METHODS FOR ASSESSING LOCAL MAP ACCURACY IN THEMATIC CLASSIFICATIONS DERIVED FROM REMOTELY SENSED IMAGES

GeoffreyEdwards
Chaire industrielle en géomatique appliquée à la foresterie
Centre de recherche en géomatique
Pavillon Casault, Université Laval, Québec, Canada, G1K 7P4
fax: (418) 656-7411
email: edwardsg@vm1.ulaval.ca

Abstract: Error evaluation for classified remotely sensed images has been limited in the past to global measures such as the confusion or error matrix, derived from comparing categories in the image with known values. This form of error evaluation gives very little information about spatial errors in the image. As a result, techniques which produce substantially better maps from imagery, such as image segmentation are difficult to evaluate in appropriate ways. Also, it is difficult to determine when one technique (spatially) outperforms another. Qualitative evaluation of the results has been the preferred method, in the absence of more quantifiable techniques. In this paper, work undertaken to characterise boundary errors in maps resulting from photointerpretation is extended to deal with boundary errors in remotely sensed imagery. A new, polygon-specific error (PSE) matrix is introduced, a superset of the traditional error matrix. Combined with estimates of boundary length, five new derivative measures are proposed which characterise the spatial information contained in the image partition with respect to the ground truth partition. The introduction of aggregation indices, in particular, allows the results of segmentation to be evaluated independently of a subsequent classification, and, indeed, to provide information about the information loss involved in the classification procedure. The six measures discussed are all easy to compute, and substantially increase the power of evaluation tools available to the user.

Key words: error evaluation, error matrix, classification, segmentation, map accuracy, boundary error.

1) Introduction

The results of classification of remote sensing imagery have been evaluated by a relatively limited set of tools for more than a decade. Results are reported in terms of an error or confusion matrix which summarizes the relationship between classification categories and ground-truth categories [10]. Other mesures, most of which are derived from the error matrix, include the classification accuracy [10], which is a single number expressing the percent of pixels "correctly" classified out of the total number of pixels; the user's and producer's errors [10], which represent, on the one hand, misclassified ground truth classes, and, on the other, grouped thematic classes; and the kappa coefficient [4,8], which corrects the classification accuracy for errors of commission and omission, and which permits comparison of a given thematic classification with a random classification. These are all global statistics in that they refer to the entire region for which ground truth data are available. They represent summary statistics concerning the thematic information, but they give little information about the spatial precision of the classified regions. Indeed, [9] identified this as a priority research topic. They stated that "techniques need to be developed for assessing the spatial structure of error in an integrated remote sensing classification product, e.g., how are errors related to polygon boundaries."

One of the consequences of this situation is that sophisticated region-based classification techniques, such as segmentation algorithms, are not well evaluated using standard techniques [11]. Segmentation algorithms are designed to extract structural (i.e. spatial) information about the image, and not just thematic information. Applying thematic statistical measures such as the error matrix may result in severly underestimating the performance of these algorithms. Many cases may be cited where reported classification accuracies for segmented images are of the same order as classification accuracies derived from pixel-based classifiers, but where a visual examination of the results reveals that the segment classifier produces much more reliable spatial information than the pixel-based [1]. For example, it is possible to conceive of a situation where a segmentation algorithm produces a totally accurate spatial partition of the image, corresponding exactly to the ground truth, but for which the classification accuracy is zero, if all the regions have been misclassified. Furthermore, many segmentation algorithms produce a spatial partition of the image, previous to carrying out the final thematic classification. Thematic statistics such as the error matrix permit only the combination of the two stages to be evaluated - there is no statistic for the first stage alone. Visual examination, while useful, cannot be applied in a systematic manner to determine map accuracy, or to cross-compare performance among different segmentation algorithms. Suitable quantitative evaluation methods for structural information are sorely lacking.

In this paper, we present new procedures for evaluating map accuracy. These tools are more suitable to the evaluation of segmentation algorithms, for they are based on an evaluation of the spatial structure of image-derived data. The new tools represent an adaption of recent developments in the analysis of boundary uncertainty in photointerpreted images, and hence we briefly introduce the latter before presenting the new remote sensing statistics.

2) Measuring Boundary Uncertainty

Despite the fact that photointerpretation has been one of the principle sources of map data for more than forty years, photointerpretation uncertainty has been poorly understood. This may be due, in part, to the difficulty of obtaining new assessment techniques. Also, until the emergence of Geographic Information Systems (GIS), the need for obtaining accurate local information was probably less pressing than it is today.

A concerted effort over the past few years on the part of myself and my colleagues has led to the development of a number of new methods for assessing uncertainty in photointerpreted imagery [2,3,5,6,7]. These methods are largely based on the analysis of multiple interpretations of the same image. The results of several interpretations are overlayed in a GIS and techniques have been developed for determining a mean interpreted boundary and a boundary width for the overlayed interpretations. The key insight that led to the development of these techniques was that boundary uncertainty requires analysis of the textures on both sides, and hence the basic unit for boundary uncertainty assessment is the boundary and its two adjacent textures, which we collectively call the 'twain'. This focus, although it appears obvious, represents a fundamental shift of attention away from polygons as entities. An analysis of image textures on either side of a given boundary [7] led to the conclusion the automated prediction of boundary uncertainty in photointerpretations, based on the image textures present in each twain, is difficult if not impossible to perform. This is due to the complexity of the factors which influence the determination of boundary uncertainty in human interpretation of images. Hence the work on boundary uncertainty assessment has focussed on empirical tools for characterising "fuzzy boundary width". Two techniques have been developed - one functions essentially with a raster grid space [2,3], while the other is designed to work with vector data [5]. Both techniques have been used on real data, but they give somewhat different results (indeed, they are based on somewhat different assumptions).

3) New Measures for Assessing Local Map Error in Classified Imagery

The key to arriving at a more suitable method for assessing local map accuracy in region-based thematic mapping is similar to that which led to the boundary uncertainty assessment techniques. Standard evaluation of thematic maps is focussed, naturally enough, on map classes. In order to assess local map accuracy, we note that our focus should be on the spatial partitionning of the image, not on the thematic classes. In particular, if we are interest in boundaries, we should examine the polygon pairs on either side of the boundary.

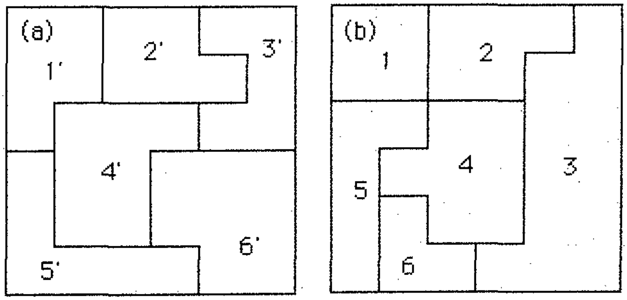

Figure 1: (a) an image partition (6 pixels by 6 pixels); (b) the corresponding ground truth

GT Segments	1	2	3	4	5	6	Total
1'	**4**	-	-	-	1	-	5
2'	-	**4**	1	-	-	-	5
3'	-	1	4	-	-	-	5
4'	-	-	-	**5**	1	**1**	7
5'	-	-	1	-	**3**	2	6
6'	-	-	**6**	2	-	-	8
Total	4	5	12	7	5	3	36

Table 1: Polygon-specific error (PSE) matrix, corresponding to situation shown in Figure 1. Bold numbers indicate diagonal elements of the PSE matrix.

A means to do this consists of constructing a*polygon-specific errormatrix* for the superposition of the classified polygons onto the ground truth polygons (Table 1). Each region in both the region-based classification and the ground truth data must be labelled separately (Figure 1). The classified polygon (hereafter called a "segment") which contain the largest numbers of pixels for each ground truth polygon is first selected as a candidate *ground truth matched* segment. This may be determined from the polygon-specific error matrix (hereafter the PSE matrix) by searching each ground truth column for the largest number of pixels and assigning the correspond segment (row label) to that ground truth polygon (bold numbers in Table 2). More than one ground-truth polygon may be matched to the same segment using this procedure (e.g. GT polygons #4 and #6 are both matched to segment 4' in Table 1), but this will be modified in a subsequent step. The segments assigned in this way to ground truth polygons correspond to the diagonal cells of a traditional error matrix (and the columns may be reshuffled to show this). The so-called off-diagonal elements then correspond to pixels which have been "misclassified" because of boundary errors or boundary uncertainty. However, a traditional error matrix is square, and in fact, a non-square PSE matrix is not easy to interpret. The measures proposed here all require the construction of a square PSE matrix.

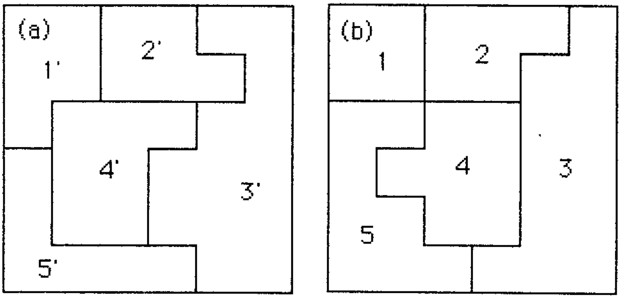

Figure 2: (a) another image partition (6 pixels by 6 pixels); (b) the corresponding ground truth

Let us begin by examining the interpretation of a PSE matrix which is already square (Figure 2 and Table 2). In such a case, the pair of cell values which are matrix transpositions of each other (i.e. symmetrically around the diagonal, the (i,j)th member of the matrix and the (j,i)th member) correspond to boundary displacement on one side or the other of the boundary between two polygons. The off-diagonal value in the same column as the ground-truth polygon corresponds to the displacement error inwards with respect to that polygon, on the part of the matched segment, whereas the transposed value, in the same row as the matched segment, corresponds to the displacement error outwards from the ground-truth polygon. Hence the PSE matrix is seen to present information both about twains and about polygons.

GT Segments	1	2	3	4	5	Total
1'	**4**	-	-	-	-	4
2'	-	**4**	1	-	-	5
3'	-	1	**10**	-	1	12
4'	-	-	2	**5**	-	7
5'	1	-	-	2	**5**	8
Total	5	5	13	7	6	36

Table 2: Polygon-specific error (PSE) matrix, corresponding to situation shown in Figure 2. Bold numbers indicate diagonal elements of the PSE matrix.

If the number of ground-truth polygons exceeds the number of segments, however, then some of the ground-truth polygons will only exhibit inwards boundary displacements, since no matched segment exists. It is possible to "square" the matrix by introducing additional "dummy" segment rows which consist only of zeros, which are matched to the extra ground-truth polygons. It is also possible to reduce the number of ground-truth polygons by merging them. This should only be carried out if there are good reasons for doing so, for instance, if it is decided that thematic differences between two ground-truth classes are not pertinent for a given study.

If the number of segments exceeds the number of ground-truth polygons, then two similar strategies exist to "square" the matrix. The first is to group the segments which have not been assigned a ground-truth polygon together with those that have. This may be done by searching each unassigned segment row for its maximum value and assigning the segment to the corresponding ground-truth column. Spatially, this consists of grouping segments which fall within the same ground-truth polygon. This procedure is consistent with determining the "best possible" boundary match from a given segment partition, presuming the subsequent segment aggregation can be carried out perfectly by a classification algorithm. The second, more conservative procedure consists of grouping all unmatched segments into a single row and matching this to a dummy ground-truth column. This corresponds to an estimate of the boundary error given the existing segmentation.

Once the PSE matrix has been squared, it is possible to determine a global boundary error measure (%BE), similar, in fact, to the global classification error (%CE) determined from the standard error matrix. The %BE consists of the sum of the off-diagonal elements divided by the total number of pixels in the region of interest (i.e. determined from the set of ground-truth polygons) and expressed as a percentage.

The %BE gives a convenient summary statistic for comparing different segment partitions with respect to the ground truth partition. It represents the percentage of pixels which contribute to the boundary error of the given partition. However, like the %CE statistic that it resembles, it says little about the metric size or the spatial distribution of the boundary error.

Therefore, it is useful to determine an additional boundary error statistic which overcomes these shortcomings. Such a boundary statistic may be determined in a direct manner from the PSE matrix. For each pair of ground-truth polygons, the inwards boundary displacement and the outwards boundary displacement are determined (Table 3).

Polygon A	1	1	1	2	2	3	3	4	4	5	Tot
Polygon B	2	4	5	3	4	4	5	5	6	6	-
BE outwards	0	0	1	1	0	2	0	2	2	2	10
BE inwards	0	0	0	1	0	0	1	0	0	0	2
bx=beo-bei	0	0	1	0	0	2	-1	2	2	2	8
bs=beo+bei	0	0	1	2	0	2	1	2	2	2	12
Length	2	0	2	2	2	4	0	2	2	2	18
BX											0,44
BS											0,67

Table 3: Boundary displacement calculation for the situation described in Figure 1 and Table 1.

Furthermore, the length of common boundary can likewise be determined. From these values, it is possible to construct two boundary error statistics, the *boundary displacement (bx)* and the *boundary dispersion (bs)*. The boundary displacement is determined by subtracting the inwards boundary displacement (bei) from the outwards boundary displacement (beo), while the dispersion is determined by adding the two quantities (note, however, that this is not a least squares dispersion measure - its nearest cousin would be a dispersion measure computed from the absolute values of the displacement). When normalised by the length of the common boundary (l), these correspond to metric estimates of boundary error, expressed in pixels or converted to meters. By summing beo and bei across all boundaries, and obtaining the total boundary length (L), two global statistics can be determined:

$$BX = [\Sigma\ (beo) - \Sigma\ (bei)]/L$$

and

$$BS = [\Sigma\ (beo) + \Sigma\ (bei)]/L$$

The values of these statistics are shown in Table 4, for the two simple data sets shown in Figures 1 and 2. Furthermore, it is possible to obtain bx and bs measures for other subsets of the ground-truth partition - for particular pairs of thematic classes, for example, or for particular types of polygons. In this way both twain-specific and polygon-specific boundary error measures may be computed, as well as class-specific boundary error measures.

Note that these measures are similar to those developed for characterising boundary uncertainty in photointerpretation [7]. In general, the bx measures indicate the presence of systematic boundary displacements, while the bs measures indicate random boundary errors. Hence a high bx value means that the bs value is probabily dominated by a systematic boundary displacement, whereas a low bx value combined with a high bs value indicates that there is considerable uncertainty in the boundary placement.

Measure	Figure 1	Figure 2
Total # of pixels	36	36
%BE	39%	22%
BX	0,44	0,25
BS	0,67	0,50
GTAI	0	0
IPAI	1	0

Table 4: Boundary error measures for the image partitions shown in Figure 1 and Figure 2

Finally, the maximum number of segment merges required to square the PSE matrix may be used as an aggregation index (the Image Partition Aggregation Index - IPAI), indicating how far a given partition is from a "best possible" partition, expressed in terms of the number of aggregation operations. Hence, a small B%BE (best %BE, after aggregating segments) may be considered less useful than a larger B%BE if the number of aggregations required to obtain the B%BE is smaller. This statistic takes care of the case of individual pixel classifications, which may have a very good B%BE but will correspond to a high IPAI - that is, they will require a large number of aggregation operations to obtain the BBEM. Likewise, it is useful to record the number of ground truth aggregations as a similar aggregation index, called the GTAI. These aggregation indices, combined with the boundary error measures, may be used as a guiding principle to determine the best post-segmentation *classification* method. Post-segmentation classification attempts to obtain a thematic classification similar to that represented by the ground truth. Most of the time, this entails a loss of information, compared to the original segmented image. This can best be seen by noting that the %BE measure tends to increase with the classification process.

The measures described here should allow us to track the introduction of additional off-diagonal pixels during the grouping process and hence may help us to keep such information loss to a minimum. Post-segmentation classification consists of rules for grouping polygons which are not necessarily spatially contiguous, and hence which do not necessarily have common elements within the appropriate columns. Each grouping of classes results in a change in the aggregation indices described above. Hence a given classification procedure may reduce one aggregation index while increasing the other. Post-segmentation classification needs to look at ways of preserving the existing information as much as possible. Clearly, however, this must be done using information available only in the image and its partition, and not information from the ground truth as such.

The polygon-specific error (PSE) matrix requires a faire amount of storage, especially for large images. However, there are ways to reduce these requirements, since most entries in the PSE matrix will occur between regions which are neighbours, and hence much of the matrix will consist of zeros. As a result, the use of hash tables or other indexing structures to represent the PSE matrix should reduce storage requirements to the minimum necessary. As described above, the boundary displacement and the boundary dispersion measures require the use of boundary length information. Boundary length computations are standard in most GIS and hence could be computed after importing the results of classification into GIS. Alternatively, computing boundary length within a raster environment is not particularly challenging either. The traditional error matrix may be computed from the PSE matrix given rules for grouping ground-truth polygons into classes, as well as rules for grouping image polygons into classes. The set of measures described applies equally well to pixel-based classification (for example, post-classification filtering) as to segment-based classifications, and should allow better quantitative comparison between different algorithms to be performed.

4) Application of the New Measures to a Segmented Scene

In a 1992 paper [1], my students, colleagues and myself published a paper reporting on segmentation experiments in the presence of small agricultural fields. Among other things, we experimented with the inclusion of partial cartographic information (e.g. cadastral boundaries) in order to enhance the mapping accuracy of additional field boundaries. Furthermore, we compared pixel-based classification accuracies and segment-based classification accuracies using the standard error matrices. However, although the classification comparison revealed significant improvements in classification accuracy with the segmentations, we were unable to conclude that the structured segmentation did a better job than the unstructured segmentation; except by visual evaluation.

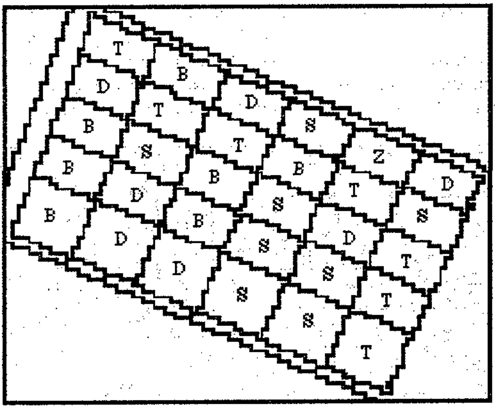

Figure 3: The ground truth for the Morocco SPOT data. The different crops are soft wheat (T), hard wheat (D), sugar beet (S), bersim (B) and alfalfa (Z).

For the purpose of illustrating the techniques proposed in this paper, the first of the two data sets reported in this paper [1] are reconsidered. Figure 3 shows the ground truth polygons for the image, which consisted of a 100 by 100 pixel subset of a SPOT image of a region in Morocco. Figure 4

6) Acknowledgements

This research was supported by the Natural Science and Engineering Research Council of Canada and the Association des industries forestières du Québec through the establishment of an Industrial Chair in Geomatics applied to forestry, of which the author is a member.

References

[1] Ait Belaid, M., Beaulieu, J.-M., Edwards, G., Jaton, A., and Thomson, K.P.B. 1992. Post-Segmentation Classification of Images Containing Small Agricultural Fields. *Geo-Carto*, Volume 7 (3), 53-60.

[2] Aubert, E. 1995. *M.Sc. Thesis*, Laval University, Québec.

[3] Aubert, E., Edwards, G., and Lowell, K.E. 1994. Quantification des erreurs de frontière en photo-interprétation forestière pour le suivi spatio-temporel des peuplements. *Proceedings of the Canadian Conference on GIS*, Ottawa, 195-205.

[4] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Volume 20(1), 37-46.

[5] Edwards, G. 1994a. Characterising and Maintaining Polygons with Fuzzy Boundaries in Geographic Information Systems. *Proceedings of the 6th International Symposium on Spatial Data Handling*, Edinburgh, 223-239.

[6] Edwards, G. 1994b. Characterising Spatial Uncertainty and Variability in Forestry Data Bases. *Proceedings of the ASPRS Conference on Spatial Data Accuracy in Natural Ressource Databases*.

[7] Edwards, G., and Lowell, K.E. 1995. Modeling photo-interpreted boundaries. *Photogrammetric Engineering and Remote Sensing*, In press.

[8] Foody, G.M. 1992. On the Compensation for Chance Agreement in Image Classification Accuracy Assessment. *Photogrammetric Engineering and Remote Sensing*, Volume 58(10), 1459-1460.

[9] Lunetta, R.S., Congalton, R.G., Fenstermaker, L.K., Jensen, J.R., McGwire, K.C., and Tinney, L.R. 1991. Remote Sensing and Geographic Information System Data Integration: Error Sources and Research Issues. *Photogrammetric Engineering and Remote Sensing*. Volume 57(6), 677-687.

[10] Story, M., and Congalton, R. 1986. Accuracy assessment: A user's perspective. *Photogrammetric Engineering and Remote Sensing*, Volume 52(3), 397-399.

[11] Thomson, K.P.B., Edwards, G., Landry, R., Jaton, A., Cadieux, S.-P., and Gwyn, H. 1990. SAR Applications in Agriculture: Multiband Correlation and Segmentation. *Canadian Journal of Remote Sensing*, Volume 16, 47-54.