SPATIAL ASSESSMENT OF CATEGORICAL MAPS A PROPOSED FRAMEWORK

E. Marinho [a, *], D. Fasbender [a], R. De Kok [a]

[a] European Commission Joint Research Centre, 21027, Ispra, Italy - eduardo.marinho@jrc.ec.europa.eu

KEY WORDS: Classification, quality, analysis, accuracy assessment, statistics

ABSTRACT: Meaningful accuracy assessment is a *sine qua non* condition for the use of increasingly precise and available thematic maps. Despite several caveats raised in the literature, the kappa coefficient is still extensively used as an accuracy assessment tool. However, it fails to take into account the spatial nature of remote sensing data, penalizes agreement on area estimation and relies upon hypotheses that are explicitly violated by classification methods such as Geographic Object-Based Image Analysis. Based on the fact that area estimation accuracy measurement is trivial, an innovative and comprehensive framework focusing on the spatial accuracy of categorical maps is built and discussed in this paper. The framework considers not only overlaps between the reference and the classification but also spatial vicinity and does not rely on hypotheses that are in contradiction with the assessment of spatial data. Two class-specific indicators are then designed and the ability of generating probability maps explained. The theoretical developments are followed by simulations illustrating specific cases.

1. INTRODUCTION

In the last decades, remote sensing data gave rise to a plethora of methods for the production of thematic maps. Large areas are now covered with an unprecedented spatial detail. At the same time, there is a constant need for accuracy assessments of the outcomes from automated classifications.

The mainstream validation technique for categorical maps was and still is the kappa coefficient. Initially proposed by Cohen (1960) as a measure of diagnostic agreement between psychologists, this measure has been widely accepted and adopted by the GIS and remote sensing community after its introduction by Congalton and Mead (1983). The idea is to build a confusion matrix summarizing the agreement in terms of spatial overlap (or common categorisation) between a reference dataset, and the thematic map to be assessed.

The reasons for the success of the kappa coefficient are four-fold: (i) when computing the proportion of agreement, it removes from consideration chance agreement, determined by the marginal distributions of the reference and the classification, (ii) it is sensitive to both position agreement (overlap) and area estimation, (iii) it varies over a convenient range going from -1 to 1 and (iv) it has a correlation-like value interpretation where 0 corresponds to the random case, 1 to a perfect agreement and -1 to a perfect disagreement (note that -1 is achieved only in the case of a reversed classification of 2 equiareal classes).

Despite its appeal, caveats have been raised in the literature with respect to kappa's (mis)use. In the field of psychology, Brennan and Prediger (1981) discussed the specific definition of chance agreement that the index relies on: marginals (frequency of each class) are assumed to be fixed. Transposed to the assessment of thematic maps, this means that the proportion of each class is a predefined value (programmed to be reached) rather than an output of the classification. Even if reasonable in some cases, this assumption significantly reduces the scope of application of most classification. Moreover, although the kappa's sensitivity to area estimation hinges on this definition of chance agreement, it is worth noting that the index is negatively - and not positively - related with higher agreement on estimated areas. This can be checked by deriving the kappa with respect to p_c (the proportion of units for which agreement is expected by chance); and holding constant p_o (the proportion of units in which the judges agreed). As p_o is a proportion thus smaller or equal to 1, equation 1 shows that the kappa decreases with increases on p_o. Now, as p_c is the sum of the reference's and classification's marginal products, it increases the higher is the agreement between the marginals which consequently has a negative impact on the kappa. In other words, for a given overlap area (unchanged sum of diagonal entries in the confusion matrix), the kappa will be maximized for higher discrepancies between the estimated area of the classes.

$$\frac{\partial \kappa}{\partial p_c} = \frac{\partial}{\partial p_c}\left(\frac{p_o - p_c}{1 - p_c}\right) = -\frac{(1 - p_o)}{(1 - p_c)^2} < 0 \qquad (1)$$

Among the assumptions that Brennan and Prediger (1981) state as underlying to the computation of the kappa coefficient, the fact that the entries of the confusion matrix should be independent is explicitly violated by Geographic Object-Based Image Analysis (GEOBIA) or any other classification method that integrates contextual information. As pixel labelling depends on its properties as well as the ones from its neighbours, independency is not fulfilled. Finally, the fact that the classification and the reference are treated symmetrically is an issue raised by Brennan and Prediger (1981) that pleads against the utilization of the kappa coefficient for the assessment of thematic maps since the reference is considered as the truth. Interestingly enough, the kappa coefficient has firstly been used by the remote sensing community (Congalton

* Corresponding author.

and Mead, 1983) to test the consistency between photointerpreters, not for accuracy assessment.

Understanding the previously mentioned properties is perhaps enough to discourage assessments of categorical maps to be based on the kappa coefficient. However, in the fields of GIS and remote sensing, the most important criticism to the index is that it neglects the very essential dimension of the data to be evaluated, viz. space. In fact, the confusion matrix between the reference map and the assessed classification focus on the respective proportions of each class and how they overlap, disregarding neighbouring pixels and object morphology. Figure 1 illustrates this problem by showing 5 classifications sharing the same kappa but with markedly different spatial patterns.

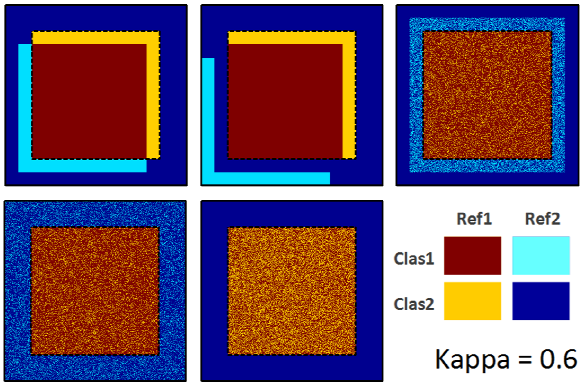

Figure 1. Classifications with identical kappa values but different spatial patterns. Category 1 is represented by red+yellow pixels surrounded by a black dashed square in the reference and by red+cyan pixels in the classification. Agreement over classes 1 and 2 is respectively shown by red and dark blue areas while yellow and cyan highlight disagreements.

Pontius (2000), in an early attempt to remedy this problem, proposed a method to decompose the agreement in "due to chance", "due to quantity" (over/under estimation of the areas) and "due to location". His idea is to use kappa-like coefficients, adjusting the expected agreement proportions due to chance and the maximum possible agreement level given different abilities to correctly estimate the quantities of each class and/or their location. Notwithstanding its undeniable contribution to the improvement of thematic map assessments, this method still relies on the hypothesis of independence of pixels, implies symmetry between the reference and the classification and, with respect to its scope to evaluate the spatial accuracy of the classification, only considers overlaps. Only in 2010, with the development of the object fate analysis by Albrecht *et al.* (2010), has spatial vicinity been explicitly taken into account by applying buffers around the classification and the reference. However, object fate analysis seeks to evaluate uncertainty of object boundaries rather than the spatial validity of a classification. With respect to the hypothesis of independence of the entries in the confusion matrix, the use of objects instead of pixels in the validation process has been proposed (Desclée, 2006; Radoux, 2010) but it only applies if there are thematic but not segmentation differences between the classification and the reference.

Conversely, this paper proposes a framework for the analysis of the spatial accuracy of a classification and distinguishes classifications such as illustrated in Figure1. Since agreement on area estimates is a trivial computation, the proposed method is insensitive to area estimation errors, and consequently makes no assumption concerning the marginals. Instead, it specifically focuses on the ability of the classification to reproduce the spatial pattern of the reference. The index is class specific, which gives additional information about the assessed classification, but can be aggregated in different ways. Furthermore, the approach is not hindered by the assumption of independence of the confusion matrix entries, making it particularly suitable for the assessment of GEOBIAs, and is not symmetric with respect to the reference and the classification. Last but not least, probability maps can be produced over the whole classification area.

2. METHODS

2.1 Theoretical Framework

The theoretical framework exposed here enables the assessment of the spatial validity of a classification (*C*) over an area (*A*) when compared to a reference (*R*). The reference is assumed to be the truth and should fully cover a subset region of the classified area (*As*). *C* and *R* are assumed to be composed of *n* classes, all of them with strictly positive areas over *As*. The set of objects from class *i* within *As* will be noted c_i and r_i for the classification and the reference respectively, while the function *a(.)* will be used to get the surface of any set of objects.

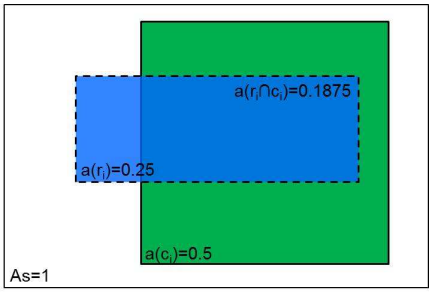

Figure 2. Reference in blue, classification in green

To start, consider the stylized example in Figure 2 representing a reference and a classification over the area *As*. Without loss of generality, let us assume that *As=1*. The class of interest consists of a blue rectangle in the reference (r_i) and a green square in the classification (c_i). The classification overestimates the area of class *i* with $a(c_i)=0.5$ versus $a(r_i)=0.25$. Their overlapping area $a(r_i \cap c_i)$ equals 0.1875 and corresponds to 75% of $a(r_i)$. Now, let us represent c_i in a bi-dimensional space where the x-axis has the proportion of *As* covered by c_i while the y-axis has the proportion of $a(r_i)$ covered by c_i. This point (*Cla*) is represented in Figure 3 with $a(c_i)/As=0.5$ and $a(r_i \cap c_i)/a(r_i)=0.75$. The reference is represented in the same space (*Ref*) with $a(r_i)/As=0.25$ and $a(r_i \cap r_i)/ a(r_i)=1$.

The distance between the two points (*Cla* and *Ref*) is a first quality indicator for the class *i* of the classification *C*: the smaller it is and the better the classification performs. This distance can be decomposed into two effects: (i) over/under estimation of the class area, measured by the distance between

the projections of the points on the x-axis; and (ii) the proportion of the class *i* that *C* could not detect, measured by the distance between the projections of the points on the y-axis.

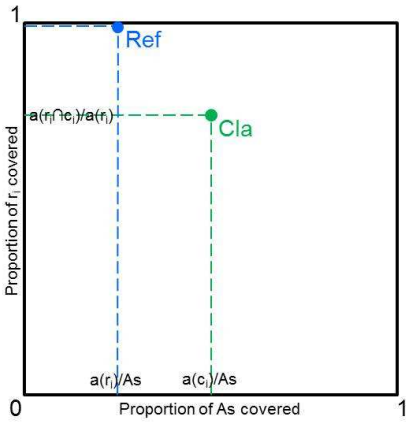

Figure 3. Representation of a reference and a classification.

Although useful, these quality indicators do not measure the spatial agreement between the structures of c_i and r_i. Indeed, the decomposed transition from *Cla* to *Ref* is made by (i) excluding from c_i only wrongly detected objects up to the point $a(c_i)= a(r_i)$ (x-axis transition) and (ii) by replacing the remaining wrong ones by correct detections in a spatial selective process (y-axis transition); and consequently by changing the spatial structure of c_i.

A way to avoid this spatial selection is to make $a(c_i)$ vary through a buffering process such as illustrated in Figure 4. In the image on the left, positive buffers (i.e. inflation) are iteratively applied to c_i elements until r_i is fully contained by the buffered c_i. In the image on the right, $a(c_i)$ is lessened up to the point it equals $a(r_i)$ through negative buffering (i.e. deflation) on c_i. The so buffered c_i will be noted $c_{i,b+}$ and $c_{i,b-}$.

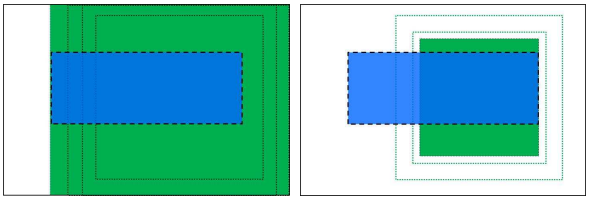

Figure 4. Positive and negative buffering processes - left and right images respectively- over c_i. Positive buffers have been applied until r_i is contained in c_i; negative buffers have been applied until $a(c_i)=a(r_i)$.

The new sets of objects $c_{i,b+}$ and $c_{i,b-}$ can then be represented in the previously used bi-dimensional space. In Figure 5, positively and negatively buffered c_i are represented by the remarkable points P_+ and P_- respectively. Since $c_{i,b+}$ contains r_i, the value of P_+ on the y-axis reaches 1 while in the x-axis appears the proportion of *As* covered by $c_{i,b+}$. In other words, this point informs about how inflated c_i should be in order to avoid omission errors in class *i*. The better is c_i and the smaller are the needed buffers, and consequently the higher is the slope between *Cla* and P_+. Conversely, P_- is the point in the curve for which $a(c_{i,b-})= a(r_i)$.

A continuum of points can be marginally generated in the same way for different sizes of negative and positive buffers, until the bounds of the graph are reached. The lower bound, (0,0), is reached when c_i totally fades away following the successive application of negative buffers and $a(c_i)=0$; while the upper bound, (1,1) is reached when positive buffers are applied up to the moment c_i covers the whole validation area and $a(c_i)=As$. The result of this succession of points forms a buffer curve (BC) in what we name the buffer box (green line in Figure 5).

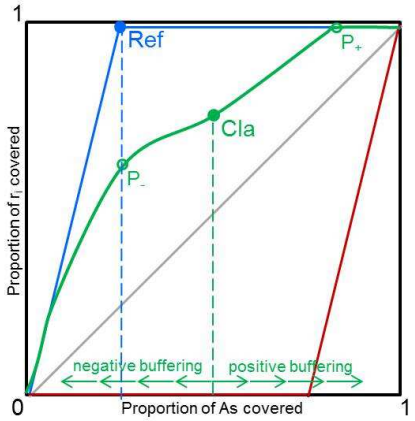

Figure 5. Buffer box representing a reference and a classification.

A similar curve can be generated for r_i with respect to itself and defines the best reachable classification (blue line in Figure 5). Contrariwise, the BC corresponding to the potentially worst c_i is obtained by the application of negative and positive buffers to the complement of r_i over the validation area. It corresponds to the red line in Figure 5 stylized case.

Buffer curves have the following properties for any classification and reference:

1. The curve passes through (0,0), when the applied buffers are negative enough and (1,1) when the buffered class fully covers *As*;

2. It is increasing but not strictly increasing;

3. The case of a random classification corresponds to the diagonal of the box (grey line on Figure 5). As elements of a random c_i are randomly located within the validation area, the proportion of *As* covered by them is expected to be the same than the proportion of r_i they cover. The same applies after the application of buffers;

4. It is insensitive to changes on the marginal class distribution of *C*. Indeed, a change in the marginal that respects the spatial structure of the class *i* merely leads to a displacement of *Cla* on the unchanged buffer curve. The same does not apply for marginal changes on *R* since $a(r_i)$ would be impacted and consequently the y-axis ;

5. The best possible spatial pattern for c_i (blue curve, built by applying buffers on r_i), is defined by two straight lines: from (0,0) to ($a(r_i)/As$,1) and then to (1,1);

6. The worst possible spatial pattern for c_i (red curve, built by applying buffers on the complementary of r_i), is defined by two straight lines: from (0,0) to (1- $a(r_i)/As$,0) and then to (1,1);

7. Its maximum slope over the domain equals $As/a(r_i)$;

8. Its first derivative is sensitive to false and missed objects. False objects tend to have a negative impact on the slope while missed ones, when detected by positive buffers, tend to increase it;

9. Its first derivative implies a probability map (as demonstrated in section 2.3);

10. The reference and the classification are treated asymmetrically. Indeed, a different curve is obtained if the classification is taken as reference and vice-versa;

2.2 Buffered Classification Indexes

From the buffer curve such as designed in the previous section, two class specific indexes can be derived in order to assess its spatial validity: the Absolute Buffered Classification Index (ABCI) and the Relative Buffered Classification Index (RBCI). The ABCI, measures the overall spatial validity of the considered class taking into account how difficult it was to find r_i features; the RBCI, evaluates the spatial validity of c_i taking into consideration the best possible classification for the given reference. Before presenting how they are computed and their properties, it is worth noting that once computed for all classes, each index can be averaged in order to obtain the overall absolute and relative assessments for the spatial validity of the classification over all classes. However, the optimal averaging method to be applied it not discussed in this article.

Absolute Buffered Classification Index (ABCI): has been designed to take into account how difficult it is to find a class. As the spatial pattern of a dominant class, covering per example 90% of *As*, is easy to reproduce, the idea is that the ABCI for this class cannot reach high values. Basically, for such a predominant class, no classification can strongly outperform a random detection of it all over the territory. Conversely, the ABCI should reach values close to 1 only for good performing classifications over very small classes. In other words, the ABCI reaches 1 only when the needle is found in a haystack. This is achieved by the following formula:

$$ABCI_i = 2 * S_i - 1 \quad (1)$$

Where S_i is the integral of the BC over [0-1] such as illustrated by the green zone in Figure 6.

The total area of the buffer box equalling to one, the ABCI tends to 1 if and only if small classes perfectly reproduce the spatial pattern of the reference. In this case the blue and red curves tend to the edges of the buffer box. Indeed, from property 5, if $a(r_i)$ tends to 0 the blue line will go from (0,0) to (0,1) and then to (1,1) while from property 6 the red line will go from (0,0) to (1,0) and then to (1,1). Consequently, if c_i corresponds to r_i, the surface S_i will tend to 1 as well as the $ABCI_i$. Conversely, the ABCI will not substantially deviate from 0 for any classification trying to reproduce a dominant class. Using again properties 5 and 6 it is easy to see that the blue and red lines will tend to the diagonal of the box meaning that S_i will take values close to 0.5 for any c_i. Finally, note that the ABCI can only reach -1 for small classes that are detected everywhere but where they are, in other words if one takes all the hay for needles and the needle for hay.

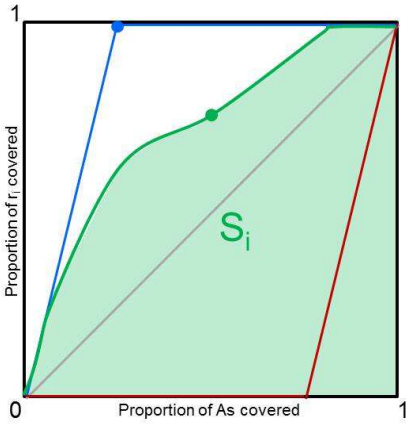

Figure 6. Surface used in to compute the ABCI.

Relative Buffered Classification Index (RBCI): is a way of normalizing the ABCI. While the latter only attributes high values to classes that are difficult to be found, the Relative Buffered Classification Index is computed with respect to the feasibility domain of a given class. This is achieved by subtracting from the integral under the BC, the integral under the red curve (that corresponds to the worst possible spatial pattern) and dividing the result by the integral of the blue curve (best possible spatial pattern) minus the one of the red curve.

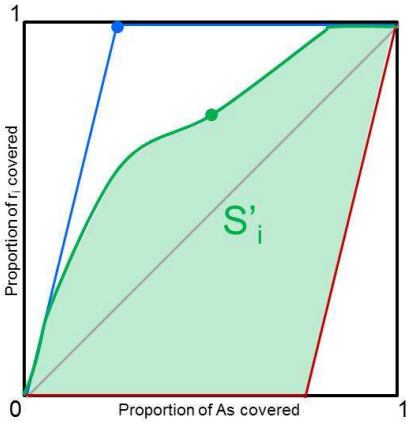

Figure 7. Surface used in to compute the RBCI.

Again, in order to get an index bounded between -1 and 1 the result is multiplied by 2 and -1 is then added:

$$RBCI_i = \frac{2 * S'_i}{(maxS_i - minS_i)} - 1 \quad (2)$$

Where $minS_i$ = the integral below the red line
$maxS_i$ = the integral below the blue line
$S'_i = S_i - minS_i$.

As a result, the RBCI can reach any value between -1 and 1 for any r_i. What matters here is only how close the spatial pattern of c_i is to the one of r_i.

2.3 Probability Maps

Another important advantage of the study of buffer curves (BC), maybe its most innovative one, is that probability maps can be generated for each assessed class. Indeed, the derivative of BC gives the increase of the proportion of r_i covered by c_i with respect to an increase of the proportion of *As* covered by c_i when a buffer is applied to it:

$$BC'_{i,b} = \frac{\partial a(r_i \cap c_{i,b}) / a(r_i)}{\partial a(c_{i,b}) / As} = \frac{\partial a(r_i \cap c_{i,b})}{\partial a(c_{i,b})} * \frac{As}{a(r_i)} \quad (3)$$

Where $BC'_{i,b}$ = derivative of the buffer function of the class i corresponding the the buffer b
$c_{i,b}$ = objects of class i after application of the buffer b

Then, if BC' is multiplied by $a(r_i)/As$, the probability of observing the assessed class i in the additional buffer b is obtained:

$$BC'_{i,b} * \frac{a(r_i)}{As} = \frac{\partial a(r_i \cap c_{i,b})}{\partial a(c_{i,b})} = \Pr_{i,b} \quad (4)$$

Formula 4 gives the proportion of the increase (decrease) of $a(c_{i,b})$ that is covered by r_i, or in another words, the probability that a random point within the marginal buffer has to hit an r_i element.

3. RESULTS

The buffered classification analysis has been used to assess the spatial validity of the 5 simulations presented in the introduction. They are composed of a simple spatial pattern in the reference (square composed by the red and yellow pixels) that different classifications try to reproduce. This simple object has been chosen for simplicity reasons when exposing the spatial properties of the classifications. All classifications have been constructed in order to get the same kappa when compared to the reference. They are presented in Figure 8 with their respective buffer curve and probability maps for category 1 (black dashed square being the reference).

In classification A, a square of the correct size and shape is detected but is penalized by a positioning shift. Classification B is identical to A but with an additional shift of the cyan area corresponding to the commission error. The commission area being further from the object to be detected in classification B than in A, the latter is expected to spatially perform better than the former. In classifications C and D, the proportions of omission and commission errors with respect to the total area are identical to classifications A and B. However, C and D are expected to have a poorer performance since the general salt and pepper shape they detect is bigger than the object corresponding to the class. Finally, classification E correctly detects the square but with a high omission rate. As the proposed indicator is insensitive to area over/under estimations, and focus only on the spatial pattern of the classification, E is expected outperforms the other 4 classifications.

By comparing the buffer curves for A and B, it can be seen that the false cyan detected object is having a negative impact on the slope for positive buffers and that the slope increases when the object merges with the correctly classified object. As expected, the RBCI is higher for classification A (0,91 against 0,89). Lower RBCI values are obtained for classifications C (0.86) and D (0.82) while its value is 0,99 for classification E.

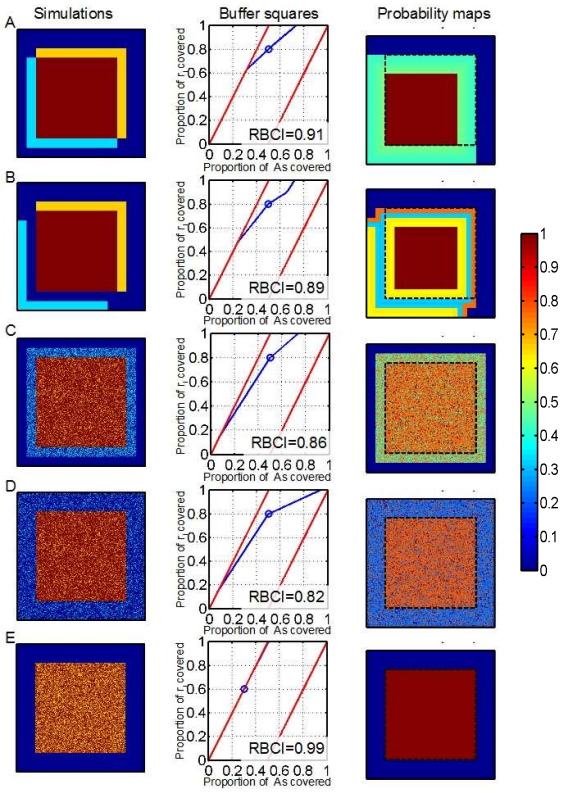

Figure 8. Spatially different classifications (cfr. Figure 1); Associated buffer curves and RBCIs; Probability maps for the assessed class 1 (red+yellow in the reference).

The derived probability maps show the uncertainties generated by errors on the spatial pattern of the classification. They can be decomposed on 3 areas: (i) red, where class 1 has a probability close to 1 of being found, (ii) blue, where class 1 has a

probability close to 0 of being found and (iii), uncertain areas. In classification E, only the first two areas are present. Indeed, despite high omission levels, category 1 detections (c_1) are always correct and randomly distributed over the area actually covered by the category (r_1). The same holds to the area obtained when small buffers are applied to c_1; till the moment it will cover almost the whole r_1 and only r_1. From this point on, the BC derivative will tend to 0 and so will the probability of finding r_1 components. Conversely, in classification A, negative buffers should be applied to c_1 in order to have it fully covered by r_1 and positive ones in order to fully cover r_1. This transition between fully covered and fully covering is represented by a slope of the BC inferior to its theoretical maximum in the buffer square, and by the green area in the probability maps. For more negative buffers, the BC slope equals its theoretical maximum and the probability of finding category 1 becomes 1. For more positive ones, the slope and the same probability both equal 0. It is worth noting that here the probability maps have been generated only in the validation areas As, but the same buffer curves can be used over the whole classified zone A if the spatial validity of the classification is assumed to be homogeneous over space.

4. CONCLUSION

In this paper, we present a new framework for the quality assessment of classification products. Contrary to existing methods (e.g. Cohen's kappa coefficient), the proposed methodology encompasses the spatial position of the classified objects relatively to the reference objects.

The methodology relies on the spatial generalization of objects for a given map. For a given class, the corresponding objects are gradually increased (resp. decreased) until the whole area is covered (resp. empty). A percentage of agreement with the reference map is then computed as the ratio between (i) the surface of agreement with the reference map and (ii) the surface covered by the reference. It is then possible to show evolution of this percentage as a function of the percentage of the total area covered by the generalized objects. By construction, this function is increasing, continuous and remains in the [0,1]x[0,1] square. From this rather simple idea, one can bring a whole quality assessment framework with a set of well-defined properties. Mainly, as it is based on a spatial construction, it allows us to have a spatial representation of the quality assessment, e.g., probability maps.

Although presented here for object-based classifications, it is worth noting that the methodology can also be adapted to pixel-based classifications.

REFERENCES

Albrecht, F., Lang, S., Holbling, D., 2010. Spatial Accuracy Assessment of Object Boundaries for Object-Based Image Analysis. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol.XXXVIII, 4/C7.

Brennan, R., Prediger, D., 1981. Coefficient kappa: Some uses, misuses and alternatives. *Educational and Psychological Measurement*, 41(3), pp. 687-699.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp. 37-46.

Congalton, R.G., Mead, R.A., 1983. A quantitative Method to Test for Consistency and Correctness in Photointerpretation. *Photogrammetric Engineering and Remote Sensing*, 49(1), pp. 96-74.

Desclée, B., Bogaert, P., Defourny, P., 2006. Forest change detection by statistical objecct-based method. *Remote Sensing of Environment*, 102(1-2), pp. 1-11.

Pontius, R., 2000. Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, 66(8), pp. 1011-1016.

Radoux, J., Bogaert, P., Fasbender, D., Defourny, P., 2010, Thematic accuracy assessment of geographic object-based image classification. *International Journal of Geographical Information science*, 25(6), pp. 895-911.

Thoonen, G., Hufkens, K., Vanden Borre, J., Spanhove, T., Scheunders, P., 2012. Accuracy assessment of contextual classification results for vegetation mapping. *International Journal of Applied Earth Observation and Geoinformation*, 15(2012), 7-15.