AN IMPROVED UNSUPERVISED IMAGE SEGMENTATION EVALUATION APPROACH BASED ON UNDER- AND OVER-SEGMENTATION AWARE

Tengfei Su [1], *

[1] Inner Mongolia Agricultural University, College of Water Conservancy and Civil Engineering, 010018 Inner Mongolia Hohhot, China - stf1987@126.com

Commission ICWG II/III

KEY WORDS: Image Segmentation Evaluation, Unsupervised, Under- and Over-segmentation Aware, Over-Segmentation Error, Under-Segmentation Error, Edge Strength

ABSTRACT:

In this paper, an unsupervised evaluation scheme for remote sensing image segmentation is developed. Based on a method called under- and over-segmentation aware (UOA), the new approach is improved by overcoming the defect in the part of estimating over-segmentation error. Two cases of such error-prone defect are listed, and edge strength is employed to devise a solution to this issue. Two subsets of high resolution remote sensing images were used to test the proposed algorithm, and the experimental results indicate its superior performance, which is attributed to its improved OSE detection model.

1. INTRODUCTION

Image segmentation has been considered as a key step in object-based image analysis (OBIA), since it has a significant impact on classification performance (Blaschke et al. 2014; Su and Zhang 2017). Accordingly, many studies attempted to improve the quality of segmentation approaches (Troya-Galvis et al. 2015; Su and Zhang 2017). However, it is not an easy task, not only because in remote sensing images, there are miscellaneous geo-objects of different sizes and various spectral signatures, which complicates the design of segmentation algorithms, but also because it is often a difficult problem to precisely and objectively evaluate the performance of segmentation results. Although there exists a number of segmentation evaluation metrics, they all have their merits and demerits which deserve to be analyzed.

The existent segmentation evaluation methods can be broadly categorized into two types: qualitative and quantitative approaches. The former mainly relies on visual inspection, which is simple to perform, but it cannot provide quantitative score and may suffer from subjectivity (Su and Zhang 2017). On the contrary, quantitative methods can produce explicit scoress to reflect the quality of a segmentation result. Thus, quantitative methods are more prevalently used in segmentation-related studies.

Generally, quantitative approaches can also be classified as two sub-types: supervised and unsupervised strategies. Their primary difference resides in whether ground truth is adopted. The former needs ground truth to yield evaluation scores, and such methods are frequently employed in related researches, since it can accurately determine which segmentation result is the best by directly comparing the resulted scores. Although supervised approaches are more popular, it is still inappropriate to conclude that unsupervised methods are useless. In fact, in terms of operational application, ground truth is often hard and expensive to obtain, and unsupervised methods are therefore useful to help tune the segmentation parameter(s), such as scale, which has been considered to be the most influential one in multi-scale segmentation approaches (Johnson and Xie 2011; Yang et al. 2015).

During recent years, many studies related to optimal scale estimation by means of unsupervised metrics have been reported. Johnson and Xie (2011) used regional spectral variance and spatial auto-correlation to determine the optimal scale. This method was further improved by Böck et al. (2017), leading to stabilized solution for the estimation of optimal scale. Spectral angle was also utilized (Yang et al. 2014; Yang et al. 2015) to identify the most meaningful scale parameter(s). Moreover, an interesting method called under- and over-segmentation aware (UOA) was proposed by Troya-Galvis et al. (2015), which, in addition to the capacity of estimating optimal scale, was able to explicitly determine whether a segment is over- or under-segmented. This approach is also extendable, since its homogeneity criteria can be adaptively modified according to user's specific needs. However, there is a defect in its design, which may result in over-estimation of over-segmentation error. Accordingly, this work aims to improve this method by correcting this shortcoming.

2. METHOD

2.1 The original UOA

Consider a segment S_i that may not accurately correspond to a real geo-object in a remotely sensed scene, it can be assumed that such a segment tends to contain under-segmentation error (USE) if its within-segment heterogeneity is very large, conversely, it may have over-segmentation error (OSE) if its intra-segment heterogeneity is very low and most importantly, there exists a neighbouring segment S_j, whose union with S_i is also of very low heterogeneity. Based on this assumption, UOA provides the following model to locally evaluate the segmentation error type of a segment S_i:

* Corresponding author

$$E_\delta(S_i) = \begin{cases} -1 & \text{if } H(S_i) > \delta \\ 1 & \text{if } H(S_i) < \delta \text{ and } \exists\, S_j \in N(S_i) \wedge H(S_i \cup S_j) < \delta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $E_\delta(S_i)$ represents the segmentation error type of segment S_i, and it indicates USE, OSE, or not an error when its value is -1, 1, 0, respectively; $H(\cdot)$ is heterogeneity measure that can be adaptively designed; δ symbolizes a heterogeneity threshold, and if $H(S_i) > \delta$, S_i is considered to be too heterogeneous; $N(S_i)$ means a set of segments that are neighbours of S_i.

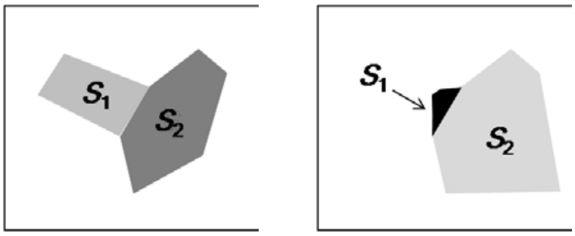
It is worth mentioning that there are various ways to design $H(\cdot)$. However, this study focuses on unveiling and correcting the defect of OSE identification detailed in the following sub-sections, and thus intra-segment spectral variance, which has been tested to be simple but effective (Troya-Galvis et al. 2015), was adopted to model $H(\cdot)$:

$$H(S_i) = \frac{1}{B}\sum_b^B \sigma_b \tag{2}$$

where B represents the number of bands; σ_b is the spectral standard deviation of bth band of the segment under processing.

2.2 The defect of UOA

To clearly explain the UOA defect in its OSE measurement strategy, Fig. 1(a) is firstly used as an illustration, where there are two spectrally similar segments. Suppose that the two segments well represent two real geo-objects. In this situation, the E_δ value of both segments should not be 1. However, if the spectral appearances of S_1 and S_2 are similar enough, making $H(S_1 \cup S_2) < \delta$, then according to equation (1), the E_δ value of S_1 and/or S_2 will be mistakenly calculated as 1, leading to over-estimation of OSE. This kind of problem can be quite evident for images capturing rural landscape, where there are often many adjacent croplands having similar spectra, despite that there exist a clear boundary between them.



(a) (b)

Fig. 1 Two examples that over-estimation of OSE may easily occur. (a) Two segments have similar spectra; (b) The sizes of two segments are very unequal

In addition, this type of error tends to take place for the neighbouring segments whose sizes are very different. Such a case can be observed in Fig. 1(b), in which a small geo-object and a large one are well singled out as S_1 and S_2, respectively. Although the spectral signatures of the two segments are very different, the size of S_1 is quite smaller than S_2, likely resulting in $H(S_1 \cup S_2) < \delta$, and consequently OSE is erroneously computed for S_1 and/or S_2.

2.3 The improved UOA

In order to remove the defect of original UOA, edge strength is used as an additional metric. The method is simple: if the segment S_i under processing is homogeneous enough ($H(S_i) < \delta$), then find all of its neighboring segments which meet the condition: $H(S_i \cup S_j) < \delta$. and put those segments into a segment collection N_H. Next, if the cardinality of N_H is not 0, and one of its element S_j has weak edge strength at the common boundary with S_i, then it is determined that the $E_\delta(S_i)$=1, indicating that S_i contains OSE. This strategy is capable of avoiding the error in the two cases introduced in Fig. 1, because there are clear edges dividing S_1 and S_2 in both sub-figures.

In the aforementioned approach, it is a key issue to identify whether there exist a weak or strong edge at the common boundary between two segments. Two steps are used to solve it. First, an edge strength map is produced. For the convenience of computation, it is suggested that the pixel values in edge strength map should be within the scope of [0,1], and higher values indicate more prominent edge. Second, equation (3) is exploited for weak edge detection:

$$M_{\mathrm{ES}}(S_i,S_j) = \begin{cases} 1 & \text{if } \left(1/\left|C(S_i,S_j)\right|\right) \cdot \sum C_{\mathrm{ES}}(S_i,S_j) < \gamma \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $M_{\mathrm{ES}}(S_i,S_j)$ is weak edge indicator, and 1 indicates that there exist a weak boundary between two neighboring segments S_i and S_j, otherwise a strong/evident boundary is detected; $C(S_i,S_j)$ represents the common boundary of two segments S_i and S_j, and $|C(S_i,S_j)|$ equals the length of that common boundary; $C_{\mathrm{ES}}(S_i,S_j)$ is a set of pixel values of the edge strength map, and the pixels are all at the common boundary of S_i and S_j; γ is a threshold with numerical range of (0,1), and higher it is, more weak edges tend to be detected.

From the above description of the proposed approach, it can be seen that there are totally two parameters, δ and γ. The former is the only parameter of the original UOA, while the latter is introduced by this work. In the experiment of this paper, γ has been extensively analyzed to show its influences on the performance of the proposed method.

3. EXPERIMENT SETUP

3.1 Dataset

Two scenes of high spatial resolution remote sensing images acquired by GaoFen-2 satellite have been adopted to test the improved UOA. The two images both have spatial resolution of 3.24 meters, and both include four channels: near infrared (NIR), red, green and blue bands. Due to that the spatial range of the two images is too large, which is difficult for detailed analysis, a small subset with width and height both equal to 400 pixels was extracted for each dataset. The subset of the first scene (T1) covers agricultural area, and its acquisition date is 15th Feb 2015. Its central coordinate of is 30.69904°N, 114.17555°E. The second subset (T2) with central coordinate of 29.62997°N, 91.00786°E, captures urban area, and it was acquired on 10th Oct 2014. The first row of Fig 2 shows these two subsets.

To enable supervised quantitative evaluation in this experiment, ground truth segmentation is required. 52 and 53 reference geo-objects were manually and accurately digitized for T1 and T2, respectively, as can be seen in the middle row of Fig. 2. These reference geo-objects have various sizes and spectral features. In addition, since the proposed method needs edge strength map to detect OSE, the vector-gradient-based approach used by Su et al. (2015) was employed for this purpose, and the results can be seen in the last row of Fig. 2, where the boundaries between various geo-objects are well represented.

3.2 Experimental setup

The experiment includes two parts: 1) analysis of γ, 2) comparison to other supervised and unsupervised evaluation methods. In the first part, a series of segmentation results were produced by using the frequently used algorithm in OBIA, multi-resolution segmentation (MRS) (Baatz and Schäpe 2000). The shape and compactness coefficients of this method were set as 0.1 and 0.5, respectively. A range of scale parameters from 10 to 500 with 10 as interval were used, and as a result, 50 different segmentation results for T1 or T2 were derived. Note that with the increase of scale, OSE becomes less apparent while USE is more dominant. By applying the proposed method to the set of segmentation results, the effects of γ on the evaluation performance can be analyzed.

The second part aims to compare the proposed approach to different evaluation strategies. To achieve this goal, the original UOA, as well as two newly developed methods for global segmentation error assessment (Böck et al. 2017; Su and Zhang 2017) were implemented and adopted. The segmentation results generated in the first part were also used here, and the ability of the three methods to identify OSE and USE was analyzed and discussed.

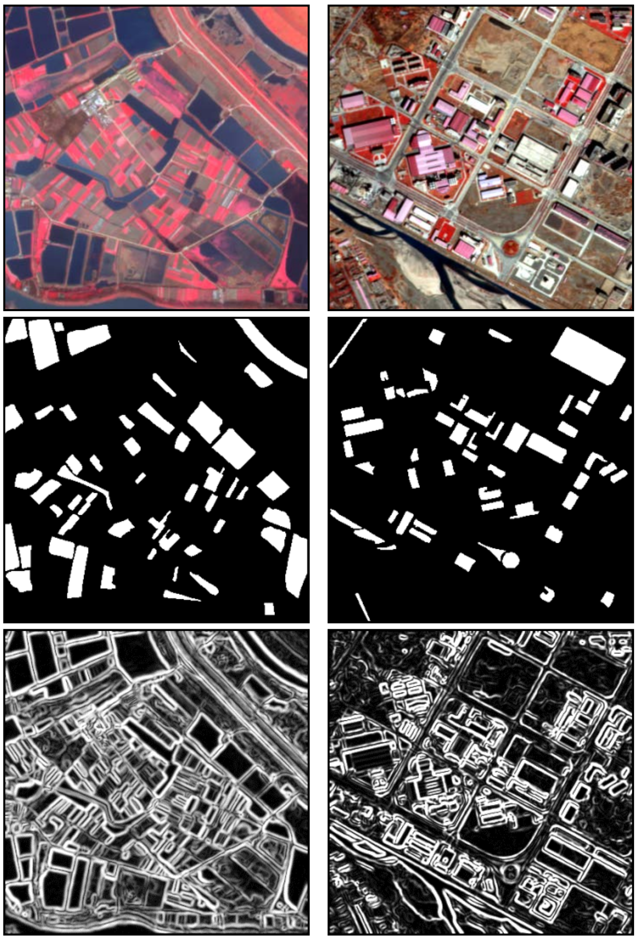


Fig. 2 The dataset used in this experiment. From left to right, the first and second column correspond to T1 and T2, respectively. From top to bottom, the first row shows the original images with color composition of R: near infrared, G: red, B: green; the second row demonstrates the ground truth geo-objects extracted by experts; the third row displays the edge strength maps with 0 value (black) indicating no edge and 1 value (white) representing edge pixels.

4. RESULTS

4.1 Parameter analysis for γ

Visual inspection is firstly conducted to observe the behaviour of γ for some segmentation results with different error patterns. For this purpose, Fig. 3 and Fig. 4 are provided. As can be seen in the left column of these two figures, the segmentation results of scale 10 exhibit evident over-segmentation error (OSE), in which almost all of the geo-objects are fragmented as small segments. On the contrary, the scale of 110 leads to obvious under-segmentation error (USE) for most geo-objects in T1 and T2, especially for the small crop fields in T1 and the small buildings in T2. Comparatively, for both subsets, the scale of 50 generates more accurate segmentation results than 10 and 110, although OSE and USE still exist, such as for T1, some large aquaculture pools being fragmented, and a few crop fields erroneously merged with adjacent ones.

For the aforementioned three different segmentation results for each dataset, three values of γ (0.1, 0.5, 0.9) were used to perform the proposed UOA, leading to nine evaluation results for T1 or T2, as shown in the right three columns of Fig. 3 and Fig 4.

From these results of T1, it is straightforward to see that with the rise of γ, more segments are assigned as containing OSE. This is easy to explain: according to equation (3), higher γ allows more weak edges to be detected, thus leading to more OSEs to be determined. For scale of 10, when γ=0.1, only some geo-objects of large size, i.e., aquaculture pools, are correctly evaluated as OSE, while most other geo-objects are considered to have no error (gray colored). When γ=0.9, it seems that OSEs are excessively identified, many small crop fields which are rightfully segmented out are mistakenly assigned as white (OSE). In comparison, γ=0.5 results in the optimal evaluation, which best reflects the segmentation error types. Similar patterns of γ can be found for the evaluation results of scale 50 and 110.

The results of T2 as illustrated in Fig. 4 demonstrate quite similar pattern as compared to those in Fig. 3, which further verifies that higher values of γ tend to result in more OSE segments.

In addition, the metric of θ, which is a global indicator of OSE as developed in the original UOA (Troya-Galvis et al. 2015), is plotted against γ to reflect more analytical information. Such results are shown in Fig. 5, where it can be seen that for both T1 and T2, with γ rising, θ becomes larger for the segmentation results of various scales. It is interesting to note that, in both cases of T1 and T2, when scale=300, the θ curve is overlapped with the horizontal axis, indicating that no OSE can be spotted.

However, there exist differences between Fig. 5(a) and (b). For all of the curves of various scales, those for T1 have slower changing rate than those of T2 when γ is below 0.5. This indicates that when γ is greater than 0.5, T1's evaluation result is more sensitive to γ than that of T2. This may be attributed to the landscape discrepancy of the two subsets. In T1, there are many small crop fields and bare soil lands surrounded by relatively weak edges, as can be seen in the last row of Fig. 2, thus varying the value of γ will obviously affect the OSE identification result of those geo-objects. Contrarily, T2 contains much less weak-edge geo-objects, and most ones such as buildings and bare soil lands are surrounded by roads, which

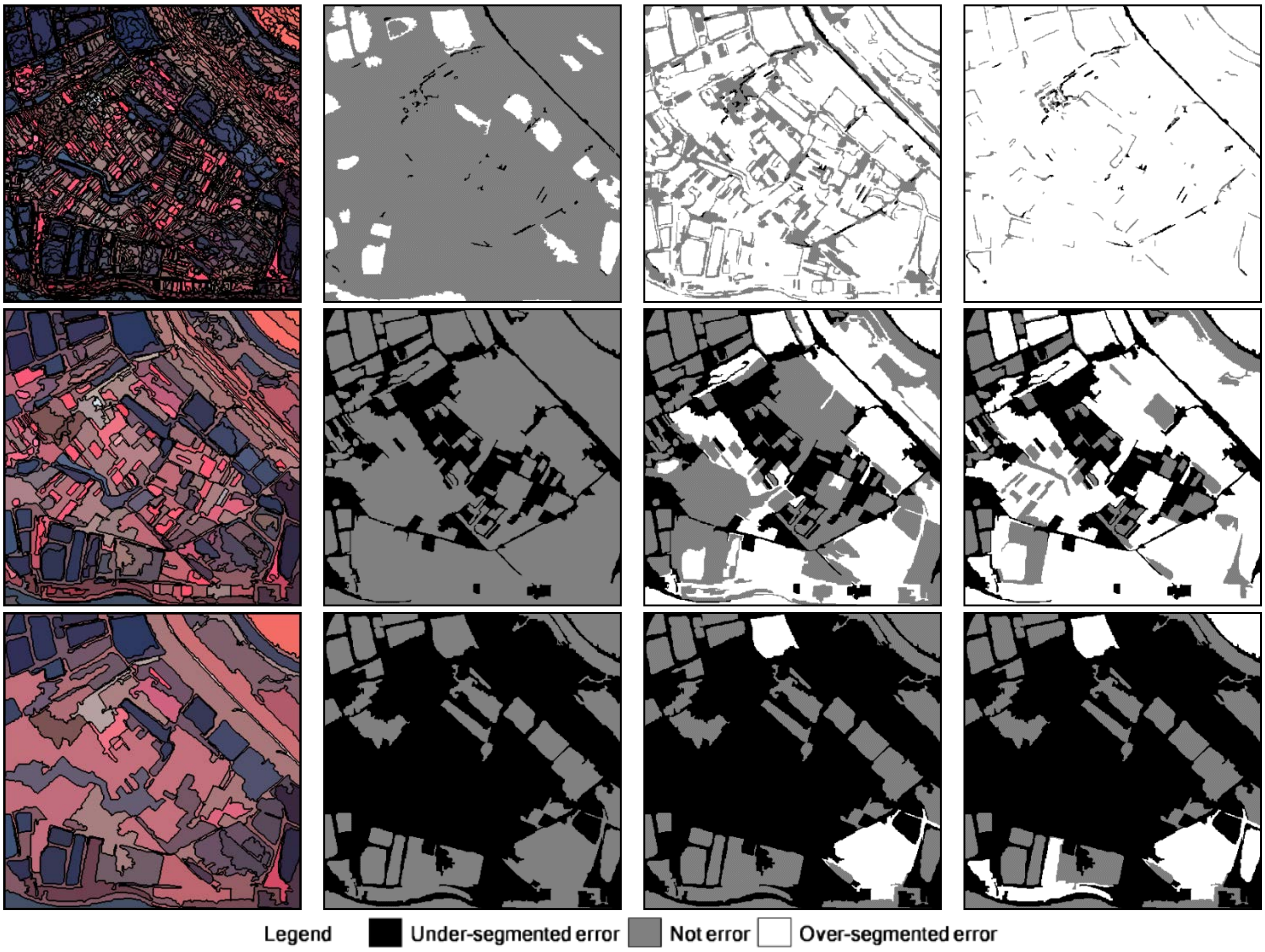


Fig. 3 T1's analytical results produced by using the proposed UOA method parameterized by different γs. From top to bottom, the first, second and third row correspond to segmentation results of scale 10, 50 and 110, respectively. From left to right, the first column represents segmentation results, and the second, third, and fourth column are evaluation results produced by using γ=0.1, γ=0.5, and γ=0.9, respectively.

	The improved UOA	The original UOA	JX	SZ
T1	60	70	280	30
T2	60	70	380	40

Table 1. Optimal scales determined by different methods

represent strong cues of edge feature. Accordingly the θ value of T2 is less sensitive to γ when this parameter is large enough.

According to the results of this sub-section, it can be understood that γ has a serious influence on the evaluation effects of the proposed method. It is also suggested that γ should not be valued too low nor too high, since the two cases may lead to under- or over- estimation of OSE, respectively. As a consequence, γ=0.5 is used in the following experiment due to its sufficiently good performance reflected in this sub-section.

4.2 Comparative study

To fully evaluate the proposed method, three different methods are adopted for comparative experiment, including the original UOA, a popular unsupervised evaluation approach developed by Johnson and Xie (JX) (Johnson and Xie 2011), and a recently proposed supervised evaluation strategy of Su and Zhang (2017) (SZ). In light of the fact that the proposed method is an improved version of the original UOA, a direct comparison of them is firstly carried out, which is shown in Fig. 6 for T1 and Fig. 8 for T2.

Fig. 6 illustrates T1's segmentation and evaluation results of three scales which are determined as optimal by the three different methods. The optimal scales determined by different evaluation approaches are listed in Table 1, according to the curves shown in Fig. 7 and Fig. 9. More details on optimal scale selection are described in the subsequent paragraph. By comparing the figures of the middle and the right column of Fig. 6 and Fig. 8, it is apparent that the original UOA results in more OSEs than its counterpart. In the segmentation result of scale 60 for T1, most of aquaculture pools and bare soil lands are measured to be OSE according to the original UOA, which is not very consistent with the actual situation. The improved UOA method, on the other hand, reflects the error pattern more accurately: for T1, the aquaculture pool situated at the right bottom corner of the test image is successfully singled out by using scale 60, thus its color should be gray and the proposed method has correctly evaluated this segment; however, the original UOA assigns white to it, indicating its inferior performance of OSE identification. As for the results for T2, the superiority of the improved UOA is not as conspicuous as that for T1. In Fig. 8, the proposed UOA only correctly identifies the OSE of a few more bare soil lands and small buildings than the old UOA, probably because the average size of the geo-objects in T2 is apparently larger than those in T1, thus when the same scales are used for segmentation, more OSEs tend to be produced for T2.

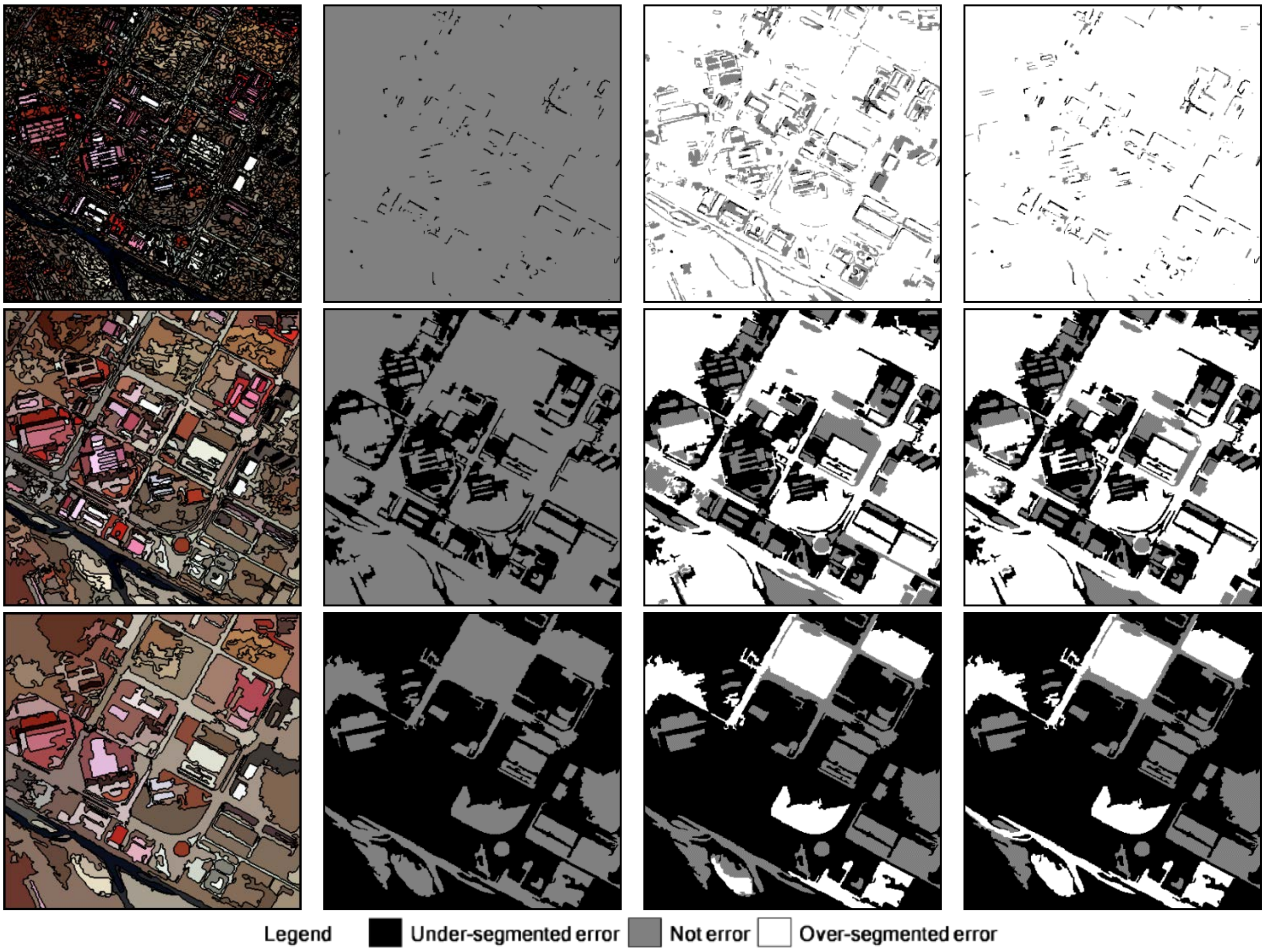


Fig. 4 T2's analytical results produced by using the proposed UOA method parameterized by different *γ*s. From top to bottom, the first, second and third row correspond to segmentation results of scale 10, 50 and 110, respectively. From left to right, the first column represents segmentation results, and the second, third, and fourth column are evaluation results produced by using *γ*=0.1, *γ*=0.5, and *γ*=0.9, respectively.

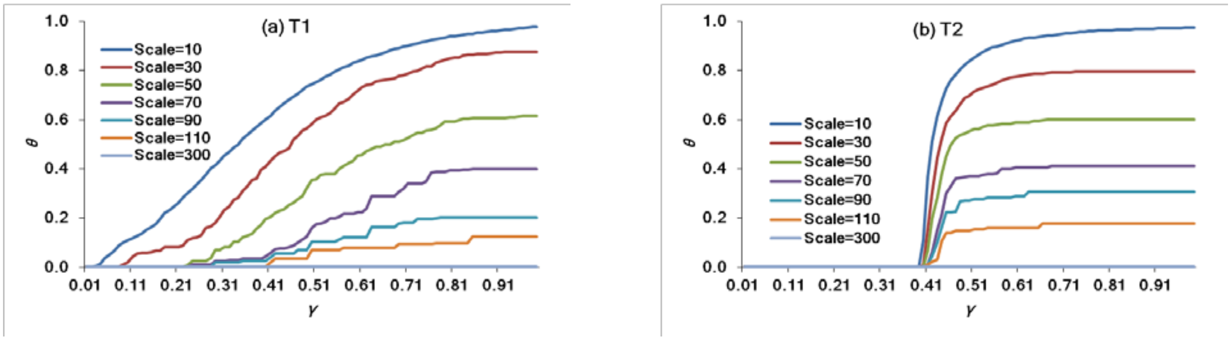


Fig. 5 Variation of OSE (*θ*) with the increase of *γ* for a set of segmentation results. (a) and (b) correspond to T1 and T2, respectively.

In addition to the visual comparison of the two UOA methods, comparative study is further conducted by using the aforementioned approaches. In this experiment, global evaluation scores of the four methods (the original UOA, JX, SZ, and the proposed UOA) are plotted. For the two UOA strategies, four scores including φ, θ, Σ, and L2 are adopted. φ and θ respectively correspond to global area-weighted aggregation of USE and OSE, while Σ and L2 are both global evaluation scores considering both OSE and USE, their difference can be seen in (Troya-Galvis et al. 2015). Note that the optimal scale determined by UOA has a Σ value closest to 0 and has a L2 value that is the minimum. As for the JX method, an improved version that has been recently proposed (Böck et al. 2017) is implemented and used, and it consists of three scores: weighted variance (WV), Moran's index (MI), and goodness score (GS). The optimal scale selected by JX corresponds to the lowest GS. Different form the other three approaches, SZ is supervised, and its global scores (GOSE, GUSE and F1 measure) are calculated by using the ground truth segmentation as shown in Fig. 2(b). According to this method, the optimal scale corresponds to the segmentation result with the highest F1 score.

Fig. 7 and Fig. 9 display the plots of global scores of the four schemes for T1 and T2, respectively. It is interesting to note that four different optimal scales are identified by the four methods, as listed in Table 1. For both test images, the two UOA and the

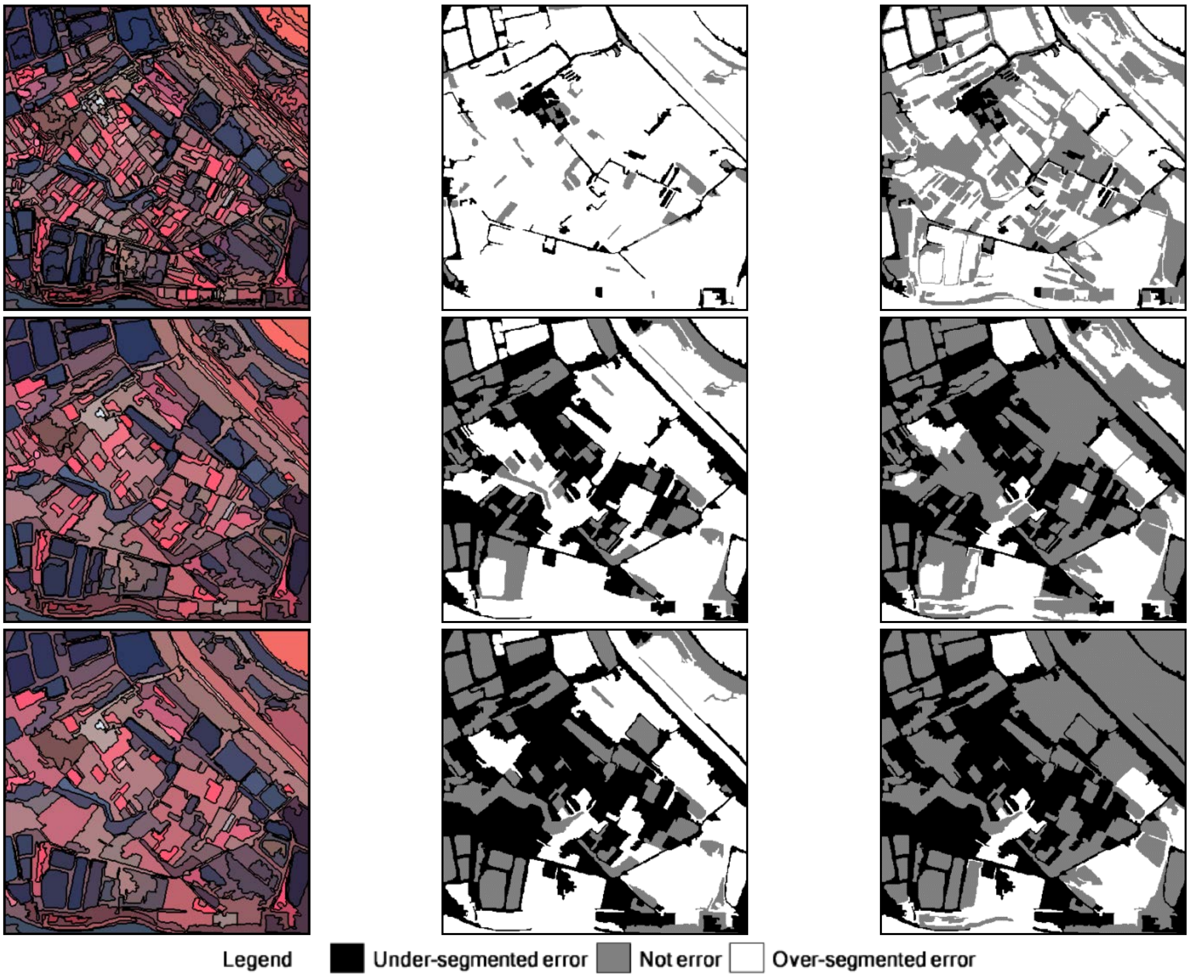
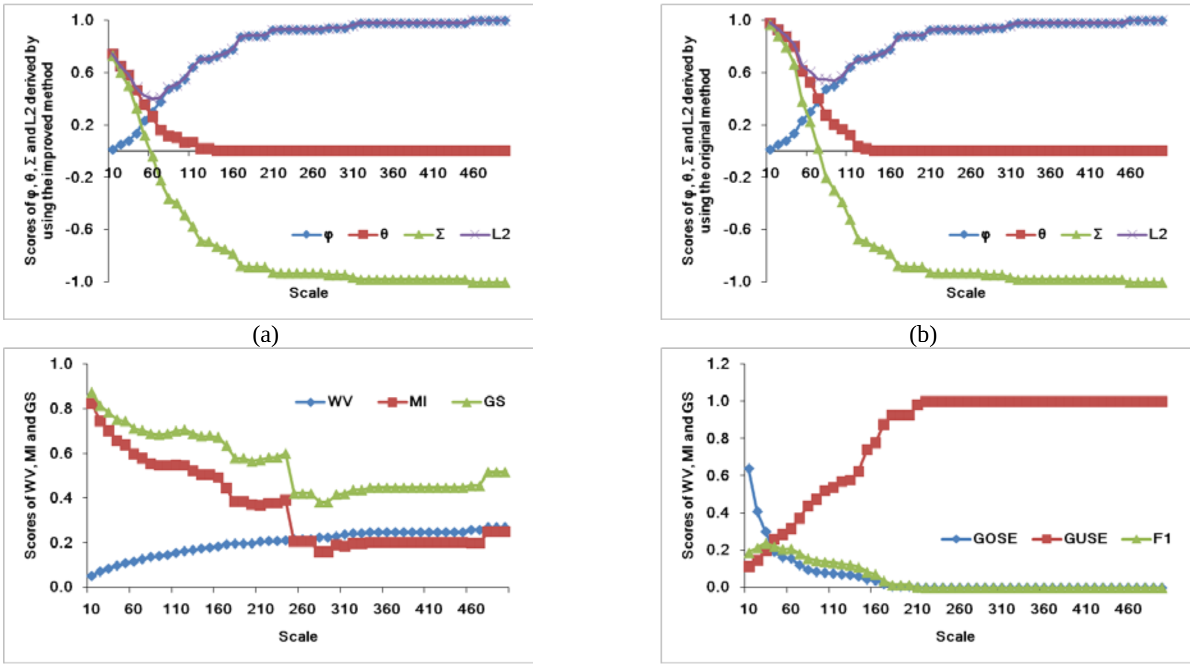


Fig. 6 T1's evaluation results produced by using the original and the newly developed UOA approaches. From top to bottom, the first, second and third row correspond to segmentation results of 30, 60 and 70, respectively. From left to right, the first column shows the segmentation results, while the second and third column are evaluation results generated by employing the original UOA and the proposed UOA, respectively.



(a)

(b)

(c)

(d)

Fig. 7 T1's variation of global evaluation scores of four different methods with the increase of scale. (a) The proposed method, (b) the original UOA, (c) the JX unsupervised evaluation method, (d) the SZ supervised evaluation method.

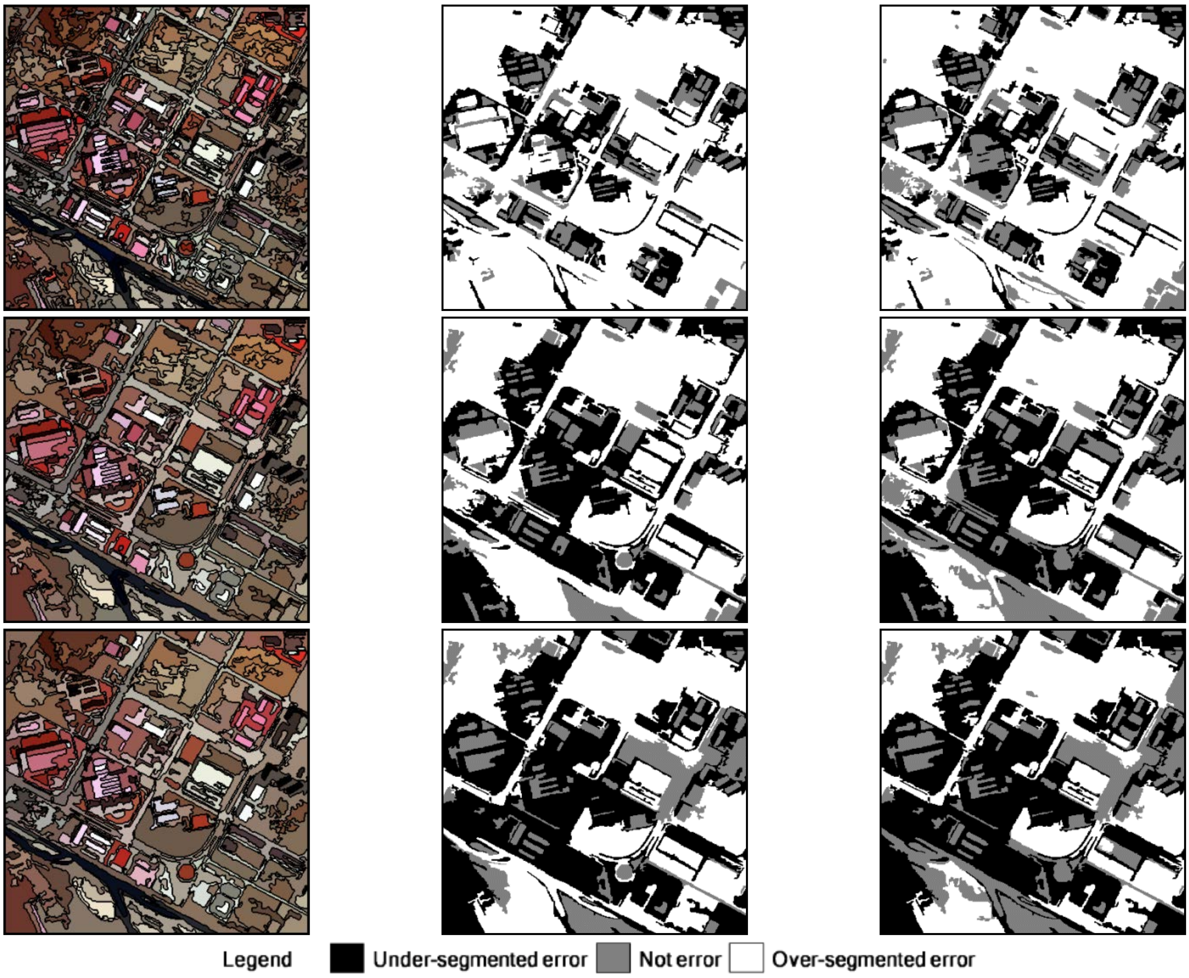
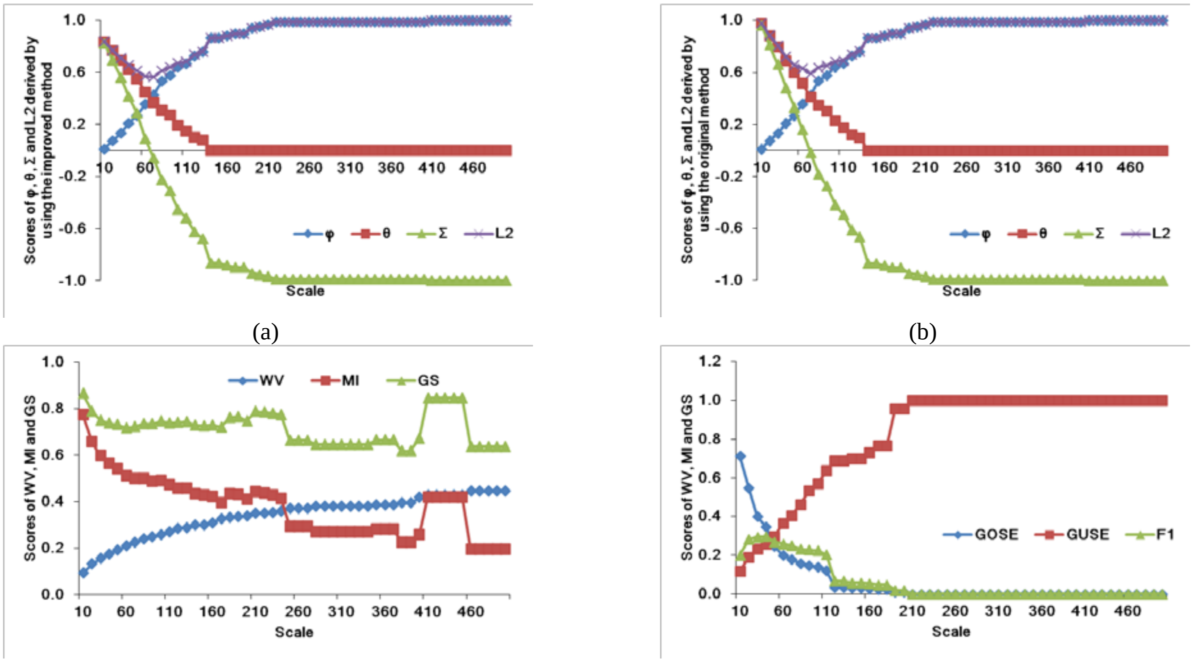


Fig. 8 T2's evaluation results produced by using the original and the newly developed UOA approaches. From top to bottom, the first, second and third row correspond to segmentation results of 40, 60 and 70, respectively. From left to right, the first column shows the segmentation results, while the second and third column are evaluation results generated by employing the original UOA and the proposed UOA, respectively.



(a)

(b)

(c)

(d)

Fig. 9 T2's variation of global evaluation scores of four different methods with the increase of scale. (a) The proposed method, (b) the original UOA, (c) the JX unsupervised evaluation method, (d) the SZ supervised evaluation method.

supervised strategies have relatively similar results, while JX's results are far more deviant. This may be due to the large difference in numerical range of weighted variance and global Moran's index. It is also worth noting that SZ produced the lowest optimal scale for both T1 and T2, which is probably due to the fact that most small area geo-objects are selected as reference segments. Since the improved UOA has lower extent of over-estimating OSE, the optimal scale selected by this method is smaller, which is nearer to the one identified by the supervised SZ method. From this point, the proposed UOA is more accurate than the old version.

5. CONCLUSION

This paper presents an unsupervised segmentation evaluation algorithm, which is mainly based on a method called under- and over-segmentation aware (UOA). The proposed approach is superior to the original UOA in terms of over-segmentation error (OSE) identification. Edge strength is utilized to avoid over-estimation of OSE. Two subsets of high spatial resolution remote sensing images were adopted for the experiment of segmentation evaluation. The experimental results indicates that 1) the parameter γ should not be set too low or too high, and 0.5 is good enough for the test images used in this work; 2) the new UOA is superior to the old version since it can identify OSE with higher precision.

In future studies, the proposed method will be applied to remote sensing images of other types, to further test its performance. Moreover, efforts may be made to integrate the new UOA with some segmentation algorithms, with an aim of enhancing segmentation accuracy and automation.

ACKNOWLEDGEMENT

This work has been funded by national natural science foundation of China under grant of 61701265.

REFERENCES

Baatz, M., Schäpe, M., 2000. Multiresolution segmentation - an optimization approach for high quality multi-scale image segmentation. *Angewandte Geographische InformationsVerarbeitung XII.* pp. 12-23.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., vander Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* 87, pp. 180-191.

Böck, S., Immitzer, M., Atzberger, C., 2017. On the objectivity of the objective function—problems with unsupervised segmentation evaluation based on global score and a possible remedy. *Remote Sens.* 9, pp. 1-9.

Johnson, B., Xie, Z., 2011. Unsupervised image segmentation evaluation and refinement using a multi-scale approach. *ISPRS J. Photogramm. Remote Sens.* 66, pp. 473-483.

Su, T., Li, H., Zhang, S., Li, Y., 2015. Image segmentation using mean shift for extracting croplands from high-resolution remote sensing imagery. *Remote Sens. Lett.* 6(12), pp. 952-961.

Su, T., Zhang, S., 2017. Local and global evaluation for remote sensing image segmentation. *ISPRS J. Photogramm. Remote Sens.* 130, pp. 256-276.

Troya-Galvis, A., Gançarski, P., Passat, N., Berti-Équille, L., 2015. Unsupervised quantification of under- and over-segmentation for object-based remote sensing image analysis. *IEEE J. Sele. Top. Appl. Earth Obs. Remote Sens.* 8(5), pp. 1936-1945.

Yang, J., He, Y., Weng Q., 2015. An automated method to parameterize segmentation scale by enhancing intrasegment homogeneity and intersegment heterogeneity. *IEEE Geosci. Remote Sens. Lett.* 12(6), pp. 1282-1286.

Yang, J., Li, P., He, Y., 2014. A multi-band approach to unsupervised scale parameter selection for multi-scale image segmentation. *ISPRS J. Photogramm. Remote Sens.* 94, pp. 13-24.