Supervised methods of image segmentation accuracy assessment in land cover mapping

Hugo Costa, Giles M. Foody, and Doreen S. Boyd

School of Geography, University of Nottingham, Nottingham NG7 2RD, UK

Abstract

Land cover mapping via image classification is sometimes realized through object-based image analysis. Objects are typically constructed by partitioning imagery into spatially contiguous groups of pixels through image segmentation and used as the basic spatial unit of analysis. As it is typically desirable to know the accuracy with which the objects have been delimited prior to undertaking the classification, numerous methods have been used for accuracy assessment. This paper reviews the state-of-the-art of image segmentation accuracy assessment in land cover mapping applications. First the literature published in three major remote sensing journals during 2014-2015 is reviewed to provide an overview of the field. This revealed that qualitative assessment based on visual interpretation was a widely-used method, but a range of quantitative approaches is available. In particular, the empirical discrepancy or supervised methods that use reference data for assessment are thoroughly reviewed as they were the most frequently used approach in the literature surveyed. Supervised methods are grouped into two main categories, geometric and non-geometric, and are translated here to a common notation which enables them to be coherently and unambiguously described. Some key considerations on method selection for land cover mapping applications are provided, and some research needs are discussed.

Keywords: OBIA, GEOBIA, empirical goodness methods, quality, classification

1 Introduction

Land cover mapping is a very common application of remote sensing and has been increasingly conducted through object-based image analysis (Blaschke, 2010). Object-based image analysis has been described as an advantageous alternative to conventional per-pixel image classification, and adopted in a diverse range of studies (Bradley, 2014; Feizizadeh et al., 2017; Matikainen et al., 2017; Strasser and Lang, 2015).

Objects are typically discrete and mutually exclusive groups of neighbouring pixels and used as the basic spatial unit of analysis. Objects may be delimited or obtained via a range of sources (e.g. cadastral data), but typically are constructed through an image segmentation analysis, and thus often called segments. In this paper the terms “object” and “segment” are used synonymously. Image segmentation is performed by algorithms with the purpose of constructing objects corresponding to geographical features distinguishable in the remotely sensed data, which may be useful for applications such as land cover mapping.

Constructing objects poses a set of challenges. For example, it is necessary to select a segmentation algorithm from the numerous options available, but comparative studies (e.g. Basaeed et al., 2016; Neubert et al., 2008) are uncommon. Also each of the segmentation algorithms is typically able to produce a vast number of outputs depending on the parameter settings used. Selecting the most appropriate segmentation is, therefore, difficult.

Multiple methods have been proposed to assess the accuracy of an image segmentation and are normally grouped in two main categories: empirical discrepancy and empirical goodness methods, also commonly referred to as supervised and unsupervised methods respectively (Zhang, 1996). Most of the supervised methods essentially compare a segmentation output to a reference data set and measure the similarity or discrepancy between the two representations (e.g. overlapping area) (Clinton et al., 2010). Unsupervised methods measure some desirable properties of the segmentation outputs (e.g. object's spectral homogeneity), thus measuring their quality (Zhang et al., 2008).

There is no standard approach for image segmentation accuracy assessment, and some studies have compared accuracy assessment methods. Supervised and unsupervised methods are normally compared separately. For example, with regard to supervised methods, Clinton et al. (2010), Räsänen et al. (2013), and Whiteside et al. (2014) compared dozens of methods, all of them focused on some geometric property of the objects, such as positional accuracy relative to the reference data. These and other studies highlight the differences and similarities obtained from the methods compared so the reader gains a perspective of the field. However, many other supervised methods have been proposed yet are barely compared against previous counterparts; these tend to be newly proposed methods (e.g. Costa et al., 2015; Liu and Xia, 2010; Marpu et al., 2010; Su and Zhang, 2017). Furthermore, the methods are often described using a notation suitable for the specific case under discussion, which makes the cross-comparison of methods difficult.

Studies like Clinton et al. (2010) are valuable in reviewing the field of image segmentation accuracy assessment, but they often focus on the geometry of the objects evaluated and ignore that a supervised but non-geometric approach may be followed (e.g. Wang et al. 2004). Moreover, supervised methods are typically compared within a specific study case without discussion of further and important issues, such as the suitability of the methods as a function of context. As image segmentation is increasingly used in a wide range of applications, the behaviour and utility of specific methods is expected to vary in each case. Thus, selecting a method to assess the accuracy of image segmentation may be based on an incomplete understanding of the available options and ultimately problematic.

This paper reviews the state-of-the-art of image segmentation accuracy assessment in land cover mapping applications. The literature published in three major remote sensing journals in 2014-2015 is reviewed to provide an overview of the field, namely the methods used and their popularity. In particular, the supervised methods are thoroughly reviewed as they are widely used. A comprehensive description of which supervised methods are available is presented with the aim of providing a basis on which the remote sensing community may consider and select a suitable method for particular applications. A discussion on which methods should be used is provided, and research needs are highlighted.

2 Background

Image objects are typically expected to delimit features of the Earth's surface such as land cover patches that are remotely sensed using an air/spaceborne imaging system. Image segmentation cannot, however, deliver results exactly according to the desired outcome for multiple reasons, such as unsuitable definition of segmentation algorithm parameter settings, and insufficient spectral and spatial resolution of the data. Thus, image segmentation error is common, namely under- and over-segmentation. Under-segmentation error occurs when image segmentation fails to define individual objects to represent different contiguous land cover classes, thus constructing a single object that may contain more than one land cover class. On the contrary, over-segmentation error occurs when unnecessary boundaries are delimited, and thus multiple contiguous objects, potentially of the same land cover class, are formed.

Segmentation errors have been traditionally identified through visual inspection, but it has some drawbacks, especially when assessing large areas and comparing numerous segmentation outputs. Specifically, visual interpretation is time consuming, subjective, and the results produced by the same or different operators may not be reproducible (Coillie et al., 2014; Lang et al., 2010). As a result, objective and quantitative methods for the assessment of image segmentation accuracy may be necessary and have become more popular in recent years.

The literature published during 2014-2015 in three remote sensing journals was reviewed to provide an overview of the state-of-the-art of image segmentation accuracy assessment. The journals were *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and *Remote Sensing Letters*. These journals were selected to represent the variety of current publication outlets in the field. Historically, the former journal has had the greatest impact factor among the remote sensing journals. The second journal has been particularly active in publishing papers on object-based image analysis. The latter journal is a relatively young journal dedicated to rapid publications. The papers that included specific terms (namely “obia”, “geobia”, “object-based”, and “object-oriented”) in the title, abstract, and key words were retained for analysis. A total of 55 out of 67 papers that matched the search terms were identified as relevant, each describing techniques for estimating objects which were used as the basic spatial unit in land cover mapping applications.

These 55 papers were analysed, and it was noticeable that 17 papers (30.9%) do not document if or how the accuracy of the image segmentation outputs was assessed. This shows that image segmentation accuracy assessment is often overlooked as an important component of an image segmentation analysis protocol. It is speculated that visual interpretation was used in most of the cases that provide no information accuracy, as having used no sophisticated method may reduce any motivation for documenting the topic. The remaining 38 papers explicitly described the methods used, and often more than one method was adopted. Visual interpretation was widely used, with 15 papers (25.3% of the total of papers) describing that the qualitative appearance of the segmentations influenced the assessment of the results (e.g. Qi et al., 2015). Details were typically not given, such as the time dedicated to visual interpretation and number of interpreters.

When a quantitative alternative to subjective visual interpretation was explicitly adopted, the methods used varied widely. A rudimentary strategy of assessing the accuracy of image segmentations, and used in five papers (9.1%), was to use simple descriptive statistics, such as the average of some attributes of the objects like area, to get an impression of the segmentation output. The statistics were used in a supervised or unsupervised fashion. In the former situation, the statistics were compared to the statistics of a reference data set depicting desired polygonal shapes, and small differences were regarded as indicative of large segmentation accuracy (e.g. Liu et al., 2015). When no reference data were used (i.e. unsupervised fashion), the statistics identified the image segmentation from the set obtained with the most desirable properties, such as a target mean size (i.e. area) of the objects (Hultquist et al., 2014). Although descriptive statistics can measure some quantitative properties of an image segmentation, they provide a very limited sense of the accuracy of the objects, for example in the spatial domain, and here they are not regarded as a true accuracy assessment method. The latter are typically more evolved and normally grouped into supervised and unsupervised methods.

Supervised methods were found in 21 (38.2%) of the papers reviewed (e.g. Zhang et al., 2014). Although there was no dominant method, the Area Fit Index (Lucieer and Stein, 2002) and Euclidean distance 2 (Liu et al., 2012) were the supervised methods that were most used with three appearances each (Belgiu and Drăguţ, 2014; Drăguţ et al., 2014; Witharana et al., 2014; Witharana and Civco, 2014; Yang et al., 2014). Many of the other methods identified were used only once (e.g. Carleer et al. 2005). These and other supervised methods are, however, thoroughly described in the next section. Unsupervised methods were applied in 13 (23.6%) of the papers surveyed (e.g. Robson et al., 2015). The unsupervised method most used in the literature reviewed was the Estimation of Scale Parameter (ESP or ESP2) tool (Drăguţ et al., 2014, 2010) available in the popular eCognition software. The segmentation algorithms available in this software were used in most of the papers surveyed (36 papers, 65.5%) to construct image objects.

Object-based image analysis has received much attention and acceptance (Blaschke et al., 2014; Dronova, 2015), but the accuracy assessment of image segmentation, which is a central stage of the analysis, appears to be in a relatively early stage of maturation. Although procedures for image segmentation accuracy assessments have not been standardized, a more harmonized approach is desirable. Using subjective visual interpretation may be acceptable and suitable for some applications; the reasons are seldom explained in the literature. Among the quantitative methods proposed for image segmentation accuracy assessment, supervised approaches seem to be the most frequently adopted, hence reviewed hereafter.

3 Supervised methods

Supervised methods for image segmentation accuracy assessment use reference data to estimate the accuracy of the objects constructed. Often the reference data are formed by polygons extracted from the remotely sensed data in use (e.g. based on visual interpretation) or collected externally (e.g. a field boundary map). Approaches for assessing accuracy based on reference data are herein grouped into two main categories: geometric and non-geometric. Geometric methods are the most widely used and typically focus on the geometry of the objects and polygons to determine the level of similarity among them. Ideally, there should be no difference among objects and polygons in terms of area, position, and shape. Note that the land cover class(es) associated with the objects and polygons typically need not be known.

With non-geometric methods the land cover class(es) associated with the objects must be known, and reference data polygons are not always used. The properties of the objects such as the spectral content are used in a variety of ways, depending on the specific method. Ideally, the content of the objects representing different land cover classes should be as different as possible. When polygons are also used, the content of objects and polygons representing the same land cover class should be identical. Note that the spatial or geometric correspondence between objects and polygons need not be known. Fuller details on both geometric and non-geometric approaches are given in the sub-sections that follow. Rudimentary strategies (for example used in 9.1% of the papers reviewed in the previous section) are not covered however.

3.1 Geometric methods

Geometric methods rely on quantitative metrics that describe aspects of the geometric correspondence between objects and polygons, often based on difference in area and position (Winter, 2000). Figure 1 illustrates a typical case involving an object and polygon for which the larger the overlapping area and/or the shorter the distance between their centroids, the larger the accuracy with which the object has been delimited.

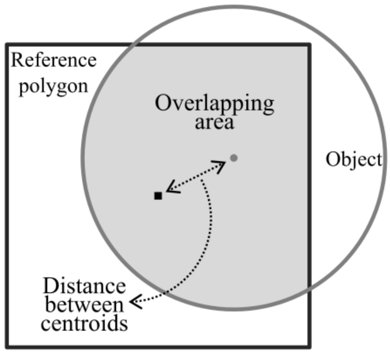

Figure 1. Geometric comparison between an object and polygon based on the overlapping area (shaded area) and/or distance between centroids (dashed arrow).

3.1.1 Notation

Notation is necessary to assist the description of the metrics used by geometric methods. The notation presented hereafter uses that defined in Clinton et al. (2010). Therefore, the notation is transcribed below together with additional elements necessary to describe all the methods covered.

The *m* objects constructed via image segmentation are denoted by y_j (j=1, …, *m*), the *n* polygons forming a reference data set by x_i (i=1, …, *n*), and the *l* pixels of the segmented remotely sensed data by z_p (p=1, …, *l*). They define the following sets:

- $X=\{x_i: i=1, \ldots, n\}$ is the set of *n* polygons (Figure 2a)
- $Y=\{y_j: j=1, \ldots, m\}$ is the set of *m* objects of a segmentation output (Figure 2b)
- $S=X\cap Y=\{s_{ijk}: \text{area}(x_i\cap y_j)\neq 0\}$ is the set of *s* intersection objects that result from the spatial intersection (represented by symbol ∩) of X and Y; s_{ijk} is the kth object that results from the spatial intersection of the ith polygon (x_i) with the jth object (y_j) (Figure 2c)
- $Z=\{z_p: p=1,\ldots,l\}$ is the set of *l* pixels of the segmented remotely sensed data.

Set S is the result of a spatial intersection of X and Y, which can be defined using common geographical information systems. Note that the subscript k is needed to create a unique symbol as the overlay of x_i and y_i can yield more than one discontinuous polygonal area (x_1 and y_6 in Figure 2). Set Z is simply the set of pixels that form the remotely sensed data submitted to segmentation analysis, but its definition is nevertheless useful for describing clearly some metrics.

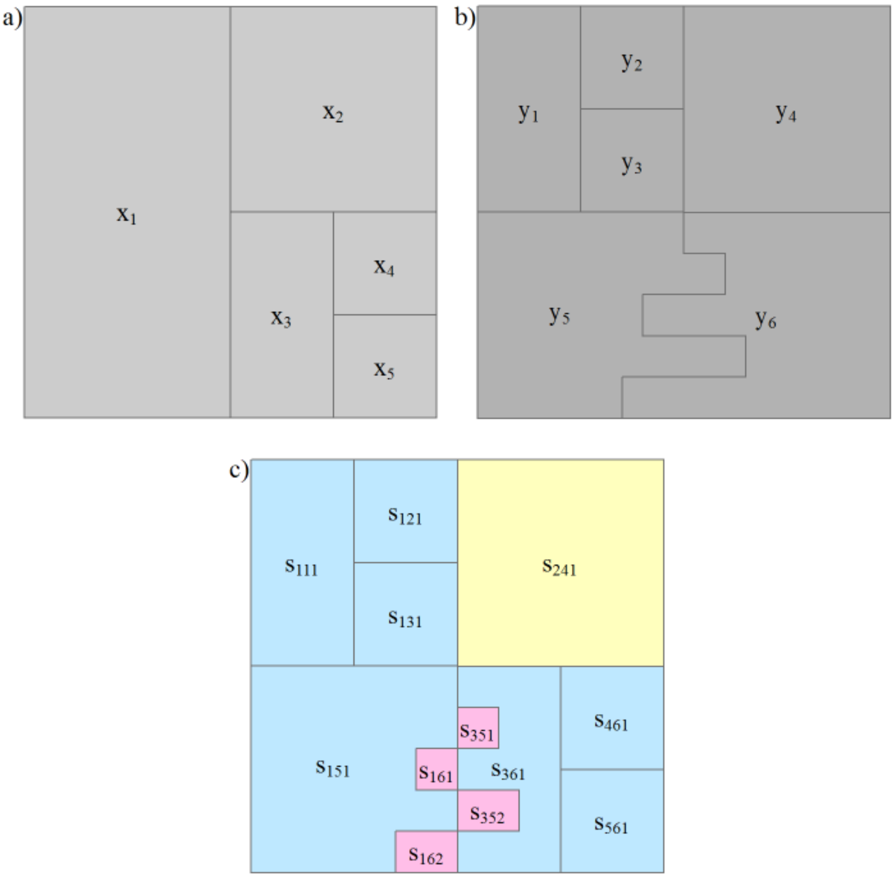

Figure 2. Sets X, Y, and S: (a) reference set X, (b) segmentation Y, and (c) intersection S=X∩Y. In (c) yellow denotes one-to-one, blue denotes one-to-many, and pink denotes many-to-many (Section 3.1.1.3).

The description of the methods also requires the use of symbols that characterize the sets X, Y, S, and Z, and their members. For example, size() denotes the number of an item identified in brackets, for example the number of objects that belong to Y – size(Y) – or the number of pixels of an object – size(y_j); and dist() is the distance between two items identified in brackets, for example the centroids of y_j and x_i – dist(centroid(x_i), centroid(y_j)). This basic notation is used to express more complex cases. For example, area($x_i \cap y_j$) is the area of the geographical intersection of polygon x_i and object y_j. Other self-explanatory cases are used in the notation adopted. Furthermore, mathematical symbols are also used, such as ¬ which is the logical negation symbol and read as “not”, \ which is the complement symbol used in set theory and reads as “minus” or “without”, and ∪ which is the union symbol.

Subsets of X, Y, and S must be defined to assist the description of methods that follow four different strategies: (i) Y is compared to X, (ii) X is compared to Y, (iii) S is compared to both X and Y, and (iv) X and Y are compared to Z. In all of the cases, the definition of subsets of X, Y, and S are used to decide which polygons x_i, objects y_j, and intersection objects s_{ijk} corresponds to each other or to pixel z_p, which is central to the calculation of geometric metrics (presented in Section 3.1.2).

3.1.1.1 Set Y compared to set X

In image segmentation accuracy assessment most often the set Y is compared to set X. This strategy typically involves the calculation of geometric metrics for the members of X, and thus there is the need to identify which member(s) of Y correspond to each member of X. For example, Figure 3a shows the set of objects that overlap and thus can be considered as corresponding to a polygon x_i. The specific objects that are actually considered as corresponding depends on the method used, and the calculations related to each polygon x_i consider only the objects regarded as corresponding. Thus, it is useful to define the following subsets of Y for each member of X:

- $\tilde{Y}_i$ is the subset of Y such that $\tilde{Y}_i=\{y_j: \text{area}(x_i \cap y_j) \neq 0\}$
- Ya_i is a subset of $\tilde{Y}_i$ such that $Ya_i=\{y_j: \text{the centroid of } x_i \text{ is in } y_j\}$
- Yb_i is a subset of $\tilde{Y}_i$ such that $Yb_i=\{y_j: \text{the centroid of } y_j \text{ is in } x_i\}$
- Yc_i is a subset of $\tilde{Y}_i$ such that $Yc_i=\{y_j: \text{area}(x_i \cap y_j)/\text{area}(y_j)>0.5\}$
- Yd_i is a subset of $\tilde{Y}_i$ such that $Yd_i=\{y_j: \text{area}(x_i \cap y_j)/\text{area}(x_i)>0.5\}$
- Ye_i is a subset of $\tilde{Y}_i$ such that $Ye_i=\{y_j: \text{area}(x_i \cap y_j)/\text{area}(y_j)=1\}$
- Yf_i is a subset of $\tilde{Y}_i$ such that $Yf_i=\{y_j: \text{area}(x_i \cap y_j)/\text{area}(y_j)>0.55\}$
- Yg_i is a subset of $\tilde{Y}_i$ such that $Yg_i=\{y_j: \text{area}(x_i \cap y_j)/\text{area}(y_j)>0.75\}$
- $Y_i^* = Ya_i \cup Yb_i \cup Yc_i \cup Yd_i$
- Y_i' is a subset of $\tilde{Y}_i$ such that $Y_i'=\{y_j: \max(\text{area}(x_i \cap y_j))\}$.

The definition of subsets of Y expresses the variety of criteria of correspondence that has been used. For example, some methods require the centroid of the objects to fall inside the polygons, and Yb_i denotes the set of objects whose centroid falls inside a specific polygon x_i. However, most of the criteria of correspondence used define a threshold of overlapping area between polygons and objects. For example, at least half of the object's area may have to overlap a polygon for a positive correspondence to be considered; Yc_i denotes the set of objects that comply with this criterion for a specific polygon x_i. The selection of a specific subset of Y depends on the method used.

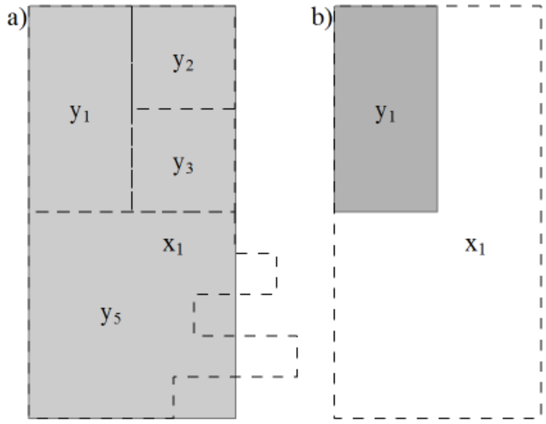

Figure 3. Comparison between X and Y of Figure 2: (a) four potential objects (dashed lines) corresponding to polygon x_1 (grey background) when Y compares to X; (b) one potential polygon (dashed line) corresponding to object y_1 (grey background) when X compares to Y.

3.1.1.2 Set X compared to set Y

When set X is compared to Y, geometric metrics are calculated for the members of Y, and thus there is the need to identify which member(s) of X correspond to each member of Y. For example, Figure 3b shows that one polygon overlap and thus can be considered as corresponding to an object y_j. The calculations related to each object y_j consider only the polygons regarded as corresponding, depending on the method used. Thus, it is useful to define the following subsets of X for each member of Y:

- $\tilde{X}_j$ is the subset of X such that $\tilde{X}_j=\{x_i: area(y_j \cap x_i) \neq 0\}$
- Xc_j is a subset of $\tilde{X}_j$ such that $Xc_j=\{x_i: area(y_j \cap x_i)/area(y_j)>0.5\}$
- X'_j is a subset of $\tilde{X}_j$ such that $X'_j=\{x_i: \max(area(y_j \cap x_i))\}$
- X''_j is a subset of $\tilde{X}_j$ such that $X''_j=\{x_i: \max(area(y_j \cap x_i)/area(y_j \cup x_i))\}$.

The subsets of X defined above represent the criteria of correspondence that have been used when X is compared to Y. All the criteria define a threshold of overlapping area between polygons and objects. For example, a polygon may have to overlap more than half of the object's area for a positive correspondence between objects and polygons; Xc_i denotes the set of polygons that comply with this criterion for a specific object y_j. The selection of a specific subset of X depends on the method used.

To describe two particular methods found in the literature (Costa et al., 2015; Liu and Xia, 2010), it is useful to define X not as the set of *n* reference polygons, but the set of *t* thematic classes represented in X. For example, if x_3 and x_4 in Figure 2 are two polygons representing the same thematic class, c_i, and both intersect the same object, y_6, notation like $area(c_i \cap y_6)$ can be used, where $area(c_i)=area(x_3 \cup x_4)$. Thus, similarly to above:

- $C=\{c_i: i=1, \ldots, t\}$ is the set of *t* thematic classes represented in X; classes c_i can also be denoted as d_i as it is useful to describe a specific method (Costa et al., 2015).

When comparing C to Y, the following subset of C is identified for each y_j:

- $\tilde{C}_j$ is the subset of C such that $\tilde{C}_j=\{c_i: area(c_i \cap y_j) \neq 0\}$.

3.1.1.3 Set S compared to both sets X and Y

When set S is compared to both sets X and Y, three types of hierarchical relations between polygons and objects emerge. The three types are one-to-one, one-to-many, and many-to-many relations (Figure 2). The first type occurs when x_i and y_j match perfectly. One-to-many relations occur when x_i intersects several objects or *vice-versa*. Many-to-many relations occur when several discontinuous intersection objects correspond to a same x_i and y_j (e.g. sliver intersection objects s_{ijk} along the edges of x_i and y_j).

Given the three types of hierarchical object relations, the following subsets of S are defined:

- $S_1=\{s_{ijk}: area(x_i \cap y_j) = area(x_i \cup y_j)\}$ is the subset of all one-to-one objects
- $S_{2a}=\{s_{ijk}: (\text{one } x_i \cap \text{many } y_j) \vee (\text{many } x_i \cap \text{one } y_j)\}$ is the subset of all one-to-many relations
- $S_{2b}=\{s_{ijk}: (\text{one } x_i \cap \text{many } y_j) \vee (\text{many } x_i \cap \text{one } y_j); \max(area(s_{ijk}))\}$ is the subset of the largest one-to-many relations
- $S_3=\{s_{ijk}: \text{one } x_i \cap \text{one } y_j \text{ over discontinuous areas}; \max(area(s_{ijk}))\}$ is the subset of the largest many-to-many relations.

Based on the above subsets, it is useful to define the subsets $Sa=S_1 \cup S_{2a} \cup S_3$, and $Sb=S_1 \cup S_{2b} \cup S_3$. Finally, subsets of Sa and Sb are defined for each x_i and y_j:

- $Sax_i=\{s_{ijk}: area(s_{ijk}\cap x_i)\neq 0\}$
- $Say_j=\{s_{ijk}: area(s_{ijk}\cap y_j)\neq 0\}$
- $Sbx_i=\{s_{ijk}: area(s_{ijk}\cap x_i)\neq 0\}$
- $Sby_j=\{s_{ijk}: area(s_{ijk}\cap y_j)\neq 0\}$

The definition of subsets Sax_i, Say_j, Sbx_i, and Sby_j are used in Möller et al. (2013) and Costa et al. (2015).

3.1.1.4 Sets X and Y compared to set Z

To describe two particular methods found in the literature (Martin, 2003; Zhang et al., 2015a, 2015b), it is useful to consider the assessment framework at the pixel level and thus define the following subsets of X and Y that correspond to each member of Z:

- Xa_p is the subset of X such that $Xa_p=\{x_i$: the centroid of z_p is in $x_i\}$
- Ya_p is the subset of Y such that $Ya_p=\{y_j$: the centroid of z_p is in $y_j\}$.

3.1.2 Available metrics

Geometric metrics are presented in Table 1 using the notation defined above, except four cases that would require the definition of unnecessarily complex notation, and thus are described as text (metrics 6, 7, 13 and 28). The metrics express the fundamental calculation involving objects and polygons; each object, polygon, or intersection object receives a metric value, which will tell something about the individual geometric accuracy of the objects constructed. Assessing each areal entity individually is often referred to as local evaluation or validation (Möller et al., 2013, 2007; Persello and Bruzzone, 2010). The subscripts i and j used in the name of the metrics in Table 1 (e.g. $Precision_{ij}$) indicate that the metrics are calculated for the local level. These subscripts come from those used to identify the specific polygon x_i and object y_j involved in the calculations.

Place table 1 near here. **See Table 1 after the references.**

Local metric values are commonly aggregated in a variety of ways to produce a single value to express the accuracy of a segmentation output as a whole. This is often referred to as global evaluation or validation (Möller et al., 2013, 2007; Persello and Bruzzone, 2010). Table 1 provides details on how the local metric values are aggregated for the global level in the column headed Notes. Typically, the local values are summed or averaged in either one or two steps, which in Clinton et al. (2010) is referred to as weighted and unweighted measures respectively. In the first case, all the local values are aggregated in a straightforward fashion (e.g. SimSize, metric 15). In the second case, the aggregation is undertaken first for each individual polygon or object (depending of the strategy of comparison), and then for the whole segmentation. For example, metric PI_{ij} (metric 22) is first aggregated for each polygon, and then for the whole segmentation. Therefore, if for a given polygon, say x_1, there are two corresponding objects, y_1 and y_2, then PI_{11} and PI_{12} are calculated according to metric 22. Then, PI_{11} and PI_{12} are summed to calculate a single PI_1 value for polygon x_1. This produces n PI_i values (one for each polygon x_i). Finally, the n PI_i values can be averaged to express image segmentation accuracy as a whole, denoted as PI (without any subscript).

Showing the metrics for the local level facilitates comparison, but it was not possible to write them all in the same style. For example, the LP_i formula (metric 31) shows only the subscript i (i.e. the subscript j is missing). This specific metric, calculated for polygons x_i, needs immediately to involve all the corresponding objects. In other cases, such as NSR (metric 39), the metric's name in Table 1 shows no subscripts because the metric is calculated directly as a global value for the whole segmentation output.

Oftentimes the purpose of calculating metrics, such as those of Table 1, is to combine them later for the definition of further metrics. These are hereafter referred to as combined metrics (Table 2). Several approaches have been proposed to combine geometric metrics, such as metrics sum, and root mean square. The combination of metrics is done at either the local or global level. For example, the index D (metric 56) combines two geometric metrics at the local level (OS_{ij} and US_{ij}) to produce a set of D_{ij} values, which is then aggregated for the global level. The F-measure (metric 55) combines two metrics at the global level (Precision and Recall). A few more complex strategies have also been proposed for combining metrics, namely clustering (CI, metric 58) and comparison of the cumulative distribution of the metrics combined (M^g and M^j, metrics 60 and 63).

Place table 2 near here. **See Table 2 after the references**.

Further methods are found in the literature. Most of them are essentially the same as those presented in Table 1 and Table 2. They are omitted here as are ambiguously described in the original publications; for example, the correspondence between objects and polygons is frequently unclear. Thus, they could not be translated to the notation defined in Section 3.1.1. Methods not described here are, however, potentially useful and include those found in Winter (2000); Oliveira et al. (2003); Radoux and Defourny (2007); Esch et al. (2008); Corcoran et al. (2010); Korting et al. (2011); Verbeeck et al. (2012); Whiteside et al. (2014); Michel et al. (2015) and Mikes et al. (2015).

3.1.3 Metrics use

Table 1 reveals that a variety of strategies has been adopted to compare objects and polygons. Specifically, often the assessment is focused on the reference data set, and thus the assessment proceeds by searching the objects that may correspond to each polygon (i.e. set Y is compared to set X). For example, Recall (metric 2) uses this strategy. Sometimes the assessment proceeds by searching the polygons that may correspond to each object (i.e. X is compared to Y). Precision (metric 1) adopts this latter strategy. The remaining strategies defined in Sections 3.1.1.3 and 3.1.1.4 are less frequently adopted, namely in three specific methods which calculate metrics 11-12, 40-42, and 65-66.

Once the strategy of comparison between objects and polygons is specified, several criteria may be used to determine the correspondence between objects and polygons. For example, when set Y compares to set X a simple criterion is to consider only one corresponding object for each of the polygons. This object may be the one that covers the largest extent of the polygon (e.g. Recall, metric 2). However, a set of different criteria can be used. For example, qLoc (metric 16) views an object as corresponding to a polygon if the centroid of the polygon lies inside the object or *vice versa*. As a result, several objects may be identified as corresponding to a single polygon. Only the corresponding objects and polygons are used for calculating the geometric metrics.

Most of the metrics presented in Table 1 and Table 2 are based on proportions of overlapping area. For example, Precision (metric 1) is based on the calculation of the proportion of the area that each object has in common with the corresponding polygon. On the other hand, some metrics are based on the distance between centroids. For example, qLoc (metric 16) is based on the distance between the centroid of each of the polygons to that of the corresponding objects. Metrics that focus on area are often referred to as area coincidence-based or area-based metrics. The metrics that focus on position are often referred to as boundary coincidence-based, location-based, or position-based metrics (Cheng et al., 2014; Clinton et al., 2010; Montaghi et al., 2013; Whiteside et al., 2014; Winter, 2000).

A substantial proportion of the metrics detect either under-segmentation or over-segmentation error. This may be unexpected as commonly a balanced result is desired, but it informs on what type of error dominates. This may be used, for example, to parameterize a segmentation algorithm. For this reason, normally metrics that detect and measure under- or over-segmentation error are calculated separately, but combined later (Table 2) to provide a complementary view on image segmentation accuracy. Moreover, area-based metrics and position-based metrics are sometimes combined to provide a comprehensive assessment of image segmentation accuracy from a geometric point of view (Möller et al., 2013). The combined metrics are typically the outcome of an image segmentation accuracy assessment based on a geometric approach. The possible values of these metrics are in the range between 0 and 1, and they may be used to rank a set of image segmentation outputs based on their expected suitability for image classification. To assist in the comparison of all metrics presented here, the metrics of Table 1 and Table 2 are grouped in Table 3 by type of error measured (over- and/or under-segmentation) and geometric feature considered (area and/or position).

Table 3. Geometric metrics of tables 1 and 2 grouped by type of error measured (over-segmentation and/or under-segmentation) and type of metric (area-based and/or position-based). Combined metrics of table 2 are in bold.

Type of metric	Type of error					
	Over-segmentation		Under-segmentation		Over- and under-segmentation	
Area-based	Recall	(2)	Precision	(1)	M	(5)
	uM	(3)	oM	(4)	AFI	(10)
	$LRE(x_i,y_j)_p$	(12)	$LRE(y_j,x_i)_p$	(11)	d_{sym}	(13)
	RAsub	(17)	E	(14)	SimSize	(15)
	countOver	(26)	RAsuper	(18)	G_s	(21)
	BsO	(30)	PI	(22)	F_{ij}	(23)
	OS	(34)	countUnder	(27)	m_2	(24)
	ED	(35)	A_j	(29)	qr	(25)
	FG	(36)	LP	(31)	SH	(37)
	NSR	(39)	EP	(32)	SOA	(50)
	O^R	(40)	US	(33)	MOA	(51)
	OE	(45)	PSE	(38)	OI2	(54)
	OS2	(48)	O^F	(41)	**F**	(55)
	OSE	(52)	CE	(44)	**D**	(56)
			US2	(47)	**BCE**	(57)
			TSI	(49)	**ED2**	(59)
			USE	(53)	**ADI**	(61)
					ED3	(62)
					SEI	(64)
					$BCA(x_i,y_j)_p$	(65)

					BCA	(66)
Position-based	User's BPA	(6)	Prod.'s BPA	(7)	qLoc	(16)
	C'	(8)	O'	(8)	RPsub	(19)
	P^R	(42)	P^F	(43)	RPsuper	(20)
					modD(b)	(28)
					PDI	(46)
Area- and position-based					**CI**	(58)
					$\mathbf{M^g}$	(60)
					$\mathbf{M^j}$	(63)

3.2 Non-geometric methods

A small number of non-geometric methods have been proposed (Table 4). Typically, this category of methods does not require an overlay operation between a polygonal reference data set and the image segmentation output under evaluation as they need not to be spatially coincident. Polygons may not even be used. The requirement common to all non-geometric methods is that the land cover class(es) associated with the objects are known. Note that non-geometric methods are not able to explicitly inform on which type of error, under- or over-segmentation, predominates.

Table 4. Non-geometric methods for supervised assessment of image segmentation accuracy. All metrics detect under- and over-segmentation error.

Reference	Focus of the method	Polygons needed[a]
Wang et al. (2004)	Objects' content (spectral separability of classes using the Bhattacharyya distance).	No
Laliberte and Rango (2009)	Classifier (Decision trees classification accuracy and Gini index).	No
Anders et al. (2011)	Objects' content (difference among objects and polygons on the frequency distribution of characterizing topographic attributes).	Yes
Yang et al. (2017)	Classifier (classification uncertainty)	No

[a] The reference data set used is required in the form of polygons

Non-geometric methods essentially follow two approaches to assess the accuracy of image segmentation. The first approach focuses on the content of the objects. Anders et al. (2011) compared the content of objects and polygons using the frequency distribution of their topographic attributes such as slope angle while mapping geomorphological features. Smaller differences between frequency distributions calculated from objects and polygons of the same geomorphological feature type indicated greater segmentation accuracy. However, most of the non-geometric methods dispense with polygons and only require objects with known spectral and thematic content. These objects may be represented in the spectral space used in the segmentation analysis where the objects of different land cover classes are desirable to lie in different regions so that later a classifier can allocate them to the correct class. The separability of the objects in the spectral space as a function of the land cover classes they represent is regarded as indicative of segmentation accuracy, and this can be assessed based on, for example, the Bhattacharyya distance (Fukunaga, 1990). This is possibly the most used non-geometric method (Li et al., 2015; Radoux and Defourny, 2008; Wang et al., 2004; Xun and Wang, 2015).

The second approach used in non-geometric methods assesses image segmentation using a classifier. Specifically, a series of preliminary classifications are undertaken with a set of image segmentation outputs, and the classifier is used to rank the segmentations based on their suitability for image classification. For example, a sample of the objects of the image segmentation under evaluation can be used to train a decision tree, and the impurity of the terminal nodes can be regarded as indicative of classification success; large accuracy of image segmentation is expected to be related to low node impurity (Laliberte and Rango, 2009). Most often, however, traditional estimators of classification accuracy such as overall accuracy are used (Laliberte and Rango, 2009; Smith, 2010). Thus, the classifier suggests which of a set of segmentation outputs affords the largest classification accuracy. In this case, samples of the objects constructed can be used for training and testing a classifier by means of out-of-bag estimate or cross-validation (Laliberte and Rango, 2009; Smith, 2010). Classification uncertainty rather than accuracy can also be used. If a fuzzy classifier is employed, the way in which the probability of class membership is partitioned between the classes can be used to calculate classification uncertainty, for example based on entropy measures. Segmentation accuracy may be viewed as negatively related to the magnitude of classification uncertainty (Yang et al., 2017).

The second approach of non-geometric image segmentation accuracy assessment, especially when classification accuracy expressed by traditional estimators such as overall accuracy is considered, may appear similar to traditional classification accuracy assessment, but they are different things. The former uses the training sample to assess the accuracy of the preliminary classifications while the latter assesses the quality of the final mapping product and requires an independent testing sample. Sometimes traditional classification accuracy assessment is nevertheless used to assess indirectly image segmentation accuracy (e.g. Kim et al., 2009; Li et al., 2011). When used, the focus is typically on a comparison among the accuracy values of a set of final classifications (Foody, 2009, 2004), with each produced with different image segmentation outputs. The differences are caused not only by the image segmentations used, but the entire approach to image classification. This may be well suited for applications focused on the final mapping products, but implies possibly impractical labour and resources such as multiple testing samples.

4 Selecting a method

The selection of a method to assess the accuracy of image segmentation is a complex decision, and here it is suggested to tackle that decision from two central perspectives: the application in-hand, and the pros and cons of the methods. These issues should be considered holistically although discussed separately hereafter.

4.1 Application in-hand

The purpose of the application in-hand should be considered, and there are two main situations. First, the applications are focused on just a fraction of the classes in which a landscape may be categorized. These applications use image segmentation primarily for object recognition and extraction, such as buildings and trees in urban environments (e.g. Belgiu and Drăguţ, 2014; Sebari and He, 2013). The desired characteristics of the objects are likely to be geometric, such as position and shape. Several methods may be appropriate, such as shape error (metric 37); the segmentation output indicated as optimal will in principle be formed by objects that most resemble the desired shapes represented in the reference data set. Alternatively, the relative overlapping areas between objects and polygons may be maximised. This strategy may benefit from area-based metrics designed for object recognition, such as SEI (metric 64).

The second main situation corresponds to wall-to-wall land cover classification and mapping (e.g Bisquert et al. 2015; Strasser and Lang 2015). In this case, the geometric properties of the objects may be considered important as in the first situation described above, and hence geometric methods may be used. However, the thematic information associated with the objects is commonly regarded as more important than the geometrical representation. In this context, an output that enables the maximisation of the area under analysis correctly represented in the final map is preferred. Geometric methods can still be used, and area-based methods may be appropriate, which will in principle suggest as optimal the segmentation output formed by objects that represent the largest amount of area of the corresponding polygons. This gives the classification stage the opportunity of maximising the area correctly classified and thus the overall accuracy of the map. Non-geometric methods can also be used (Table 4). There is less experience in the use of this category of methods, but it is potentially useful when the geometry of the objects does not have to meet predefined requirements.

An intermediate situation is also possible in that both the geometric and thematic properties of the objects are regarded as important. In this case, methods that combine different approaches for the accuracy assessment may be used, for example focused on the relative position and area of overlap between objects and polygons (Möller et al., 2013, 2007). However, there is no need to select just one method, and assembling multiple methods is a valid option (Clinton et al., 2010). Different methods, including geometric and non-geometric methods, can be used together to address all the specific properties of the objects considered as relevant as long as the set of methods used fits the purpose of the application in-hand.

Another relevant aspect of the application in-hand is the relative importance of under- and over-segmentation error. Image segmentation is typically conducted to trade-off and minimize under- and over-segmentation error, but over-segmentation may be needed to address conveniently the problem under analysis. Specifically, small objects, sometimes called primitive objects (Dronova, 2015), may be needed for modelling complex classes that are not directly related to spectral data, such as habitats (Strasser and Lang, 2015). The final land cover classes can be delineated later, for example, based on knowledge-driven semantic rules (Gu et al., 2017). If no primitive objects are needed, and the border of the final land over classes to be mapped are pursued in a segmentation analysis, it may be desirable nevertheless to recognize that under- and over-segmentation error are not always equally serious, especially if the application is interested more on the thematic rather than the geometric properties of the objects. Multiple authors have expressed their preference for over- rather than under-segmentation error as the latter is associated with relatively small classification accuracy (Gao et al., 2011; Hirata and Takahashi, 2011; Lobo, 1997; Wang et al., 2004). Under-segmentation error produces objects that correspond to more than one class on the ground and thus may represent an important origin of misclassification or land cover map error. Therefore, using methods able to inform on the level of over- and under-segmentation error may be convenient, such as that proposed by Möller et al. (2013).

The third and last aspect highlighted here relates to the potential importance of thematic errors associated with under-segmentation error. That is, the impact of under-segmentation error may depend on the classes associated with under-segmented objects. This is because the needs of the individual users may vary greatly in their sensitivity to misclassifications as a function of the classes involved (Bontemps et al., 2012; Comber et al., 2012). Traditionally, supervised methods consider all under-segmentation errors as equally serious, but under-segmentation errors can in fact be weighted as a function of the classes involved. This is the situation with the geometric method proposed by Costa et al. (2015) (metric 63) and non-geometric methods that use a classifier to perform a preliminary series of classifications, whose results can be expressed through weighted estimators of classification accuracy, such as the Stehman's (1999) map value V.

4.2 Methods’ pros and cons

A consideration of the potential implications associated with the approach of the assessment is advisable. Non-geometric methods do not require geo-registered reference data, which may be very practical, but are unable to explicitly inform on which type of segmentation error predominates. That information may be useful for guiding the definition of segmentation settings. If this limitation is undesirable, a geometric method suited to detecting segmentation error explicitly should be preferred. However, the need of defining criteria of correspondence between objects and polygons should be considered carefully as it impacts on the accuracy assessment. The geometric methods proposed by Yang et al. (2015) (SEI, metric 64), Su and Zhang (2017) (OSE, metric 52), and Möller et al. (2013) (M^g, metric 60) pay particular attention to this issue.

Quantitative comparisons of different methods should be undertaken. Several comparative studies dedicated to geometric methods have been published (Clinton et al., 2010; Räsänen et al., 2013; Whiteside et al., 2014; Yang et al., 2015), and some of them (e.g. Clinton et al., 2010; Verbeeck et al., 2012) observed that different methods can indicate very different segmentation outputs as optimal. Thus, special attention should be given to potential bias of the methods. For example, Radoux and Defourny (2008) found that spectral separability measures used in non-geometric methods may be insensitive to under-segmentation error, and thus indicate a segmentation as optimal while notably under-segmented; Witharana and Civco (2014) found that the sensitivity of Euclidean distance 2 (ED2, metric 59) to the accuracy of the objects depends of the scale of the analysis.

Finally, it should be noted that estimated bias in image segmentation accuracy assessment is not caused merely by unsuitable choice of methods or their potential flaws, but the protocol used for their implementation. Typically, some reference data are available for a sample of the entire area to be mapped, and thus limited data are used to infer an accuracy estimate to represent the entire area. Therefore, the nature of sampling is an issue that will impact on the results of an image segmentation accuracy assessment. The reference data must be acquired using a probability sampling design, which must incorporate a randomization component that has a non-zero probability of selection for each object into the sample. Consideration of general sampling and statistical principles for defining samples is recommended (Olofsson et al., 2014; Stehman and Czaplewski, 1998).

5 Discussion

5.1 Current status

Image segmentation accuracy assessment appears to be in a relatively early stage of maturation in land cover mapping applications. Often no information on the assessment produced is given, and qualitative assessment based on visual interpretation is widely used. This situation may be a result of several factors. For example, the lack of a solid background in image segmentation accuracy assessment and reliable recommendations for method selection may be a motivation for neglecting a quantitative accuracy assessment. Another factor may be related to the difficulty of implementing most of the methods proposed in the literature. Many analysts of remote sensing data depend on standard software and have no resources or expertise to implement new methods. This may also be a reason why comparison among methods has been addressed in a relatively small number of studies. There are some initiatives to implement supervised methods and make them available to the public (Mikes et al. 2015), but further work should be done in this respect. Clinton et al. (2010), Montaghi et al. (2013), Eisank et al. (2014), and Novelli et al. (2017) provide additional information on how to access software that includes supervised methods for image segmentation accuracy assessment.

Supervised methods were reviewed here and grouped into two categories: geometric and non-geometric methods. The former includes numerous area-based methods (Table 3), and many of them are similar. This is the case of area($x_i \cap y_j$)/area(y_j), which appears in metrics 1, 18, 33, and 47. Winter (2000) demonstrated that only seven metrics are possible to derive from an area-based approach if they are free of dimension, normalized, and symmetric (i.e. there is a single and mutual correspondence between objects and polygons). However, several correspondence criteria and strategies of comparison between objects and polygons can be specified, and thus the number of area-based metrics can proliferate. This is essentially the case of metrics 1, 18, 33 and 47, which are calculated with different criteria of correspondence between objects and polygons (X'_j, $\tilde{Y}_i$, Y'_i, and $Yc_i \cup Yd_i$, respectively). The ways the local metric values are used to produce a global accuracy value also vary. These apparently slight differences may, however, impact substantially on the assessment as different calculations are involved.

Selecting an appropriate method for image segmentation accuracy assessment is not obvious. The pros and cons of the potential methods, such as ease of use and bias, should be taken into account. However, it is noted that there is often neither a right nor wrong method. The suitability of a method will ultimately depend on how it fits with the application in-hand.

5.2 Research needs

Quantitative studies similar to Clinton et al. (2010) and Witharana and Civco (2014) should be done to exhaustively test and compare the supervised methods used in the remote sensing community. Non-geometric methods should be inspected as they have been neglected in quantitative studies. Moreover, the studies should be conducted under different contexts that may represent different types of applications, such as object recognition, and wall-to-wall mapping. Critically, research to address the relationship between segmentation and classification accuracies is required, as often relations were not simple (Belgiu and Drăguţ, 2014; Costa et al., 2017; Räsänen et al., 2013; Verbeeck et al., 2012).

Finally, the concept of over- and under-segmentation error should be revisited. Commonly, as in this paper, segmentation error is defined relative to the reference data used, and thus the concept lacks theoretical robustness. For example using reference data representing final land cover classes to be mapped or primitive objects impacts on the results. Primitive objects have a more spectral rather than thematic significance, and this may influence the assessment, including the selection of the assessment approach, supervised or unsupervised. However, theory and concepts related to object-based image analysis are generally incipient (Blaschke et al., 2014; Ma et al., 2017), and comparing supervised and unsupervised methods which often focus on thematic and primitive objects, respectively, has not received much attention.

6 Conclusions

Accuracy assessment is an important component of an image segmentation analysis, but is not mature. It has been much undertaken through visual inspection possibly for practical reasons while many quantitative approaches and methods have been proposed. Most often these methods are supervised and focus on the geometry of the objects constructed and polygons taken as reference data. However, other approaches may be used. The spectrum of methods available is large, and it is difficult to select consciously suitable methods for particular applications. There are at least three important questions that should be asked during the selection of supervised methods for image segmentation accuracy assessment: (i) the goal of the application; (ii) the relative importance of under- and over-segmentation error (including a possible varying sensitivity to thematic issues associated to under-segmentation); and (iii) the pros and cons of the methods. Answering these questions will help select suitable methods, but further research is needed to improve the standards of image segmentation accuracy assessment, otherwise there is the risk of using methods unsuitable or sub-optimal for the application in-hand.

Acknowledgements

Hugo Costa was supported by the PhD Studentship SFRH/BD/77031/2011 from the “Fundação para a Ciência e Tecnologia” (FCT), funded by the “Programa Operacional Potencial Humano” (POPH) and the European Social Fund. The paper benefited from valuable comments received from the editors and anonymous reviewers.

References

Abeyta, A., Franklin, J., 1998. The accuracy of vegetation stand boundaries derived from image segmentation in a desert environment. Photogramm. Eng. Remote Sensing 64, 59–66.

Anders, N.S., Seijmonsbergen, A.C., Bouten, W., 2011. Segmentation optimization and stratified object-based analysis for semi-automated geomorphological mapping. Remote Sens. Environ. 115, 2976–2985. doi:10.1016/j.rse.2011.05.007

Basaeed, E., Bhaskar, H., Hill, P., Al-Mualla, M., Bull, D., 2016. A supervised hierarchical segmentation of remote-sensing images using a committee of multi-scale convolutional neural networks. Int. J. Remote Sens. 37, 1671–1691. doi:10.1080/01431161.2016.1159745

Beauchemin, M., Thomson, K.P.B., Edwards, G., 1998. On the Hausdorff distance used for the evaluation of segmentation results. Can. J. Remote Sens. 24, 3–8. doi:10.1080/07038992.1998.10874685

Belgiu, M., Drăguţ, L., 2014. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. ISPRS J. Photogramm. Remote Sens. 96, 67–75. doi:10.1016/j.isprsjprs.2014.07.002

Bisquert, M., Bégué, A., Deshayes, M., 2015. Object-based delineation of homogeneous landscape units at regional scale based on MODIS time series. Int. J. Appl. Earth Obs. Geoinf. 37, 72–82. doi:10.1016/j.jag.2014.10.004

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16. doi:10.1016/j.isprsjprs.2009.06.004

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E.A., Queiroz Feitosa, R., van der 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, 180–191. doi:10.1016/j.isprsjprs.2013.09.014

Bontemps, S., Herold, M., Kooistra, L., van Groenestijn, A., Hartley, A., Arino, O., Moreau, I., Defourny, P., 2012. Revisiting land cover observation to address the needs of the climate modeling community. Biogeosciences 9, 2145–2157. doi:10.5194/bg-9-2145-2012

Bradley, B.A., 2014. Remote detection of invasive plants: A review of spectral, textural and phenological approaches. Biol. Invasions 16, 1411–1425. doi:10.1007/s10530-013-0578-9

Cardoso, J.S., Corte-Real, L., 2005. Toward a generic evaluation of image segmentation. IEEE Trans. Image Process. 14, 1773–1782. doi:10.1109/TIP.2005.854491

Carleer, A.P., Debeir, O., Wolff, E., 2005. Assessment of very high spatial resolution satellite image segmentations. Photogramm. Eng. Remote Sensing 71, 1285–1294. doi:10.14358/PERS.71.11.1285

Cheng, J., Bo, Y., Zhu, Y., Ji, X., 2014. A novel method for assessing the segmentation quality of high-spatial resolution remote-sensing images. Int. J. Remote Sens. 35, 3816–3839. doi:10.1080/01431161.2014.919678

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

Coillie, F.M.B. Van, Gardin, S., Anseel, F., 2014. Variability of operator performance in remote-sensing image interpretation: the importance of human and external factors. Int. J. Remote Sens. 35, 754–778. doi:10.1080/01431161.2013.873152

Comber, A., Fisher, P., Brunsdon, C., Khmag, A., 2012. Spatial analysis of remote sensing image classification accuracy. Remote Sens. Environ. 127, 237–246. doi:10.1016/j.rse.2012.09.005

Corcoran, P., Winstanley, A., Mooney, P., 2010. Segmentation performance evaluation for object-based remotely sensed image analysis. Int. J. Remote Sens. 31, 617–645. doi:10.1080/01431160902894475

Costa, G.A.O.P., Feitosa, R.Q., Cazes, T.B., Feijó, B., 2008. Genetic adaptation of segmentation parameters, in: Blaschke, T., Lang, S., Hay, G.J. (Eds.), Object-Based Image Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 679–695. doi:10.1007/978-3-540-77058-9_37

Costa, H., Foody, G.M., Boyd, D.S., 2015. Integrating user needs on misclassification error sensitivity into image segmentation quality assessment. Photogramm. Eng. Remote Sensing 81, 451–459. doi:10.14358/PERS.81.6.451

Costa, H., Foody, G.M., Boyd, D.S., 2017. Using mixed objects in the training of object-based image classifications. Remote Sens. Environ. 190, 188–197. doi:10.1016/j.rse.2016.12.017

Crevier, D., 2008. Image segmentation algorithm development using ground truth image data sets. Comput. Vis. Image Underst. 112, 143–159. doi:10.1016/j.cviu.2008.02.002

Drăguţ, L., Csillik, O., Eisank, C., Tiede, D., 2014. Automated parameterisation for multi-scale image segmentation on multiple layers. ISPRS J. Photogramm. Remote Sens. 88, 119–127. doi:10.1016/j.isprsjprs.2013.11.018

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. Inf. Sci. 24, 859–871. doi:10.1080/13658810903174803

Dronova, I., 2015. Object-based image analysis in wetland research: A review. Remote Sens. 7, 6380–6413. doi:10.3390/rs70506380

Eisank, C., Smith, M., Hillier, J., 2014. Assessment of multiresolution segmentation for delimiting drumlins in digital elevation models. Geomorphology 214, 452–464. doi:10.1016/j.geomorph.2014.02.028

Esch, T., Thiel, M., Bock, M., Roth, A., Dech, S., 2008. Improvement of image segmentation accuracy based on multiscale optimization procedure. IEEE Geosci. Remote Sens. Lett. 5, 463–467. doi:10.1109/LGRS.2008.919622

Feitosa, R.Q., Ferreira, R.S., Almeida, C.M., Camargo, F.F., Costa, G.A.O.P., 2010. Similarity metrics for genetic adaptation of segmentation parameters, in: 3rd International Conference on Geographic Object-Based Image Analysis (GEOBIA 2010). The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Ghent.

Feizizadeh, B., Blaschke, T., Tiede, D., Moghaddam, M.H.R., 2017. Evaluating fuzzy operators of an object-based image analysis for detecting landslides and their changes. Geomorphology 293, Part A, 240–254. doi:https://doi.org/10.1016/j.geomorph.2017.06.002

Foody, G.M., 2009. Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. Remote Sens. Environ. 113, 1658–1663. doi:10.1016/j.rse.2009.03.014

Foody, G.M., 2004. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. Photogramm. Eng. Remote Sensing 70, 627–633. doi:10.14358/PERS.70.5.627

Fukunaga, M., 1990. Introduction to statistical pattern recognition, 2nd ed. Academic Press, San Diego.

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

doi:10.1080/01431161003777189

Gu, H., Li, H., Yan, L., Liu, Z., Blaschke, T., Soergel, U., 2017. An Object-Based Semantic Classification Method for High Resolution Remote Sensing Imagery Using Ontology. Remote Sens. . doi:10.3390/rs9040329

Hirata, Y., Takahashi, T., 2011. Image segmentation and classification of Landsat Thematic Mapper data using a sampling approach for forest cover assessment. Can. J. For. Res. 41, 35–43. doi:10.1139/X10-130

Hultquist, C., Chen, G., Zhao, K., 2014. A comparison of Gaussian process regression, random forests and support vector regression for burn severity assessment in diseased forests. Remote Sens. Lett. 5, 723–732. doi:10.1080/2150704X.2014.963733

Janssen, L.L.F., Molenaar, M., 1995. Terrain objects, their dynamics and their monitoring by the integration of GIS and remote sensing. IEEE Trans. Geosci. Remote Sens. 33, 749–758. doi:10.1109/36.387590

Kim, M., Madden, M., Warner, T.A., 2009. Forest type mapping using object-specific texture measures from multispectral Ikonos Imagery: Segmentation quality and image classification issues. Photogramm. Eng. Remote Sensing 75, 819–829.

Korting, T.S., Dutra, L.V., Fonseca, L.M.G., 2011. A resegmentation approach for detecting rectangular objects in high-resolution imagery. IEEE Geosci. Remote Sens. Lett. 8, 621–625. doi:10.1109/LGRS.2010.2098389

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, 1–10. doi:10.1109/TGRS.2008.2009355

Lang, S., Albrecht, F., Kienberger, S., Tiede, D., 2010. Object validity for operational tasks in a policy context. J. Spat. Sci. 55, 9–22. doi:10.1080/14498596.2010.487639

Lang, S., Kienberger, S., Tiede, D., Hagenlocher, M., Pernkopf, L., 2014. Geons – domain-specific regionalization of space. Cartogr. Geogr. Inf. Sci. 41, 214–226. doi:10.1080/15230406.2014.902755

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.

Li, D., Ke, Y., Gong, H., Li, X., 2015. Object-based urban tree species classification using bi-temporal WorldView-2 and WorldView-3 images. Remote Sens. 7, 16917–16937. doi:10.3390/rs71215861

Li, P., Guo, J., Song, B., Xiao, X., 2011. A multilevel hierarchical image segmentation method for urban impervious surface mapping using very high resolution imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4, 103–116. doi:10.1109/JSTARS.2010.2074186

Liu, D., Xia, F., 2010. Assessing object-based classification: Advantages and limitations. Remote Sens. Lett. 1, 187–194. doi:10.1080/01431161003743173

Liu, J., Li, P., Wang, X., 2015. A new segmentation method for very high resolution imagery using spectral and morphological information. ISPRS J. Photogramm. Remote Sens. 101, 145–162. doi:10.1016/j.isprsjprs.2014.11.009

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. Photogramm. Remote Sens. 68, 144–156. doi:10.1016/j.isprsjprs.2012.01.007

Lobo, A., 1997. Image segmentation and discriminant analysis for the identification of land cover units in ecology. IEEE Trans. Geosci. Remote Sens. 35, 1136–1145. doi:10.1109/36.628781

Lucieer, A., Stein, A., 2002. Existential uncertainty of spatial objects segmented from satellite sensor imagery. Geosci. Remote Sensing, IEEE Trans. 40, 2518–2521. doi:10.1109/TGRS.2002.805072

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017. A review of supervised object-based land-cover image classification. ISPRS J. Photogramm. Remote Sens. 130, 277–293. doi:10.1016/j.isprsjprs.2017.06.001

Marpu, P.R., Neubert, M., Herold, H., Niemeyer, I., 2010. Enhanced evaluation of image segmentation results. J. Spat. Sci. 55, 55–68. doi:10.1080/14498596.2010.487850

Martin, D.R., 2003. An empirical approach to grouping and segmentation. ECS Department, University of California.

Matikainen, L., Karila, K., Hyyppä, J., Litkey, P., Puttonen, E., & Ahokas, E. (2017). Object-based analysis of multispectral airborne laser scanner data for land cover classification and map updating. *ISPRS Journal of Photogrammetry and Remote Sensing*, *128*, 298–313. http://doi.org/https://doi.org/10.1016/j.isprsjprs.2017.04.005

Michel, J., Youssefi, D., Grizonnet, M., 2015. Stable mean-shift algorithm and its application to the segmentation of arbitrarily large remote sensing images. IEEE Trans. Geosci. Remote Sens. 53, 952–964. doi:10.1109/TGRS.2014.2330857

Mikes, S., Haindl, M., Scarpa, G., Gaetano, R., 2015. Benchmarking of remote sensing segmentation methods. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 2240–2248. doi:10.1109/JSTARS.2015.2416656

Möller, M., Birger, J., Gidudu, A., Gläßer, C., 2013. A framework for the geometric accuracy assessment of classified objects. Int. J. Remote Sens. 34, 8685–8698. doi:10.1080/01431161.2013.845319

Möller, M., Lymburner, L., Volk, M., 2007. The comparison index: A tool for assessing the accuracy of image segmentation. Int. J. Appl. Earth Obs. Geoinf. 9, 311–321. doi:10.1016/j.jag.2006.10.002

Montaghi, A., Larsen, R., Greve, M.H., 2013. Accuracy assessment measures for image segmentation goodness of the Land Parcel Identification System (LPIS) in Denmark. Remote Sens. Lett. 4, 946–955. doi:10.1080/2150704X.2013.817709

Neubert, M., Herold, H., Meinel, G., 2008. Assessing image segmentation quality – Concepts, methods and application, in: Blaschke, T., Lang, S., Hay, G. (Eds.), Object-Based Image Analysis. Springer Berlin Heidelberg, pp. 769–784.

Novelli, A., Aguilar, M., Aguilar, F., Nemmaoui, A., Tarantino, E., 2017. AssesSeg—A Command Line Tool to Quantify Image Segmentation Quality: A Test Carried Out in Southern Spain from Satellite Imagery. Remote Sens. 9, 40. doi:10.3390/rs9010040

Oliveira, J., Formaggio, A., Epiphanio, J., Luiz, A., 2003. Index for the Evaluation of Segmentation (IAVAS): An application to agriculture. Mapp. Sci. Remote Sens. 40, 155–169. doi:10.2747/0749-3878.40.3.155

Olofsson, P., Foody, G.M., Herold, M., Stehman, S. V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. doi:10.1016/j.rse.2014.02.015

Persello, C., Bruzzone, L., 2010. A novel protocol for accuracy assessment in classification of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48, 1232–1244. doi:10.1109/TGRS.2009.2029570

Qi, Z., Yeh, A.G.-O., Li, X., Zhang, X., 2015. A three-component method for timely detection of land cover changes using polarimetric SAR images. ISPRS J. Photogramm. Remote Sens. 107, 3–21. doi:10.1016/j.isprsjprs.2015.02.004

Radoux, J., Defourny, P., 2008. Quality assessment of segmentation results devoted to object-based classification, in: Blaschke, T., Lang, S., Hay, G.J. (Eds.), Object-Based Image Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 257–271. doi:10.1007/978-3-540-77058-9_14

Radoux, J., Defourny, P., 2007. A quantitative assessment of boundaries in automated forest stand delineation using very high resolution imagery. Remote Sens. Environ. 110, 468–475. doi:10.1016/j.rse.2007.02.031

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, 8603–8627. doi:10.1080/01431161.2013.845318

Robson, B.A., Nuth, C., Dahl, S.O., Hölbling, D., Strozzi, T., Nielsen, P.R., 2015. Automated classification of debris-covered glaciers combining optical, SAR and topographic data in an object-based environment. Remote Sens. Environ. 170, 372–387. doi:10.1016/j.rse.2015.10.001

Sebari, I., He, D.-C., 2013. Automatic fuzzy object-based analysis of VHSR images for urban objects extraction. ISPRS J. Photogramm. Remote Sens. 79, 171–184. doi:10.1016/j.isprsjprs.2013.02.006

Smith, A., 2010. Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. J. Spat. Sci. 55, 69–79. doi:10.1080/14498596.2010.487851

Stehman, S. V., 1999. Comparing thematic maps based on map value. Int. J. Remote Sens. 20, 2347–2366. doi:10.1080/014311699212065

Stehman, S. V., Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment. Remote Sens. Environ. 64, 331–344. doi:10.1016/S0034-4257(98)00010-8

Strasser, T., Lang, S., 2015. Object-based class modelling for multi-scale riparian forest habitat mapping. Int. J. Appl. Earth Obs. Geoinf. 37, 29–37. doi:10.1016/j.jag.2014.10.002

Su, T., Zhang, S., 2017. Local and global evaluation for remote sensing image segmentation. ISPRS J. Photogramm. Remote Sens. 130, 256–276. doi:10.1016/j.isprsjprs.2017.06.003

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, 4625–4644. doi:10.1080/01431160701241746

Van Coillie, F.M.B., Verbeke, L.P.C., De Wulf, R.R., 2008. Semi-automated forest stand delineation using wavelet based segmentation of very high resolution optical imagery, in: Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications. pp. 237–256. doi:10.1007/978-3-540-77058-9_13

Van Rijsbergen, C.J., 1979. Information retrieval. Butterworth-Heinemann, London.

Verbeeck, K., Hermy, M., Van Orshoven, J., 2012. External geo-information in the segmentation of VHR imagery improves the detection of imperviousness in urban neighborhoods. Int. J. Appl. Earth Obs. Geoinf. 18, 428–435. doi:10.1016/j.jag.2012.03.015

Wang, L., Sousa, W.P., Gong, P., 2004. Integration of object-based and pixel-based classification for mapping mangroves with IKONOS imagery. Int. J. Remote Sens. 25, 5655–5668. doi:10.1080/014311602331291215

Weidner, U., 2008. Contribution to the assessment of segmentation quality for remote sensing applications. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 37, 479–484.

Whiteside, T.G., Maier, S.W., Boggs, G.S., 2014. Area-based and location-based validation of classified image objects. Int. J. Appl. Earth Obs. Geoinf. 28, 117–130. doi:10.1016/j.jag.2013.11.009

Winter, S., 2000. Location similarity of regions. ISPRS J. Photogramm. Remote Sens. 55, 189–200. doi:10.1016/S0924-2716(00)00019-8

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. Photogramm. Remote Sens. 87, 108–121.

doi:10.1016/j.isprsjprs.2013.11.006

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. Photogramm. Remote Sens. 87, 1–18. doi:10.1016/j.isprsjprs.2013.10.005

Xun, L., Wang, L., 2015. An object-based SVM method incorporating optimal segmentation scale estimation using Bhattacharyya Distance for mapping salt cedar (Tamarisk spp.) with QuickBird imagery. GIScience Remote Sens. 52, 257–273. doi:10.1080/15481603.2015.1026049

Yang, J., He, Y., Caspersen, J.P., Jones, T., 2017. Delineating Individual Tree Crowns in an Uneven-Aged, Mixed Broadleaf Forest Using Multispectral Watershed Segmentation and Multiscale Fitting. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 1390–1401. doi:10.1109/JSTARS.2016.2638822

Yang, J., He, Y., Caspersen, J., Jones, T., 2015. A discrepancy measure for segmentation evaluation from the perspective of object recognition. ISPRS J. Photogramm. Remote Sens. 101, 186–192. doi:10.1016/j.isprsjprs.2014.12.015

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, 13–24. doi:10.1016/j.isprsjprs.2014.04.008

Yi, L., Zhang, G., Wu, Z., 2012. A scale-synthesis method for high spatial resolution remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens. 50, 4062–4070. doi:10.1109/TGRS.2012.2187789

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, 2953–2974. doi:10.1080/01431160500057764

Zhang, H., Fritts, J.E., Goldman, S.A., 2008. Image segmentation evaluation: A survey of unsupervised methods. Comput. Vis. Image Underst. 110, 260–280. doi:10.1016/j.cviu.2007.08.003

Zhang, X., Feng, X., Xiao, P., He, G., Zhu, L., 2015a. Segmentation quality evaluation using region-based precision and recall measures for remote sensing images. ISPRS J. Photogramm. Remote Sens. 102, 73–84. doi:10.1016/j.isprsjprs.2015.01.009

Zhang, X., Xiao, P., Feng, X., Feng, L., Ye, N., 2015b. Toward evaluating multiscale segmentations of high spatial resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 53, 3694–3706. doi:10.1109/TGRS.2014.2381632

Zhang, X., Xiao, P., Feng, X., Wang, J., Wang, Z., 2014. Hybrid region merging method for segmentation of high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 98, 19–28. doi:10.1016/j.isprsjprs.2014.09.011

Zhang, Y.J., 1996. A survey on evaluation methods for image segmentation. Pattern Recognit. 29, 1335–1346. doi:10.1016/0031-3203(95)00169-7

LIST OF FIGURE CAPTIONS

Figure 1. Geometric comparison between an object and polygon based on the overlapping area (shaded area) and/or distance between centroids (dashed arrow).

Figure 2. Sets X, Y, and S: (a) reference set X, (b) segmentation Y, and (c) intersection $S=X \cap Y$. In (c) yellow denotes one-to-one, blue denotes one-to-many, and pink denotes many-to-many (Section 3.1.1.3).

Figure 3. Comparison between X and Y of Figure 2: (a) four potential objects (dashed lines) corresponding to polygon x_1 (grey background) when Y compares to X; (b) one potential polygon (dashed line) corresponding to object y_1 (grey background) when X compares to Y.

Table 1. Geometric metrics for supervised assessment of image segmentation accuracy. All metrics are numbered and ordered chronologically. The type of metric and segmentation error are identified in columns Typ. and Err. while the minimum, maximum, and optimal values of the metrics are identified in columns Min., Max., and Opt. The subscripts of the metrics' name indicate local accuracy assessment (see notes on the corresponding global metric), and global metrics have no subscripts.

Metric		Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(1)	$\mathrm{Precision}_{ij} = \frac{\mathrm{area}(x_i \cap y_j)}{\mathrm{area}(y_j)}, x_i \in X'_j$	Van Rijsbergen (1979) and Zhang et al. (2015a).	AB	U	0	1	1	Global metric Precision is the weighted mean of all $\mathrm{Precision}_{ij}$ using $\mathrm{area}(y_j)$ as weights.
(2)	$\mathrm{Recall}_{ij} = \frac{\mathrm{area}(x_i \cap y_j)}{\mathrm{area}(x_i)}, y_j \in Y'_i$	Van Rijsbergen (1979) and Zhang et al. (2015a).	AB	O	0	1	1	Global metric Recall is the weighted mean of all Recall_{ij} using $\mathrm{area}(x_i)$ as weights.
(3)	$\mathrm{underMerging}_{ij} = \frac{\mathrm{area}(x_i) - \mathrm{area}(x_i \cap y_j)}{\mathrm{area}(x_i)}, y_j$	Levine and Nazif (1982) and Clinton et al. (2010).	AB	O	0	0.5	0	Global metric underMerging can be the mean of all $\mathrm{underMerging}_{ij}$.
(4)	$\mathrm{overMerging}_{ij} = \frac{\mathrm{area}(y_j) - \mathrm{area}(x_i \cap y_j)}{\mathrm{area}(x_i)}, y_j \in$	Levine and Nazif (1982) and Clinton et al. (2010).	AB	U	0	0.5	0	Global metric overMerging can be the mean of all $\mathrm{overMerging}_{ij}$.

Metric		Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(5)	$M_{ij} = \sqrt{\frac{\text{area}(x_i \cap y_j)^2}{\text{area}(x_i) \times \text{area}(y_j)}}, y_j \in Y_i'$	Janssen and Molenaar (1995) and Feitosa et al. (2010).	AB	UO	0	1	1	Match (M). Global metric M is the mean of all M_{ij} values.
(6)	User's BPA= proportion of boundary length defined in segmentation with corresponding real boundaries	Abeyta and Franklin (1998)	PB	O	0	1	1	Boundary positional accuracy (BPA). Boundary length are estimated based on point-type data collected via line intersect sampling. Boundaries defined in segmentation that fell within ε (epsilon) tolerances (spatial error bounds) of surveyed boundaries are considered accurate.
(7)	Producers's BPA= proportion of real boundary length with corresponding boundaries defined in segmentation	Abeyta and Franklin (1998)	PB	U	0	1	1	Boundary positional accuracy (BPA). Boundary length are estimated based on point-type data collected via line intersect sampling. Boundaries defined in segmentation that fell within ε (epsilon) tolerances (spatial error bounds) of surveyed boundaries are considered accurate.

Metric		Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(8)	$C' = 1 - \frac{\mathrm{size}\left(\bigcup_i \mathrm{vertex}(y_j)\right) - \mathrm{size}\left(\bigcup_i \left(\mathrm{dist}\left(\mathrm{vertex}(y_j); \mathrm{vertex}(x_i)\right)\right)\right)}{\mathrm{size}\left(\bigcup_i \mathrm{vertex}(x_i)\right)}, x_i \in X$	Beauchemin et al. (1998).	PB	O	0		0	dist() represents the partial directed Hausdorff distance, which calculates the fraction of vertexes of the objects of Y that are each within a distance of some vertex of the polygons of X.
(9)	$O' = 1 - \frac{\mathrm{size}\left(\bigcup_i \left(\mathrm{dist}\left(\mathrm{vertex}(x_i); \mathrm{vertex}(y_j)\right)\right)\right)}{\mathrm{size}\left(\bigcup_i \mathrm{vertex}(x_i)\right)}, y_j \in Y$	Beauchemin et al. (1998).	PB	U	0	1	0	dist() represents the partial directed Hausdorff distance, which calculates the fraction of vertexes of the polygons of X that are each within a distance of some vertex of the objects of Y.
(10)	$AFI_{ij} = \frac{\mathrm{area}(x_i) - \mathrm{area}(y_j)}{\mathrm{area}(x_i)}, y_j \in Y_i'$	Lucieer and Stein (2002) and Clinton et al. (2010).	AB	UO			0	Area fit index (AFI). Global metric AFI is the mean of all AFI_{ij} values. AFI<0 and AFI>0 indicate under- and over-segmentation.
(11)	$LRE(y_j, x_i)_p = \frac{\mathrm{size}(\neg x_i \cap y_j)}{\mathrm{size}(y_j)}, x_i \in Xa_p \wedge y_j \in Ya_p$	Martin (2003) and Zhang et al. (2015a).	AB	U	0	1	0	Local refinement error (LRE). This metric was not proposed to be aggregated for the entire segmentation output (see metric 57 in Table 2).
(12)	$LRE(x_i, y_j)_p = \frac{\mathrm{size}(x_i \cap \neg y_j)}{\mathrm{size}(x_i)}, x_i \in Xa_p \wedge y_j \in Ya_p$	Martin (2003) and Zhang et al. (2015a).	AB	O	0	1	0	Local refinement error (LRE). This metric was not proposed to be aggregated for the entire segmentation output (see metric 57 in Table 2).

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(13) d_{sym} = minimal number of pixels that must be removed from both X and Y so that they are identical in the remaining pixels.	Cardoso and Corte-Real (2005).	AB	UO	0	1	0	d_{sym} is normalized to 0-1 by dividing by *l*-1. $d'_{sym} = 1 - d_{sym}$ in Zhang et al. (2015a).
(14) $E_{ij} = \frac{area(y_j) - area(x_i \cap y_j)}{area(y_j)} \times 100, x_i \in X'_j$	Carleer et al. (2005).	AB	U	0	50	0	Global metric E is the weighted mean of all E_{ij}, using $area(y_j)$ as weights. A refinement of E is also presented in Carleer et al. (2005).
(15) $SimSize_{ij} = \frac{\min(area(x_i), area(y_j))}{\max(area(x_i), area(y_j))}, y_j \in Y_i^*$	Zhan et al. (2005) and Clinton et al. (2010).	AB	UO	0	1	1	Global metric SimSize can be the mean of all $SimSize_{ij}$.
(16) $qLoc_{ij} = dist(centroid(x_i), centroid(y_j)), y_j \in Y_i^*$	Zhan et al. (2005) and Clinton et al. (2010).	PB	UO	0		0	dist() represents Euclidean distance. Global metric qLoc can be the mean of all $qLoc_{ij}$.
(17) $RAsub_{ij} = \frac{area(x_i \cap y_j)}{area(x_i)}, y_j \in \tilde{Y}_i$	Möller et al. (2007) and Clinton et al. (2010).	AB	O	0	1	1	Relative area (RA). This metric was not proposed to be aggregated for the whole segmentation output (see metric 58 in Table 2).

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(18) $RAsuper_{ij} = \frac{area(x_i \cap y_j)}{area(y_j)}, y_j \in \tilde{Y}_i$	Möller et al. (2007) and Clinton et al. (2010).	AB	U	0	1	1	Relative area(RA). This metric was not proposed to be aggregated for the whole segmentation output (see metric 58 in Table 2).
(19) $RPsub_{ij} = dist(centroid(x_i), centroid(y_j)), y_j \in \tilde{Y}_i$	Möller et al. (2007) and Clinton et al. (2010).	PB	UO	0		0	Relative position (RP). This metric was not proposed to be aggregated for the whole segmentation output (see metric 58 in Table 2).
(20) $RPsuper_{ij} = \frac{dist(centroid(x_i), centroid(y_j))}{\max(RPsub_{ij})}, y_j \in Y_i^*$	Möller et al. (2007) and Clinton et al. (2010).	PB	UO	0	1	0	Relative position (RP). dist() represents Euclidean distance. This metric was not proposed to be aggregated for the whole segmentation output (see metric 58 in Table 2).
(21) $G_s = \frac{\sum_i \sum_j area(x_i \cap y_j)}{area(X) \times e^{\sum_i \sum_j (area(x_i \cup y_j) - area(x_i \cap y_j)) / area(X)}}, y_j \in Yc_i$	Tian and Chen (2007).	AB	UO	0	1	1	
(22) $PI_{ij} = \frac{area(x_i \cap y_j)^2}{area(y_j) \times area(x_i)}, y_j \in \tilde{Y}_i$	Coillie et al. (2008).	AB	U	0	1	1	Purity Index (PI). Global metric PI is the mean of all summed PI_{ij} over all x_i.

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(23) $F_{ij}=\frac{area(y_j)+area(x_i)-2\times area(y_j\cap x_i)}{area(y_j)}, x_i \in X'_j$	Costa et al. (2008).	AB	UO	0		0	Fitness function (F). Global metric F is the mean of all summed F_{ij} over all y_j.
(24) $m_{2_{ij}}=\frac{area(y_j\cap x_i)}{area(y_j\cup x_i)}, x_i \in X''_j$	Crevier (2008) and Yi et al. (2012).	AB	UO	0	1	1	Global metric m_2 is the sum of all m_{2ij}.
(25) $qr_{ij}=1-\frac{area(x_i\cap y_j)}{area(x_i\cup y_j)}, y_j \in Y_i^*$	Weidner (2008) and Clinton et al. (2010).	AB	UO	0	1	0	Quality rate (qr). Global metric qr can be the mean of all qr_{ij}
(26) $countOver = size(X), \frac{area(y_j)}{area(x_i)}<1 \wedge AFI_{ij}>0 \wedge y_j \in Y_i^*$	Clinton et al. (2010).	AB	O	0	size(x)	0	$AFI_{ij}=\frac{area(x_i)-area(y_j)}{area(x_i)}, y_j \in Y'_i$ (see metric 10).
(27) $countUnder = size(X), \frac{area(y_j)}{area(x_i)}=1 \wedge AFI_{ij}<0 \wedge y_j \in Y_i^*$	Clinton et al. (2010).	AB	U	0	size(x)	0	$AFI_{ij}=\frac{area(x_i)-area(y_j)}{area(x_i)}, y_j \in Y'_i$ (see metric 10).

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(28) $modD(b)_i$ = mean Euclidean distance between each vertex of x_i and the closest vertex in every $y_j \in Y_i^*$	Clinton et al. (2010).	PB	UO	0		0	Global metric modD(b) can be the mean of all $modD(b)_i$.
(29) $A_j = \frac{\max(area(c_i \cap y_j))}{area(y_j)}, c_i \in \tilde{C}_j$	Liu and Xia (2010).	AB	U	0	1	1	Segmentation accuracy (A). Global metric A is the weighted mean of all A_j using $area(y_j)$ as weights.
(30) $BsO_i = \max\left(\frac{area(y_j) - area(\neg x_i \cap y_j)}{area(x_i)}\right) \times 100, y_j \in Yf_i$	Marpu et al. (2010).	AB	O	0	100	100	Biggest sub-object (BsO). Global BsO can be descriptive statistics of all BsO_i (e.g. quartiles).
(31) $LP_i = \frac{area(x_i) - \sum_j area(x_i \cap y_j)}{area(x_i)} \times 100, y_j \in Yf_i$	Marpu et al. (2010).	AB	U	0	100	0	Lost pixels (LP). Global LP can be descriptive statistics of all LP_i (e.g. quartiles).
(32) $EP_{ij} = \frac{area(y_j) - area(x_i \cap y_j)}{area(x_i)} \times 100, y_j \in Yf_i$	Marpu et al. (2010).	AB	U	0	100	0	Extra pixels (EP). Global EP can be descriptive statistics of all summed EP_{ij} over all x_i (e.g. quartiles).

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(33) $US_{ij} = 1 - \frac{\mathrm{area}(x_i \cap y_j)}{\mathrm{area}(y_j)}, y_j \in Y_i'$	Persello and Bruzzone (2010) and Clinton et al. (2010).	AB	U	0	1	0	undersegmentation error (US). Global metric US can be the mean of all US_{ij}. Clinton et al. (2010) consider subset Y_i^*.
(34) $OS_{ij} = 1 - \frac{\mathrm{area}(x_i \cap y_j)}{\mathrm{area}(x_i)}, y_j \in Y_i'$	Persello and Bruzzone (2010) and Clinton et al. (2010).	AB	O	0	1	0	oversegmentation error (OS). Global metric OS can be the mean of all OS_{ij}. Clinton et al. (2010) consider subset Y_i^*.
(35) $ED_{ij} = 1 - \frac{\mathrm{perim}(x_i) \cap \mathrm{perim}(y_j)}{\mathrm{perim}(x_i)}, y_j \in Y_i'$	Persello and Bruzzone (2010).	AB	O	0	1	0	Edge location (ED). Global metric ED can be the mean of all ED_{ij}.
(36) $FG_i = \frac{\mathrm{size}(\tilde{Y}_i) - 1}{\mathrm{area}(x_i) - 1}$	Persello and Bruzzone (2010).	AB	O	0	1	0	Fragmentation error (FG). Global metric FG can be the mean of all FG_i.
(37) $SH_{ij} = \left\| \mathrm{sf}(x_i) - \mathrm{sf}(y_j) \right\|, y_j \in Y_i'$	Persello and Bruzzone (2010).	AB	UO	0		0	Shape error (SH). $\|\cdot\|$ denotes the absolute value of '·' and sf(·) denotes a shape factor of '·' such as compactness and sphericity. Global metric SH can be the mean of all SH_{ij}.

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(38) $PSE_{ij}=\frac{area(\neg x_i \cap y_j)}{area(x_i)}, y_j \in Yc_i \cup Yd_i$	Liu et al. (2012).	AB	U	0		0	Potential segmentation error (PSE). Global metric PSE is the weighted mean all PSE_{ij}, using $area(x_j)$ as weights. A refinement of PSE is presented in Novelli et al. (2017).
(39) $NSR=\frac{\left\lvert size(X)-size\left(\bigcup_i (Yc_i \cup Yd_i)\right)\right\rvert}{size(X)}$	Liu et al. (2012).	AB	O	0		0	Number-of-segments ratio (NSR). $\lvert\cdot\rvert$ denotes the absolute value of '·'. A refinement of NSR is presented in Novelli et al. (2017).
(40) $O^R_{ijk}=\frac{area(s_{ijk})}{area(x_i)}, s_{ijk} \in Sax_i \vee Sbx_i$	Möller et al. (2013).	AB	O	0	1	1	This metric was not proposed to be aggregated for the whole segmentation output (see metric 60 in Table 2).
(41) $O^F_{ijk}=\frac{area(s_{ijk})}{area(y_j)}, s_{ijk} \in Say_j \vee Sby_j$	Möller et al. (2013).	AB	U	0	1	1	This metric was not proposed to be aggregated for the whole segmentation output (see metric 60 in Table 2).
(42) $P^R_{ijk}=1-\frac{dist(centroid(s_{ijk}), centroid(x_i))}{d^x_{max}}, s_{ijk} \in Sax_i \vee Sbx_i$	Möller et al. (2013).	PB	O	0	1	1	$d^x_{max}=max(dist(centroid(s_{ijk}))), s_{ijk} \in Sax_i \vee Sbx_i$. dist() represents Euclidean distance. This metric was not proposed to be aggregated for the whole segmentation output (see metric 60 in Table 2).

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(43) $P_{ijk}^{F} = 1 - \frac{dist(centroid(s_{ijk}), centroid(y_j))}{d_{max}^{y}}, s_{ijk} \in Say_j \vee Sby$	Möller et al. (2013).	PB	U	0	1	1	$d_{max}^{y} = max(dist(centroid(s_{ijk}))), s_{ijk} \in Say_j \vee Sby_j$. dist() represents Euclidean distance. This metric was not proposed to be aggregated for the whole segmentation output (see metric 60 in Table 2).
(44) $CE_{ij} = \frac{area(y_i) - area(x_i \cap y_j)}{area(x_i)} \times 100, y_j \in Yb_i \cap Yc_i$	Cheng et al. (2014).	AB	U	0	50	0	Commission error (CE). Global metric $CE_{overall}$ is the weighted mean of all CE_{ij}, using $area(x_j)$ as weights.
(45) $OE_{ij} = \frac{area(x_i \cap y_j)}{area(x_i)} \times 100, y_j \in \tilde{Y}_i \setminus Yb_i \cap Yc_i$	Cheng et al. (2014).	AB	O	0	50	0	Omission error (OE). Global metric $OE_{overall}$ is the weighted mean of all OE_{ij}, using $area(x_j)$ as weights.
(46) $PDI_{ij} = dist(centroid(x_i), centroid(y_j)), y_j \in Yb_i \cup Yc_i$	Cheng et al. (2014).	PB	UO	0		0	Position discrepancy index (PDI). Global metric $PDI_{overall}$ is the mean of all averaged PDI_{ij} over all x_i.
(47) $US2_{ij} = 1 - \frac{area(x_i \cap y_j)}{area(y_j)}, y_j \in Yc_i \cup Yd_i$	Yang et al. (2014).	AB	U	0	1	0	Global metric US is the sum of all summed US_{ij} over each x_i.
(48) $OS2_{ij} = 1 - \frac{area(x_i \cap y_j)}{area(x_i)}, y_j \in Yc_i \cup Yd_i$	Yang et al. (2014).	AB	O	0	1	0	Global metric OS is the sum of all summed OS_{ij} over each x_i.

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(49) $TSI_j = \sum_{c_i} \left(\frac{area(c_i)}{area(y_j)} \sum_{d_i} \left(\frac{area(d_i)}{area(y_j)} w_{c_i d_i} \right) \right), c_i \wedge d_i \in \tilde{C}_j$	Costa et al. (2015).	AB	U	0	1	1	Thematic similarity index (TSI). Global metric TSI is the weighted mean of all TSI_j. using $area(y_j)$ as weights.
(50) $SOA_{ij} = \frac{area(x_i \cap y_j) \times 2}{area(x_i) + area(y_j)}, y_j \in \tilde{Y}_i$	Zhang et al. (2015b).	AB	UO	0	1	1	Single-scale object accuracy (SOA). This metric was not proposed to be aggregated for the whole segmentation output, but only for each x_i, which is $SOA_i = \max(SOA_{ij})$
(51) $MOA_i = \max(\{SOA_i\}), size(\{SOA_i\}) = h$	Zhang et al. (2015b).	AB	UO	0	1	1	Multiscale object accuracy (MOA). Metric developed to assess multiscale segmentation, that is, several sets Y are created (Y_1, Y_2,… Y_h), from which a set of *h* metrics SOA_i are calculated for each x_i. SOA_i corresponds to metric 50. Global metric MOA is the weighted mean of all MOA_i, using $area(x_j)$ as weights.
(52) $OSE_i = \begin{cases} \frac{1}{1 - \frac{1}{area(x_i)}} \left(1 - \frac{area(x_i \cap y_j)}{area(x_i)} \right)_{ij}, y_j \in Yg_i \\ 0 \quad , y_j \notin Yg_i \end{cases}$	Su and Zhang (2017).	AB	O	0	1	0	Over-segmentation error (OSE). Global metric OSE (called GOSE) is the weighted mean of all OSE_j. using $area(x_i)$ as weights.

Metric	Reference	Typ.[a]	Err.[b]	Min.	Max.	Opt.	Notes
(53) $USE_i = \frac{\min\left(\begin{array}{l} area(x_i), \\ \left(area(x_i) - \sum_j area(x_i \cap y_j)\right) + \\ \left(\sum_j area(x_i \cup y_j) - area(x_i)\right) \end{array}\right)}{area(x_i)}, y_j \in Yg_i$	Su and Zhang (2017).	AB	U	0	1	0	Under-segmentation error (USE). Global metric USE (called GUSE) is the weighted mean of all USE_j. using $area(x_i)$ as weights.
(54) $OI2_i = \max\left(\frac{area(x_i \cap y_j)}{area(x_i)} \times \frac{area(x_i \cap y_j)}{area(y_j)}\right), y_j \in \tilde{Y}_i$	Yang et al. (2017).	AB	UO	0	1	1	Overlap index (OI2). This metric was not proposed to be aggregated for the whole segmentation output.

[a] area-based (AB) or position-based (PB)

[b] under-segmentation (U), over-segmentation (O), or both (UO)

Table 2. Combined geometric metrics based on those described in Table 1. The information associated with each of the columns is presented as in Table 1. All metrics detect under-segmentation and over-segmentation error.

Combined metric	Reference	Typ.	Min.	Max.	Opt.	Notes
(55) $\mathrm{F-measure} = \frac{1}{\frac{\alpha}{\mathrm{Precision}} + (1-\alpha)\frac{1}{\mathrm{R}}}$	Van Rijsbergen (1979) and Zhang et al. (2015a).	AB	0	1	1	α=0.5 in Zhang et al. (2015a). Further combined metrics based on Precision and recall (metrics 1-2) are found in Zhang et al. (2015a).
(56) $D_{ij} = \sqrt{\frac{OS_{ij}^2 + US_{ij}^2}{2}}$	Levine and Nazif (1982) and Clinton et al. (2010).	AB	0	1	0	Index D (D). Global metric D can be the mean of all D_{ij}. More similar combined metrics are found in Clinton et al. (2010). See metrics 33-34.
(57) $BCE_p = \max\left(LRE(x_i, y_j)_p, LRE(y_j, x_i)_p\right)$	Martin (2003) and Zhang et al. (2015a).	AB	0	1	0	Bidirectional consistency error (BCE). Global metric BCE is the mean of all BCE_p. See metrics 11-12.

Combined metric	Reference	Typ.	Min.	Max.	Opt.	Notes
(58) $CI = \frac{\sum_{i=1}^{k}(C_i \times A_{C_i})}{k}$	Möller et al. (2007).	AB and PB	0	100	100	Comparison index (CI). C_i is the comparison class, which represents clustered and ranked object metrics of over- and under-segmentation such as RAsub and RAsuper (metrics 17-18). C_i can be calculated with a clustering algorithm such as *K*-means. A_{Ci} is equivalent to the proportion of C_i within the reference space.
(59) $ED2 = \sqrt{PSE^2 + NSR^2}$	Liu et al. (2012).	AB	0		0	Euclidian distance 2 (ED2). See metrics 38-39.

Combined metric	Reference	Typ.	Min.	Max.	Opt.	Notes
(60) $M^g = D^- - D^+$	Möller et al. (2013).	AB and PB	0	1	1	D^- and D^+ are the distance between the cumulative distribution functions of metrics G_{ijk}^R and G_{ijk}^F measured by a Kolmogorov–Smirnov test, in which the null hypothesis is that the distribution function of G_{ijk}^R is not less or not greater than that of G_{ijk}^F, respectively. $G_{ijk}^R = \sqrt{O_{ijk}^R \times P_{ijk}^R}$ (see metrics 40 and 42) and $G_{ijk}^F = \sqrt{O_{ijk}^F \times P_{ijk}^F}$ (see metrics 41 and 43).
(61) $ADI_{ij} = \sqrt{OE_{ij}^2 + CE_{ij}^2}$	Cheng et al. (2014).	AB	0		0	Area discrepancy index (ADI). Global metric $ADI_{overall} = \sqrt{OE_{overall}^2 + CE_{overall}^2}$ (see metrics 44-45).

Combined metric	Reference	Typ.	Min.	Max.	Opt.	Notes
(62) $ED3_{ij} = \sqrt{\frac{(OS2_{ij})^2 + (US2_{ij})^2}{2}}$	Yang et al. (2014).	AB	0	1	0	Euclidean distance 3 (ED3). Global metric ED3 is the sum of all summed $ED3_{ij}$ over each x_i. See metrics 47-48. ED3 modified in Yang et al. (2015b).
(63) $M^j = D^- - D^+$	Costa et al. (2015).	AB and PB	0	1	1	M^j is analogous to M^g (metric 60) and D^- and D^+ are the distance between the cumulative distribution functions of metrics J^F_{ijk} and J^R_{ijk} measured by a Kolmogorov–Smirnov test, in which the null hypothesis is that the distribution function of J^F_{ijk} is not less or not greater than that of J^R_{ijk}, respectively. $J^R_{ijk} = \sqrt{G^R_{ijk}}$ and $J^F_{ijk} = \sqrt{G^F_{ijk} \times TSI_j}$ (see metric 49 and notes of metrics 60).
(64) $SEI_i = \begin{cases} ED3_{ij}, y_j \in Yc_i \cap Yd_i \\ 1 \quad , y_j \notin Yc_i \cap Yd_i \end{cases}$	Yang et al. (2015a).	AB	0	1	0	Segmentation evaluation index (SEI). Global metric SEI is the mean of all SEI_i.

Combined metric	Reference	Typ.	Min.	Max.	Opt.	Notes
(65) $BCA(x_i, y_j)_p = BCA(y_j, x_i)_p = \min\left(1 - LRE(x_i, y_j)_p, 1 - LRE(y_j, x_i)_p\right)$	Zhang et al. (2015b).	AB	0	1	0	Bidirectional consistency accuracy (BCA). This metric was not proposed to be aggregated for the global level (see metrics 11, 12, and 66).
(66) $BCA_p = \max\left(\left\{BCA(x_i, y_j)_p\right\}\right), \mathrm{size}\left(\left\{BCA(x_i, y_j)_p\right\}\right) = h$	Zhang et al. (2015b).	AB	0	1	0	Bidirectional consistency accuracy (BCA). Metric developed to assess multiscale segmentation, that is, several sets Y are created (Y_1, Y_2,… Y_h), from which a set of *h* metrics $BCA(x_i,y_j)_p$ (see metric 65) are calculated for each z_p. Global metric BCA is the mean of all BCA_p.