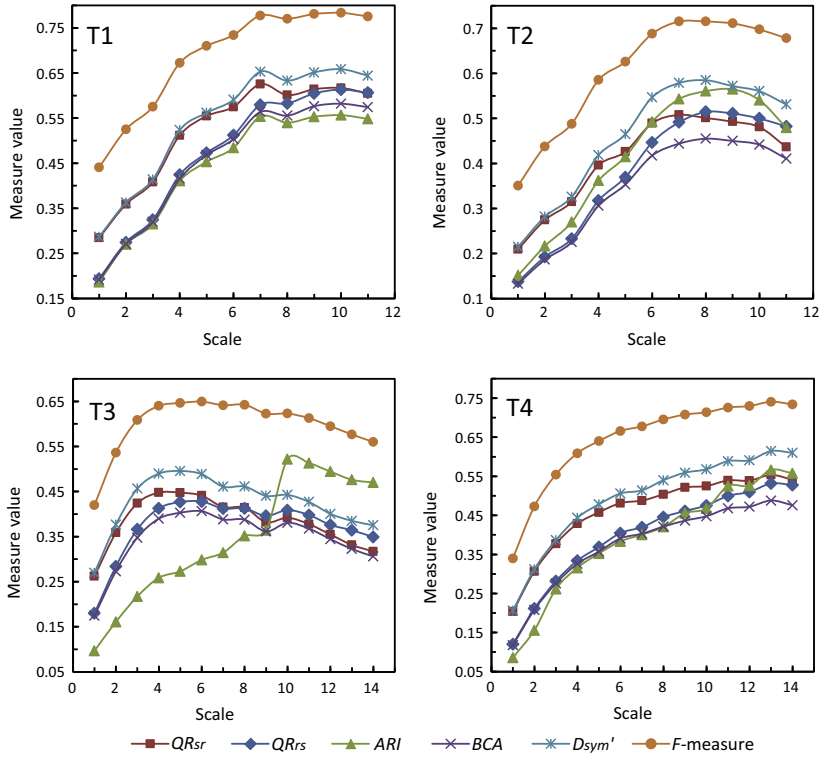

Fig. 11. Comparing the *F*-measure with existing evaluation measures (QR_{sr}, QR_{rs}, D_{sym}', *BCA*, and *ARI*).

and 7, it is interesting to see that the decreasing rate of precision values in Fig. 6 indicates close relation to the change of segmentation quality as scale coarsening in Fig. 7. In Fig. 6, the decreasing rates of precision values for T1 and T4 are relatively slow but the rates for T2 and T3 are large. Accordingly in Fig. 7, the combined indicators show that the segmentation quality of T1 and T4 tends to be improved when the scale is getting coarse, whereas the segmentation quality of T2 and T3 is improved at initial stages but decrease apparently at coarse scales. This reveals that the large decreasing rate of precision values in precision–recall curve indicates the poor performance at coarse segmentation scales.

The precision and recall measures are combined by four different methods to indicate segmentation quality and to capture the trade-off between the two measures. This can resolve the difficulty of comparing two segmentations when the precision and recall values of one segmentation are not both higher or both lower than those of the other. Because of their different isocurve shapes, as shown in Fig. 4, the different combined measures have some inconsistencies. A point located in the upper-right part of an isocurve indicates higher quality than does a point in the bottom left. We analyze the effectiveness of the proposed combined measures in the experiments and the results show that the *F*-measure, *ED′*, and *SUM* measures perform more effectively than does *ED*. The effectiveness of the combined measures is confirmed by sample segmentations and by comparing them with existing discrepancy measures. Based on our analysis in Section 3.2, we suggest using the *F*-measure because it shows the greatest sensitivity to over- and under-segmentation. Moreover, because the combined measures are sensitive to a combination strategy, it deserves to develop other effective combination strategies.

In this study, precision and recall measures are proposed based on a comparison of two partitions. The measures are calculated by summing all regions in the image by means of Eqs. (1) and (2). However, extending these measures to segmentation evaluation tasks based on selected objects is not difficult. In this case, we simply alter Eqs. (1) and (2) by summing the regions that correspond to the selected objects.

The proposed measures focus on single-scale segmentation optimization. However, in many cases multiple scales should be used jointly for image analysis because representing the various objects in high spatial resolution images by a single segmentation is difficult. Some studies examined the matter of optimizing multiscale segmentations through supervised (Trias-Sanz et al., 2008), unsupervised (Drăguţ et al., 2010; Drăguţ et al., 2014), and system-level (Johnson and Xie, 2013) strategies. To apply the proposed measures for evaluating multiscale segmentations, one possible way is to prepare different groups of reference objects at multiple scales (Drăguţ et al., 2014). Then we can optimize the segmentation parameters for each scale in the groups of reference objects. In this case, the measures may be altered so that calculations are based on selected objects. On the other hand, we will focus on extending the proposed measures to multiscale segmentation optimization based on the presentation of different objects at their respective scales in the future.

6. Conclusion

Region-based precision and recall measures have been proposed for evaluating segmentation quality. The two measures are jointly used to reveal segmentation quality by points or curves in a precision–recall space. In addition, to indicate segmentation quality clearly, four combined measures *F*-measure, *SUM*, *ED′* and *ED* are proposed and evaluated according to the shapes of their isocurves. The MRS method embedded in eCognition Developer 8 is adopted for evaluation and a set of high spatial resolution images is used in the experiments to show the effectiveness of the proposed measures. The experimental results show that the precision–recall curve can reflect changes in segmentation scale and show the influence of shape parameters for MRS. The combined measures can clearly reveal segmentation quality when different scale and shape parameters are set. The effectiveness of the combined indicators is further proven by comparing them with four existing evaluation measures. Finally, we recommend combining the precision–recall curves and the *F*-measure for segmentation evaluation. In addition, we suggest using them to compare different segmentation methods for high spatial resolution remote sensing images.

Acknowledgements

This work was supported by the National Basic Research Program of China (Grant No. 2011CB952001), the Qinglan Project of Jiangsu Province, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The authors would like to acknowledge the anonymous reviewers for their constructive suggestions.

References

Albrecht, F., 2010. Uncertainty in image interpretation as reference for accuracy assessment in object-based image analysis. In: 9th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Leicester, United Kingdom.

Baatz, M., Schäpe, M., 2000. Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. Angewandte Geographische Informations – Verarbeitung XII. Wichmann Verlag, Karlsruhe, pp. 12–23.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS J. Photogram. Remote Sens. 58 (3), 239–258.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogram. Remote Sens. 65 (1), 2–16.

Blaschke, T., Strobl, J., 2001. What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. GIS – Z. Geoinform. Syst. 14, 12–17.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., et al., 2014. Geographic object-based image analysis-towards a new paradigm. ISPRS J. Photogram. Remote Sens. 87, 180–191.

Burnett, C., Blaschke, T., 2003. A multi-scale segmentation/object relationship modelling methodology for landscape analysis. Ecol. Model. 168 (3), 233–249.

Cardoso, J.S., Corte-Real, L., 2005. Toward a generic evaluation of image segmentation. IEEE Trans. Image Process. 14 (11), 1773–1782.

Carleer, A.P., Debeir, O., Wolff, E., 2005. Assessment of very high spatial resolution satellite image segmentations. Photogram. Eng. Remote Sens. 71 (11), 1285–1294.

Chabrier, S., Emile, B., Rosenberger, C., Laurent, H., 2006. Unsupervised performance evaluation of image segmentation. EURASIP J. Appl. Sign. Process. 2006, 1–12.

Clinton, N., Holt, A., Scarborough, J., Yan, L., Gong, P., 2010. Accuracy assessment measures for object-based image segmentation goodness. Photogram. Eng. Remote Sens. 76 (3), 289–299.

Corcoran, P., Winstanley, A., Mooney, P., 2010. Segmentation performance evaluation for object-based remotely sensed image analysis. Int. J. Remote Sens. 31 (3), 617–645.

Dey, V., Zhang, Y., Zhong, M., 2010. A review on image segmentation techniques with remote sensing perspective. In: Proceedings of the International Society for Photogrammetry and Remote Sensing Symposium, vol. 38, pp. 31–42.

Drăguţ, L., Tiede, D., Levick, S.R., 2010. ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. Int. J. Geogr. Inform. Sci. 24 (6), 859–871.

Drăguţ, L., Csillik, O., Eisank, C., Tiede, D., 2014. Automated parameterisation for multi-scale image segmentation on multiple layers. ISPRS J. Photogram. Remote Sens. 88, 119–127.

Dronova, I., Gong, P., Clinton, N.E., Wang, L., Fu, W., Qi, S., Liu, Y., 2012. Landscape analysis of wetland plant functional types: the effects of image segmentation scale, vegetation classes and classification methods. Remote Sens. Environ. 127, 357–369.

Espindola, G.M., Câmara, G., Reis, I.A., Bins, L.S., Monteiro, A.M., 2006. Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation. Int. J. Remote Sens. 27 (14), 3035–3040.

Estrada, F.J., Jepson, A.D., 2009. Benchmarking image segmentation algorithms. Int. J. Comput. Vision 85 (2), 167–181.

Faur, D., Gavat, I., Datcu, M., 2009. Salient remote sensing image segmentation based on rate-distortion measure. IEEE Geosci. Remote Sens. Lett. 6 (4), 855–859.

Gao, Y., Mas, J.F., Kerle, N., Navarrete Pacheco, J.A., 2011. Optimal region growing segmentation and its effect on classification accuracy. Int. J. Remote Sens. 32 (13), 3747–3763.

Hay, G.J., Castilla, G., 2008. Geographic Object-Based Image Analysis (GEOBIA): a new name for a new discipline. Object-Based Image Analysis. Springer, Berlin Heidelberg, pp. 75–89.

Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P.J., Bunke, H., Goldgof, D.B., Bowyer, K., Eggert, D.W., Fitzgibbon, A., Fisher, R.B., 1996. An experimental comparison of range image segmentation algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 18 (7), 673–689.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2 (1), 193–218.

Johnson, B., Xie, Z., 2011. Unsupervised image segmentation evaluation and refinement using a multi-scale approach. ISPRS J. Photogram. Remote Sens. 66 (4), 473–483.

Johnson, B., Xie, Z., 2013. Classifying a high resolution image of an urban area using super-object information. ISPRS J. Photogram. Remote Sens. 83, 40–49.

Laliberte, A.S., Rango, A., 2009. Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery. IEEE Trans. Geosci. Remote Sens. 47 (3), 761–770.

Levine, M.D., Nazif, A.M., 1982. An experimental rule based system for testing low level segmentation strategies. In: Preston, K., Uhr, L. (Eds.), Multicomputers and Image Processing: Algorithms and Programs. Academic Press, New York, pp. 149–160.

Levine, M.D., Nazif, A.M., 1985. Dynamic measurement of computer generated image segmentations. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-7 (2), 155–164.

Liu, Y., Bian, L., Meng, Y., Wang, H., Zhang, S., Yang, Y., Shao, X., Wang, B., 2012. Discrepancy measures for selecting optimal combination of parameter values in object-based image analysis. ISPRS J. Photogram. Remote Sens. 68, 144–156.

Lucieer, A., 2004. Uncertainties in Segmentation and their Visualisation. Doctoral Dissertation, Utrecht University and International Institute for Geo-Information Science and Earth Observation (ITC).

Marpu, P.R., Neubert, M., Herold, H., Niemeyer, I., 2010. Enhanced evaluation of image segmentation results. J. Spatial Sci. 55 (1), 55–68.

Martin, D.R., 2003. An Empirical Approach to Grouping and Segmentation. Doctoral Dissertation, Computer Science Division, University of California, Berkeley.

Martin, D.R., Fowlkes, C.C., Malik, J., 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. 26 (5), 530–549.

Möller, M., Lymburner, L., Volk, M., 2007. The comparison index: a tool for assessing the accuracy of image segmentation. Int. J. Appl. Earth Obs. Geoinform. 9 (3), 311–321.

Neubert, M., Herold, H., Meinel, G., 2008. Assessing image segmentation quality – concepts, methods and application. Object-Based Image Analysis. Springer, Berlin Heidelberg, pp. 760–784.

Persello, C., Bruzzone, L., 2010. A novel protocol for accuracy assessment in classification of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48 (3), 1232–1244.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66 (336), 846–850.

Räsänen, A., Rusanen, A., Kuitunen, M., Lensu, A., 2013. What makes segmentation good? A case study in boreal forest habitat mapping. Int. J. Remote Sens. 34 (23), 8603–8627.

Schiewe, J., Tufte, L., Ehlers, M., 2001. Potential and problems of multi-scale segmentation methods in remote sensing. GIS – Z. Geoinform. Syst. 6, 34–39.

Smith, A., 2010. Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. J. Spatial Sci. 55 (1), 69–79.

Stein, A., De Beurs, K., 2005. Complexity metrics to quantify semantic accuracy in segmented Landsat images. Int. J. Remote Sens. 26 (14), 2937–2951.

Tian, J., Chen, D.M., 2007. Optimization in multi-scale segmentation of high-resolution satellite images for artificial feature recognition. Int. J. Remote Sens. 28 (20), 4625–4644.

Trias-Sanz, R., Stamon, G., Louchet, J., 2008. Using colour, texture, and hierarchical segmentation for high-resolution remote sensing. ISPRS J. Photogram. Remote Sens. 63 (2), 156–168.

Unnikrishnan, R., Pantofaru, C., Hebert, M., 2007. Toward objective evaluation of image segmentation algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 929–944.

Van Rijsbergen, C.J., 1979. Information Retrieval, second ed. Butterworths, London.

Weidner, U., 2008. Contribution to the assessment of segmentation quality for remote sensing applications. Int. Arch. Photogram. Remote Sens. Spatial Inform. Sci. 37 (B7), 479–484.

Witharana, C., Civco, D.L., 2014. Optimizing multi-resolution segmentation scale using empirical methods: exploring the sensitivity of the supervised discrepancy measure Euclidean distance 2 (ED2). ISPRS J. Photogram. Remote Sens. 87, 108–121.

Witharana, C., Civco, D.L., Meyer, T.H., 2014. Evaluation of data fusion and image segmentation in earth observation based rapid mapping workflows. ISPRS J. Photogram. Remote Sens. 87, 1–18.

Xiao, P., Feng, X., An, R., Zhao, S., 2010. Segmentation of multispectral high-resolution satellite imagery using log Gabor filters. Int. J. Remote Sens. 31 (6), 1427–1439.

Yang, L., Albregtsen, F., Lønnestad, T., Grøttum, P., 1995. A supervised approach to the evaluation of image segmentation methods. In: Proceedings of 6th International Conference: Computer Analysis of Images and Patterns, pp. 759–765.

Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., Schirokauer, D., 2006. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. Photogram. Eng. Remote Sens. 72, 799–811.

Zhan, Q., Molenaar, M., Tempfli, K., Shi, W., 2005. Quality assessment for geo-spatial objects derived from remotely sensed data. Int. J. Remote Sens. 26 (14), 2953–2974.

Zhang, Y.J., 1996. A survey on evaluation methods for image segmentation. Pattern Recogn. 29 (8), 1335–1346.

Zhang, Y., 2002. Problems in the fusion of commercial high-resolution satellite images as well as Landsat 7 images and initial solutions. Int. Arch. Photogram. Remote Sens. Spatial Inform. Sci. 34 (4), 587–592.

Zhang, H., Fritts, J.E., Goldman, S.A., 2008. Image segmentation evaluation: a survey of unsupervised methods. Comput. Vis. Image Underst. 110 (2), 260–280.

Zhang, X., Xiao, P., Feng, X., 2012. An unsupervised evaluation method for remotely sensed imagery segmentation. IEEE Geosci. Remote Sens. Lett. 9 (2), 156–160.