1-1-2011

Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment

Robert Gilmore Pontius
Clark University, rpontius@clarku.edu

Marco Millones
Clark University

Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment

ROBERT GILMORE PONTIUS JR*† and MARCO MILLONES†

†School of Geography, Clark University, Worcester MA, USA

Corresponding author. Email address: rpontius@clarku.edu

Abstract

The family of Kappa indices of agreement claim to compare a map's observed classification accuracy relative to the expected accuracy of baseline maps that can have two types of randomness: 1) random distribution of the quantity of each category, and 2) random spatial allocation of the categories. Use of the Kappa indices has become part of the culture in remote sensing and other fields. This article examines five different Kappa indices, some of which were derived by the first author in 2000. We expose the indices' properties mathematically and illustrate their limitations graphically, with emphasis on Kappa's use of randomness as a baseline, and the often ignored conversion from an observed sample matrix to the estimated population matrix. This article concludes that these Kappa indices are useless, misleading, and/or flawed for the practical applications in remote sensing that we have seen. After more than a decade of working with these indices, we recommend that the profession abandoned the use of Kappa indices for purposes of accuracy assessment and map comparison, and instead summarize the crosstabulation matrix with two much simpler summary parameters: quantity disagreement and allocation disagreement. This article shows how to compute these two parameters using examples taken from peer-reviewed literature.

Keywords

analysis, classification, error, kappa, matrix, statistics, thematic mapping.

1 Introduction

Proportion of observations classified correctly is perhaps the most commonly used measurement to compare two different expressions of a set of categories, for example to compare land cover categories expressed in a map and to reference data collected for the map's accuracy assessment. There are good reasons for the popularity of the proportion correct measurement. Proportion correct is simple to compute, easy to understand, and helpful to interpret. Nevertheless, it has become customary in the remote sensing literature to report the Kappa index of agreement along with proportion correct, especially for purposes of accuracy assessment, since Kappa also compares two maps that show a set of categories. Kappa is usually attributed to Cohen (1960), but Kappa has been derived independently by others and citations go back many years (Galton 1892, Goodman and Kruskal 1954, Scott 1955). It became popularized in the field of remote sensing and map comparison by Congalton (1981), Congalton *et al.* (1983), Monserud and Leemans (1992), Congalton and Green (1999), Smits *et al.* (1999), and Wilkinson (2005), to name a few. In particular, Congalton and Green (2009) state that "Kappa analysis has become a standard component of most every accuracy assessment (Congalton et al., 1983; Rosenfield and Fitzaptrick-Linz, 1986; Hudson and Ramm, 1987; Congalton 1991) and is considered a required component of most image analysis software packages that include accuracy assessment procedures." Indeed, Kappa is published frequently and has been incorporated into many software packages (Eastman 2009, Erdas Inc 2008, Visser and de Nijs 2006).

The use of Kappa continues to be pervasive in spite of harsh criticisms for decades from many authors (Brennan and Prediger 1981, Aickin 1990, Foody 1992, Ma and Redmond 1995, Stehman 1997, Stehman and Czaplewski 1998, Turk 2002, Jung 2003, Foody 2002, Di Eugenio and Glass 2004, Foody 2004, Allouche *et al.* 2006, Foody 2008). Congalton and Green (2009) acknowledge some of these criticisms, but they report that Kappa "must still be considered a vital accuracy assessment measure". If Kappa were to reveal information that is different from proportion correct in a manner that has implications concerning practical decisions about image classification, then it would be vital to report both proportion correct and Kappa; however, Kappa does not reveal such information. We do not know of any cases where the proportion correct was interpreted, and then the interpretation was changed due to the calculation of Kappa. In the cases that we have seen, Kappa gives information that is redundant or misleading for practical decision making.

Pontius (2000) exposed some of the conceptual problems with the standard Kappa described above and proposed a suite of variations on Kappa in an attempt to remedy the flaws of the standard Kappa. After a decade of working with these variations, we have found that they too posses many of the same flaws as the original standard Kappa. The standard Kappa and its variants are frequently complicated to compute, difficult to understand, and unhelpful to interpret. This paper exposes problems with the standard Kappa and its variations. It also recommends that our profession replace these indices with a more useful and simpler approach that focuses on two components of disagreement between maps in terms of the quantity and spatial allocation of the categories. We hope this paper marks the end of the use of Kappa and the beginning of the use of these two components: quantity disagreement and allocation disagreement.

2 Methods

2.1 Maps to show concepts

We illustrate our points by examining the maps in figure 1. Each map consists of nine pixels and each pixel belongs to either the white category denoted by 0 or the black category denoted by 1. The rectangle with the abbreviation “refer.” in the bottom row indicates the reference map, which we compare to all of the other maps, called the comparison maps. The comparison maps are arranged from left to right in order of the quantity of the black pixels they contain. We can think of this quantity as the amount of black ink used to print the map. We introduce this ink analogy because the analogy is helpful to explain the concepts of quantity disagreement and allocation disagreement. All the maps within a single column contain an identical quantity of black pixels, indicated by the number at the bottom of the column. Within a column, the order of the maps from bottom to top matches the order of the amount of disagreement. Specifically, the maps in the bottom row show an optimal spatial allocation that minimizes disagreement with the reference map, given the quantity of black pixels. While the maps at the top row of each column show a spatial allocation that maximizes disagreement with the reference map, given the quantity of black pixels, i.e., given the amount of black ink in the map. The concepts of quantity and allocation have been expressed by different names in other literature. In the field of landscape ecology, the word “composition” describes the quantity of each category, and the word “configuration” describes the allocation of the categories in terms of spatial pattern (Gergel and Turner 2002, Remmel 2009). In figure 1, each different column has a unique composition of black and white, while there are various configurations within each column. There are a few other possible configurations of black and white to construct the comparison maps in addition to those shown in figure 1; however we do not show those configurations because figure 1 gives a set of comparison maps that demonstrate all possible combinations of quantity disagreement and allocation disagreement.

[Insert figure 1 here]

We define quantity disagreement as the amount of difference between the reference map and a comparison map that is due to the less than perfect match in the proportions of the categories. For example, the reference map in figure 1 has three black pixels and six white pixels. The three comparison maps above the reference map in figure 1 have zero quantity disagreement with the reference map because they also have three black pixels and six white pixels. Each comparison map in a different column than the reference map has positive quantity disagreement, which is equal to the absolute value of the comparison map’s number of black pixels minus three. We can think of quantity disagreement as the difference in the amount of black ink used to produce the reference map versus the amount of black ink used to produce the comparison map. This ink analogy extends to a multi-category case, where each category is a different color of ink.

We define allocation disagreement as the amount of difference between the reference map and a comparison map that is due to the less than optimal match in the spatial allocation of the categories, given the proportions of the categories in the reference and comparison maps. Again, the ink analogy is helpful since we can envision various ways in which the ink can be allocated spatially within the comparison map, where some allocations have a better match with the reference map than other allocations. For example, each column of comparison maps in figure 1 are ordered from bottom to top in terms of increasing allocation disagreement. Allocation disagreement is always an even number of pixels, because allocation disagreement always occurs in pairs of misallocated pixels. Each pair consists of one pixel of omission for a particular category and one pixel of commission for the same category. A pixel is called omission for the black category when the pixel is black in the reference map and not black in the comparison map. A pixel is called commission for the black category when the pixel is black in the comparison map and not black in the reference map. If a comparison map has pixels of both omission and commission for a single category, then it is possible to envision swapping the positions of the omitted and committed pixels within the comparison map so that the rearranged allocation has a better match with the reference map. If it is possible to perform such swapping, then there exists a positive amount of allocation disagreement in the original comparison map (Alo and Pontius 2008). Previous literature calls this type of disagreement “location disagreement”, but we have found that scientists frequently misinterpret this term by calling any disagreement in a map “location disagreement”. Therefore, we recommend that the profession begin using the term “allocation disagreement” instead of “location disagreement”, as this paper does. Figure 1 highlights a particular comparison map that this article uses to explain the concepts in depth. This particular comparison map has one pixel of quantity disagreement and two pixels of allocation disagreement for a total disagreement of three pixels.

2.2 Disagreement space

Figure 2 plots the total disagreement versus the quantity of the black category for the maps in figure 1. Circles denote the maps in the bottom row of figure 1 that have zero allocation disagreement, such that the total disagreement is attributable entirely to the less than perfect match between the reference map and the comparison map in terms of the quantity of black and white pixels. Quantity disagreement is the name for this type of less than perfect match, and it is measured as the distance between the horizontal axis and the diagonally-oriented boundary of quantity disagreement. For all plotted points above the quantity disagreement boundary, the corresponding comparison map contains a positive amount of allocation disagreement. The total disagreement is the sum of the quantity disagreement and the allocation disagreement. In other words, the allocation disagreement is the total disagreement minus the quantity disagreement, as shown in figure 2 for the comparison map highlighted in figure 1. Triangles in figure 2 denote the maps in the top of each column in figure 1, which have the maximum possible allocation disagreement. It is mathematically impossible for any maps to fall outside the rectangle defined by the quantity disagreement and maximum disagreement boundaries. All of the diamonds denote maps that have two pixels of allocation disagreement, and all of the squares denote maps that have four pixels of allocation disagreement. The dashed line in figure 2 shows the statistical expectation of disagreement for a comparison map where the spatial allocation is random, given the quantity of black pixels. The central asterisk shows the statistical expectation of disagreement for a comparison map where both quantity and allocation of the pixels in the comparison map are random.

[Insert figure 2 here]

2.3 Mathematical notation for an unbiased matrix

A crosstabulation matrix is the mathematical foundation of proportion correct and the various Kappa indices. The crosstabulation matrix has many other names, including confusion matrix, error matrix, and contingency table. It is essential that the matrix gives unbiased information concerning the entire study area in order to derive unbiased summary statistics. If reference data are available for all pixels, as is the case in figure 1, then the matrix gives unbiased information concerning the relationship between the reference map and the comparison map, hence the matrix is analyzed directly. However, reference information for an entire study area frequently does not exist in practice due to time limitations, financial constraints, inaccessibility, or unavailability. In those cases, a sampling strategy is typically implemented to collect a sample of reference data from the landscape (Stehman and Czaplewski 1998, Stehman 2009). This subsection gives the mathematical notation for the popular stratified sampling design, where the strata are the categories in the comparison map. We present the mathematics to convert the observed sample matrix into an estimated unbiased population matrix, because we have found that this crucial step is frequently ignored in practice.

In our notation, the number of categories is J, so the number of strata is also J in a typical stratified sampling design. Each category in the comparison map is denoted by an index i, which ranges from 1 to J. The number of pixels in each stratum is N_i. Random selection of the pixels within each stratum assures that the sample from each stratum is representative of that stratum. Reference information is collected for each observation in the sample. Each observation is tallied based on its category i in the comparison map and its category j in the reference information. The number of such observations is summed to form the entry n_{ij} in row i and column j of the sample matrix.

Table 1 gives the matrix for this stratified design. The information within each row is representative of that particular stratum because sampling is random within the stratum, but it does not make sense to compute summary statistics within a column by summing tallies from different rows in table 1, because the sampling intensity might be different in each row. In particular, the proportion correct and producer's accuracies are likely to be biased when they are computed directly from the entries in the sample matrix of table 1. It is necessary to convert the sample matrix into a matrix that represents the entire study area in order to compute unbiased summary statistics. Table 2 accomplishes this goal by applying equation 1 to express each entry p_{ij} as the estimated proportion of the study area that is category i in the comparison map and category j in the reference landscape. Thus table 2 gives unbiased estimates of the proportions for the entire study area, so table 2 can be used to compute unbiased summary statistics, including proportion correct, the various Kappa indices, omission error, commission error, producer's accuracy, user's accuracy, quantity disagreement, and allocation disagreement.

$$p_{ij} = \left(\frac{n_{ij}}{\sum_{j=1}^{J} n_{ij}}\right)\left(\frac{N_i}{\sum_{i=1}^{J} N_i}\right) \quad \text{equation 1}$$

[Insert table 1 here]

[Insert table 2 here]

2.4 Parameters to summarize the population matrix

There are numerous possible parameters to summarize the information in the population matrix (Ma and Redmond 1995, Fielding and Bell 1997, Stehman 1997, Liu et al. 2007). This article focuses on the Kappa indices of agreement and two simpler measures: quantity disagreement and allocation disagreement (Pontius 2000, Pontius 2002, Pontius and Suedmeyer 2004, Pontius *et al.* 2007). All the calculations derive directly from the proportions in table 2. Equation 2 computes the quantity disagreement q_g for an arbitrary category g, since the first summation in equation 2 is the proportion of category g in the reference map and the second summation is the proportion of category g in the comparison map. Equation 3 computes the overall quantity disagreement Q incorporating all J categories. Equation 3 must divide the summation of the category-level quantity disagreements by two, because an overestimation in one category is always accompanied by an underestimation in another category, so the summation double counts the overall quantity disagreement. For the example in figure 1, the overall quantity disagreement is equal to the quantity disagreement for black plus the quantity disagreement for white, then divided by two. Equation 4 computes the allocation disagreement a_g for an arbitrary category g, since the first argument within the minimum function is the omission of category g and the second argument is the commission of category g. The multiplication by two and the minimum function are necessary in equation 4, because allocation disagreement for category *g* comes in pairs, where commission of *g* is paired with omission of *g*, so the pairing is limited by the smaller of commission and omission (Pontius *et al.* 2004). Equation 5 gives the overall allocation disagreement *A* by summing the category-level allocation disagreements. Equation 5 divides the summation by two because the summation double counts the overall allocation difference, just as the summation of equation 3 double counts the overall quantity difference. Equation 6 computes the proportion correct *C*. Equation 7 shows how the total disagreement *D* is the sum of the overall quantity disagreement and overall allocation disagreement. The appendix gives a mathematical proof of equation 7.

$$q_g = \left|\left(\sum_{i=1}^{J} p_{ig}\right) - \left(\sum_{j=1}^{J} p_{gj}\right)\right| \qquad \text{equation 2}$$

$$Q = \frac{\sum_{g=1}^{J} q_g}{2} \qquad \text{equation 3}$$

$$a_g = 2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right) - p_{gg},\ \left(\sum_{j=1}^{J} p_{gj}\right) - p_{gg}\right] \qquad \text{equation 4}$$

$$A = \frac{\sum_{g=1}^{J} a_g}{2} \qquad \text{equation 5}$$

$$C = \sum_{j=1}^{J} p_{jj} \qquad \text{equation 6}$$

$$D = 1 - C = Q + A \qquad \text{equation 7}$$

Equations 8 – 10 begin to construct the calculations to compute the Kappa indices. Equation 8 gives the expected agreement e_g for category *g*, assuming random spatial allocation of category *g* in the comparison map, given the proportions of category *g* in the reference and comparison maps. Equation 9 gives the overall expected agreement *E* assuming random spatial allocation of all categories in the comparison map, given the proportions of those categories in the reference and comparison maps. Equation 9 defines *E* for convenience because *E* is eventually used in the equations for some of the Kappa indices. Equation 10 defines the overall expected disagreement *R* as equal to 1 – *E*, so we can express the Kappa indices as ratios of disagreement, as opposed to ratios of agreement, which will be helpful when we explain the figures in the results section.

$$e_g = \left(\sum_{i=1}^{J} p_{ig}\right)\left(\sum_{j=1}^{J} p_{gj}\right) \quad \text{equation 8}$$

$$E = \sum_{g=1}^{J} e_g \quad \text{equation 9}$$

$$R = 1 - E \quad \text{equation 10}$$

Equations 11-15 define five types of Kappa indices. Each Kappa is an index that attempts to describe the observed agreement between the comparison map and the reference map on a scale where one means that the agreement is perfect and zero means that the observed agreement is equivalent to the statistically expected random agreement. Some Kappa indices accomplish this goal better than others. Equation 11 defines the standard Kappa κ_{standard} first as a ratio of agreement using *C* and *E*, then as a ratio of disagreement using *R* and *D*. The standard Kappa can be initially appealing to many authors because Kappa is usually defined in the literature as an index of agreement that accounts for the agreement due to chance, meaning that Kappa compares the observed accuracy of the classification to the expected accuracy of a classification that is generated randomly. However, this definition is only partially true, and this imprecise definition has caused tremendous confusion in the profession. A more complete description is that the standard Kappa is an index of agreement that attempts to account for the expected agreement due to random spatial reallocation of the categories in the comparison map, given the proportions of the categories in the comparison and reference maps, regardless of the size of the quantity disagreement. Equation 12 defines Kappa for no information κ_{no}, which is identical to κ_{standard}, except that $1/J$ is substituted for E. The motivation to derive Kappa for no information is that $1/J$ is the statistically expected overall agreement when both the quantity and allocation of categories in the comparison map are selected randomly (Brennan and Prediger 1981, Foody 1992). Equation 13 defines Kappa for allocation $\kappa_{\text{allocation}}$, which is identical to κ_{standard}, except that $(1\text{-}Q)$ is substituted for 1 in the denominator. The motivation to derive $\kappa_{\text{allocation}}$ is to have an index of pure allocation, where one indicates optimal spatial allocation as constrained by the observed proportions of the categories, and zero indicates that the observed overall agreement is equal the agreement expected under random spatial reallocation within the comparison map given the proportions of the categories in the comparison and reference maps (Brennan and Prediger 1981, Pontius 2000). Equation 14 defines κ_{histo}, which is identical in format to κ_{standard}, except $1\text{-}Q$ is substituted for C (Hagen 2002). The name κ_{histo} reflects that κ_{histo} is a function of the histogram of the matrix's marginal totals, i.e., the proportions of the categories. The derivation of κ_{histo} represents an effort to separate the concepts of quantity and allocation, since κ_{histo} multiplied by $\kappa_{\text{allocation}}$ equals κ_{standard}. Equation 15 defines Kappa for quantity κ_{quantity} in a format similar to the other Kappa indices, meaning that κ_{quantity} is a ratio of differences. However, the terms that generate the differences are complex, as shown in equations 16 and 17 and as explained in Pontius (2000). The original motivation to derive κ_{quantity} was to have an index of pure quantity, analogous to how $\kappa_{\text{allocation}}$ describes that accuracy of the allocation, in the context of land change modeling. Table 3 summarizes conceptually the meaning of each ratio for each Kappa index in the context of figures 2-9.

$$\kappa_{\text{standard}} = \frac{C-E}{1-E} = \frac{(1-Q-A)-(1-R)}{1-(1-R)} = \frac{R-(Q+A)}{R} = \frac{R-D}{R} \qquad \text{equation 11}$$

$$\kappa_{\text{no}} = \frac{C-(1/J)}{1-(1/J)} = \frac{(1-Q-A)-(1/J)}{(1-1/J)} = \frac{(1-1/J)-(Q+A)}{(1-1/J)} = \frac{(1-1/J)-D}{(1-1/J)} \qquad \text{equation 12}$$

$$\kappa_{\text{allocation}} = \frac{C-E}{(1-Q)-E} = \frac{(1-Q-A)-(1-R)}{(1-Q)-(1-R)} = \frac{R-(Q+A)}{R-Q} = \frac{R-D}{R-Q} \qquad \text{equation 13}$$

$$\kappa_{\text{histo}} = \frac{(1-Q)-E}{1-E} = \frac{(1-Q)-(1-R)}{1-(1-R)} = \frac{R-Q}{R} \qquad \text{equation 14}$$

$$\kappa_{\text{quantity}} = \frac{C-Z}{Y-Z} \qquad \text{equation 15}$$

$$Y = \left\{\sum_{j=1}^{J}\left[\left(\sum_{i=1}^{J} p_{ij}\right)^2\right]\right\} + \kappa_{\text{allocation}}\left\{1 - \sum_{j=1}^{J}\left[\left(\sum_{i=1}^{J} p_{ij}\right)^2\right]\right\} \qquad \text{equation 16}$$

$$Z = \{1/J\} + \kappa_{\text{allocation}}\left\{\sum_{j=1}^{J} \min\left[(1/J), \sum_{i=1}^{J} p_{ij}\right] - (1/J)\right\} \qquad \text{equation 17}$$

[Insert table 3 here]

2.5 Application to published matrices

All the parameters in this article derive entirely from the crosstabulation matrix, so we can compute the statistics easily for cases where authors publish their matrices. We compute the two components of disagreement and the standard Kappa index of agreement for five examples taken from two articles in International Journal of Remote Sensing to show how the concepts work in practice.

Ruelland et al. (2008) analyzed six categories in West Africa for three points in time: 1975, 1985, and 2000. The comparison maps derive from Landsat data and a recursive thresholding algorithm that seeks to maximize overall accuracy and κ_{standard}.

The reference data consist of control points that were selected based on practical criteria, such as being invariant since the 1970s and being close to trails. The paper does not contain sufficient information to understand whether the sample is representative of the population, and the authors preformed no conversion from the observed sample matrices to estimated population matrices. The paper does not report any Kappa indices. The paper reports percent agreement in terms that imply that the overall % of the reference data that disagrees with the map of 1975, 1985, and 2000 is respectively 24, 28, and 21. The paper then analyzes the net quantity differences among the maps' categories over the three years, and reports that there is 4.5 % net quantity difference between the map of 1985 and 2000. Thus the reported overall error in each map is about five times larger than the size of the reported difference between the maps. The paper states "results indicate relatively good agreement between the classifications and the field observations", but the paper never defines a criterion for relatively good. Our results section below reveals the insight that is possible when one examines the two components of disagreement.

Wundrum and Löffler (2008) analyze five categories in the Norwegian mountains using two matrices that derive from a supervised method and an unsupervised method of classification. The paper reports that 256 reference data points were collected randomly, in which case the summary statistics that derive from the sample matrices are unbiased. The paper reports κ_{standard} for each method, and interprets κ_{standard} by saying that the value is higher for the unsupervised method, which the reported overall proportion correct already reveals. The paper's tables show 34 % error for the supervised classification and 23 % error for the unsupervised classification, and the paper reports "The results of supervised and unsupervised vegetation classification were not consistently good", but the paper never defines a quantitative criterion for not consistently good. The results section compares Wundrum and Löffler (2008) to Ruelland et al. (2008) with respect to components of disagreement and κ_{standard}.

3 Results

3.1 Fundamental Concepts

We analyze figure 1 by plotting results in a space similar to figure 2. In figures 3-9, the vertical axis is the proportion disagreement between the comparison map and the reference map, the horizontal axis is the proportion black in the comparison map, and each number plotted in the space is an index's value for a particular comparison map. Q from equation 3 defines the quantity disagreement boundary, R from equation 10 defines the random allocation line, and D from equation 4 defines the vertical coordinate for the plotted value for each comparison map. The value at coordinates (0.22, 0.33) is the highlighted comparison map from figure 1, which we use to help to explain the results. Figure 3 shows the quantity disagreement Q plotted in this space. There is a column of zeros where the quantity in the black category is one third, because the reference map has three black pixels among its nine pixels. The numbers within each column are identical in figure 3 because the quantity disagreement is dictated completely by the proportion of the black category in each comparison map. Figure 4 shows the allocation disagreement A, which measures the distance above the quantity disagreement boundary. Quantity disagreement and allocation disagreement sum to the total disagreement D.

[Insert figures 3-4 here]

Figure 5 shows results for κ_{standard}. Values are positive below the random allocation line, zero on the line, and negative above the line, by design of the formula for κ_{standard}. The highlighted comparison map in figure 1 has $\kappa_{\text{standard}} = 0.18$, which is a ratio with a numerator of 0.41 – 0.33 and a denominator of 0.41, according to equation 11 and the vertical intervals between the left ends of the braces in figure 5. A single row of numbers in figure 5 contains different values for κ_{standard}, which indicates that κ_{standard} does not give the same result for comparison maps that have the same amount of total disagreement with the reference map. For example, κ_{standard} ranges from -0.36 to 0.12, when total disagreement is 0.56, i.e., when five of the nine pixels disagree. This range shows how κ_{standard} can indicate allocation disagreement more than quantity disagreement. The value of -0.36 shows how κ_{standard} does not reward for small quantity disagreement and penalizes strongly for allocation disagreement, and the 0.12 shows how κ_{standard} does not penalize strongly for large quantity disagreement and rewards for small allocation disagreement (Pontius 2000).

[Insert figure 5 here]

Figure 6 gives κ_{no}, which indicates where the comparison maps' total disagreement is relative to $1/J$, which is 0.5 in the case study that has two categories. If disagreement is zero, then κ_{no} is one; if disagreement is less than 0.5, then κ_{no} is positive; if disagreement is greater than 0.5, then κ_{no} is negative. κ_{no} has the same value within any given row of numbers in figure 6, because κ_{no} is a linear function of total disagreement. The highlighted comparison map has $\kappa_{\text{no}} = 0.33$, which is a ratio with a numerator of 0.50

– 0.33 and a denominator of 0.50, according to equation 12 and the vertical intervals within the braces of figure 6.

[Insert figure 6 here]

Figure 7 gives $\kappa_{\text{allocation}}$. If allocation disagreement is zero, then $\kappa_{\text{allocation}}$ is one. $\kappa_{\text{allocation}}$ is positive below the random allocation line, zero on the random allocation line, and negative above the random allocation line. When the proportion black is zero or one, then $\kappa_{\text{allocation}}$ is undefined, because the concept of allocation has no meaning when one category occupies the entire map. The highlighted comparison map has $\kappa_{\text{allocation}} = 0.25$, which is a ratio with a numerator of 0.41 – 0.33 and a denominator of 0.41, according to equation 13 and the braces in figure 7.

[Insert figure 7 here]

Figure 8 gives results for κ_{histo}. The values are identical within each individual column, because κ_{histo} is a function exclusively of the quantity disagreement boundary *Q* and the random allocation line *R*. Furthermore *R* is a function of only the quantity of each category in the reference and comparison maps. κ_{histo} is one when quantity disagreement is zero, and κ_{histo} is zero when the comparison map consists of entirely one category. κ_{histo} is never negative, so κ_{histo} does not have the characteristic that negative values indicate worse than random agreement. κ_{histo} is not equivalent to quantity disagreement, because κ_{histo} treats an overestimation of the quantity of a category differently than an underestimation. Consider the row of values where proportion disagreement is 0.33. When the comparison map has three fewer black pixels than the reference map, κ_{histo} is zero; but when the comparison map as three more black pixels than the reference map, then κ_{histo} is 0.4. The highlighted comparison map has κ_{histo} = 0.73, which is a ratio with a numerator of 0.41 – 0.11 and a denominator of 0.41, according to equation 13 and figure 8.

[Insert figure 8 here]

Figure 9 gives κ_{quantity}. A single column contains different values, which indicates that κ_{quantity} is not a function exclusively of the quantity disagreement. For example, κ_{quantity} ranges from -0.25 to 0.27 when proportion black in the comparison map is 0.22, i.e., when there is one less black pixel in the comparison map than in the reference map. When quantity disagreement is zero, κ_{quantity} ranges from 0 to 1. κ_{quantity} is undefined when the comparison map is either all black or all white, in spite of the fact that quantity disagreement has a clear interpretation at those points. These counterintuitive characteristics of κ_{quantity} relate in part to the fact that κ_{quantity} was originally derived to inform predictive land change modeling, and not for simple map comparison or accuracy assessment (Pontius 2000). κ_{quantity} attempts to assess how accurate the specification of quantity is in the comparison map, while taking into consideration a land change model's ability to predict the spatial allocation. The highlighted comparison map has κ_{quantity} = 0.73, which is a ratio with a numerator of 0.67 – 0.58 and a denominator of 0.89 – 0.58, according to equation 14 and figure 9.

[Insert figure 9 here]

3.2 Applications to peer-reviewed literature

Figure 10 shows the two components of disagreement and κ_{standard} for five matrices in peer-reviewed literature. The two components of disagreement are stacked to show how they sum to the total disagreement, thus the figure conveys information about proportion correct, since proportion correct is 1 minus the total proportion disagreement.

The results for Ruelland et al. (2008) show that the relative ranking of κ_{standard} is identical to the relative ranking of proportion correct among their three matrices, which demonstrates how κ_{standard} frequently conveys information that is redundant with proportion correct. Each bar for Ruelland et al. (2008) also demonstrates that quantity disagreement accounts for less than a quarter of the overall disagreement. This is important because one of the main purposes of their research is to estimate the net quantity of land cover change among the three points in time, in which case allocation disagreement is much less important than quantity disagreement. The separation of the overall disagreement into components of quantity and allocation reveals that their maps are actually much more accurate for their particular purpose than implied by the reported overall errors of more than 20 %. The κ_{standard} indices do not offer this type of insight.

Figure 10 demonstrates some additional characteristics of κ_{standard} described above. Specifically, the Ruelland et al. (2008) application to 1985 has 25 % total disagreement and the Wundram and Löffler (2008) application to the unsupervised case has 23 % total disagreement, while κ_{standard} for both is 0.65. κ_{standard} fails to reveal that the Wundram and Löffler (2008) application to unsupervised classification has more quantity disagreement than the Ruelland et al. (2008) application to 1985. Quantity disagreement accounts for more than a quarter of the total disagreement within the Wundram and Löffler (2008) application to unsupervised classification, which is important to know for practical applications, but κ_{standard} is designed neither to penalize substantially for large quantity disagreement nor to reward substantially for small quantity disagreement.

[Insert figure 10 here]

4 Discussion

4.1 Reasons to abandon Kappa

We have revealed several detailed reasons why it is more helpful to summarize the crosstabulation matrix in terms of quantity disagreement and allocation disagreement, as opposed to proportion correct or the various Kappa indices. This discussion section provides three main overarching rationales.

First, each Kappa index is a ratio, which can introduce problems in calculation and interpretation. If the denominator is zero, then the ratio is undefined, so interpretation is difficult or impossible. If the ratio is defined and large, then it is not immediately clear whether the ratio's size is attributable to a large numerator or a small denominator. Conversely, when the ratio is small, it is not clear whether the ratio's size is attributable to a small numerator or a large denominator. In particular, κ_{quantity} can demonstrate this problem, in some cases leading to nearly uninterpretable values of κ_{quantity} that are less than negative 1 or greater than 1 (Schneider and Pontius 2001). Kappa's ratio is unnecessarily complicated because usually the most relevant ingredient to Kappa is only one part of the numerator, i.e., the total disagreement as seen in the right sides of equations 11-14. This total disagreement can be expressed as the sum of two components of quantity disagreement and allocation disagreement in a much more interpretable manner than Kappa's unitless ratio, since both components express a proportion of the study area.

Second, it is more helpful to understand the two components of disagreement than to have a single summary statistic of agreement when interpreting results and devising the next steps in a research agenda. The two components of disagreement begin to explain the reasons for the disagreement based on information in the matrix. Examination of the relative magnitudes of the components can be used to learn about sources of error. A statement that the overall Kappa is *X* or proportion correct is *P* does not give guidance on how to improve the classification, since such statements offer no insight to the sources of disagreement. When one shifts focus from overall agreement to components of disagreement, it orients one's mind in an important respect. For example, Ruelland et al. (2008) report that an agreement of 72 % is good, while Wundram and Loffler (2008) report that a disagreement of 23 % is not good. Perhaps they came to these conclusions because Ruelland et al. (2008) focused on agreement and Wundram and Loffler (2008) focused on disagreement. It is much more common in the culture of remote sensing to report agreement than disagreement, which is unfortunate. If Ruelland et al. (2008) would have examined the two components of disagreement, then they could have interpreted the accuracy of their maps relative to their research objective, which was to examine the differences among maps from three points in time. It is usually more helpful to focus on the disagreement and to wonder how to explain the error, which is what the two components of disagreement do, rather than to focus on the agreement and to worry that randomness might explain some of the correctness, which is what the Kappa indices of agreement do.

Third, and most importantly, the Kappa indices attempt to compare observed accuracy relative to a baseline of accuracy expected due to randomness, but in the applications that we have seen, randomness is an uninteresting, irrelevant, and/or misleading baseline. For example, the κ_{standard} addresses the question, “What is the observed overall agreement relative to the statistically expected agreement that we would obtain by random spatial reallocation of the categories within the comparison map, given the proportions of the categories in the comparison and reference maps, regardless of the size of the quantity disagreement?” κ_{standard} answers this question on a scale where zero indicates that the observed agreement is equal to the statistically expected agreement due to random spatial reallocation of the specified proportions of the categories, and one indicates that the observed agreement derives from perfect specification of both the spatial allocation and the proportions of the categories. We cannot think of a single application in remote sensing where it is necessary to know the answer to that question as measured on that scale in order to make a practical decision, especially given that a simpler measure of accuracy, such as proportion correct, is already available. We know of only two cases in land change modeling where $\kappa_{\text{allocation}}$ can be somewhat helpful (Pontius *et al.* 2003, Pontius and Spencer 2005), because $\kappa_{\text{allocation}}$ answers that question on a scale where zero indicates that the observed agreement is equal to the statistically expected agreement due to random spatial reallocation of the specified proportions of the categories, and one indicates that the observed agreement is due to optimal spatial allocation of the specified proportions of the categories. Furthermore, we know of no papers where the authors come to different conclusions when they interpret proportion correct vis-à-vis κ_{standard}, which makes us wonder why authors usually present both proportion correct and κ_{standard}.

We suspect the remote sensing profession is enamored with κ_{standard} because the comparison to a baseline of randomness, i.e., chance, is a major theme in university courses concerning statistical theory, so the concept of κ_{standard} sounds appealing initially. However, comparison to randomness in statistical theory is important when sampling, but sampling is an entirely different concept than the selection of a parameter to summarize a crosstabulation matrix. The Kappa indices are parameters that attempt to account for types of randomness that are conceptually different than the randomness due to sampling. Specifically, if the underlying matrix derives from a sample of the population, then each different possible sample matrix (Table 1) might produce a different estimated population matrix (Table 2), which will lead to different a different statistical value for a selected parameter. The sampling distribution for that parameter indicates the possible variation in the values due to the sampling procedure. We have not yet derived the sampling distributions for quantity disagreement and allocation disagreement, which is a potential topic for future work.

4.2 A more appropriate baseline

There is a clear need to have a baseline for an accuracy assessment of a particular classified map. The unfortunate cultural problem in the remote sensing community is that 85 % correct is frequently used as a baseline for a map to be considered good. It makes no sense to have a universal standard for accuracy in practical applications (Foody 2008), in spite of temptations to establish such standards (Landis and Koch 1977, Monserud and Leemans 1992), because a universal standard is not related to any specific research question or study area. Perhaps some investigators think κ_{standard} avoids this problem, because randomness can generate a baseline value that reflects the particular case study. However, the use of any Kappa index assumes that randomization is an appropriate and important baseline. We think that randomness is usually not a reasonable baseline, because a reasonable baseline should reflect the alternative second-best method to generate the comparison map, and that second-best method is usually not randomization. So, what is an appropriate baseline? The baseline should be related to a second-best method to create the comparison map in a manner that uses the calibration information for the particular study site in a quick and/or naïve approach.

For example, Wu *et al.* (2009) compared eight mathematically sophisticated methods to generate a map of nine categories. If both quantity and allocation were predicted randomly, then the completely random prediction would have a proportion correct of 1/9 (Brennan and Prediger 1981, Foody 1992); however the authors wisely did not use this random value as a baseline. They intelligently used two naïve methods to serve as baselines in a manner that considered how they separated calibration data from validation data. The calibration data consisted 89 % of a single category. Thus one naïve baseline was to predict that all the validation points were that single category, which produced a baseline with 11 % quantity disagreement and zero allocation disagreement. A second naïve baseline was to predict that each validation point was the same category as the nearest calibration point, which produced a second baseline with almost zero quantity disagreement and 20 % allocation disagreement. Only one of the eight mathematically sophisticated methods was more accurate than both of the naïve baselines, while seven of the eight sophisticated models were more accurate than a completely random prediction.

Pontius *et al.* (2007) presented an example from land change modeling in the Amazon where a naïve model predicted that deforestation occurs simply near the main highway, and a null model predicted that no deforestation occurs. Both the naïve and the null models were more accurate than a prediction that deforestation occurs randomly in space. They concluded that the question “How is the agreement less than perfect?” is an entirely different and more relevant question than “Is the agreement better than random?” The components of disagreement answer the more important former question, while the Kappa indices address the less important latter question.

The two components of disagreement have many applications regardless of whether the components derived from a sample of the population, or from comparison of maps that have complete coverage. For example, Pontius et al. (2008a) show how to use the components for various types of map comparisons, while Pontius et al. (2008b) show how to compute the components for maps of a continuous real variable.

5 Conclusions

This article reflects more than a decade of research on the Kappa indices of agreement. We have learned that the two simple measures of quantity disagreement and allocation disagreement are much more useful to summarize a crosstabulation matrix than the various Kappa indices for the applications that we have seen. We know of no cases in remote sensing where the Kappa indices offer useful information, because the Kappa indices attempt to compare accuracy to a baseline of randomness, but randomness is not a reasonable alternative for map construction. Furthermore, some Kappa indices have fundamental conceptual flaws, such as being undefined even for simple cases, or having no useful interpretation. The first author apologizes for publishing some of the variations of Kappa in 2000 and asks that the professional community do not use them. Instead, we recommend that the profession adopt the two measures of quantity disagreement and allocation disagreement, which are much simpler and more helpful for the vast majority of applications. These measurements can be computed easily by entering the crosstabulation matrix into a spreadsheet available for free at www.clarku.edu/~rpontius. These two measurements illuminate a much more enlightened path, as we look forward to another decade of learning.

Acknowledgements

The United States' National Science Foundation (NSF) supported this work through its Coupled Natural Human Systems program via grant BCS-0709685. NSF supplied additional funding through its Long Term Ecological Research network via grant OCE-0423565 and a supplemental grant DEB-0620579. Any opinions, findings, conclusions, or recommendation expressed in this paper are those of the authors and do not necessarily reflect those of the funders. Clark Labs produced the GIS software *Idrisi*, which computes the two components of disagreement that this paper endorses. Anonymous reviewers supplied constructive feedback that helped to improve this paper.

Appendix

This is a mathematical proof of equation 7. We begin with equation 6 that expresses overall total disagreement D as 1 minus the overall total agreement C, then multiply and divide by 2, and then use the fact the sum of all p_{ij} equals 1.

$$D = 1 - C = 1 - \sum_{g=1}^{J} p_{gg} = \frac{2 - 2\left(\sum_{g=1}^{J} p_{gg}\right)}{2}$$

$$= \frac{\left\{\left[\sum_{j=1}^{J}\left(\sum_{i=1}^{J} p_{ij}\right)\right] + \left[\sum_{i=1}^{J}\left(\sum_{j=1}^{J} p_{ij}\right)\right]\right\} - \left(2\sum_{g=1}^{J} p_{gg}\right)}{2}$$

$$= \frac{\sum_{g=1}^{J}\left\{\left(\sum_{i=1}^{J} p_{ig}\right) + \left(\sum_{j=1}^{J} p_{gj}\right)\right\} - \left(\sum_{g=1}^{J} 2p_{gg}\right)}{2}$$

equation A1

The next expression is true because $y + z = |y - z| + 2\min[y, z]$.

$$\frac{\sum_{g=1}^{J}\left\{\left(\sum_{i=1}^{J} p_{ig}\right) + \left(\sum_{j=1}^{J} p_{gj}\right)\right\} - \left(\sum_{g=1}^{J} 2p_{gg}\right)}{2}$$

$$= \frac{\sum_{g=1}^{J}\left\{\left|\left(\sum_{i=1}^{J} p_{ig}\right) - \left(\sum_{j=1}^{J} p_{gj}\right)\right| + 2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right),\ \left(\sum_{j=1}^{J} p_{gj}\right)\right]\right\} - \left(\sum_{g=1}^{J} 2p_{gg}\right)}{2}$$

equation A2

By the associative law of addition, we get

$$\frac{\sum_{g=1}^{J}\left\{\left|\left(\sum_{i=1}^{J} p_{ig}\right)-\left(\sum_{j=1}^{J} p_{gj}\right)\right|+2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right),\ \left(\sum_{j=1}^{J} p_{gj}\right)\right]\right\}-\left(\sum_{g=1}^{J} 2p_{gg}\right)}{2}$$

$$=\frac{\sum_{g=1}^{J}\left\{\left|\left(\sum_{i=1}^{J} p_{ig}\right)-\left(\sum_{j=1}^{J} p_{gj}\right)\right|+2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right),\ \left(\sum_{j=1}^{J} p_{gj}\right)\right]-2p_{gg}\right\}}{2}$$

$$=\frac{\sum_{g=1}^{J}\left|\left(\sum_{i=1}^{J} p_{ig}\right)-\left(\sum_{j=1}^{J} p_{gj}\right)\right|}{2}+\frac{\sum_{g=1}^{J}\left\{2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right),\ \left(\sum_{j=1}^{J} p_{gj}\right)\right]-2p_{gg}\right\}}{2}$$

$$=\frac{\sum_{g=1}^{J}\left|\left(\sum_{i=1}^{J} p_{ig}\right)-\left(\sum_{j=1}^{J} p_{gj}\right)\right|}{2}+\frac{\sum_{g=1}^{J}\left\{2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right)-p_{gg},\ \left(\sum_{j=1}^{J} p_{gj}\right)-p_{gg}\right]\right\}}{2}$$

equation A3

Finally, by equations 2-5, we get

$$\frac{\sum_{g=1}^{J}\left|\left(\sum_{i=1}^{J} p_{ig}\right)-\left(\sum_{j=1}^{J} p_{gj}\right)\right|}{2}+\frac{\sum_{g=1}^{J}\left\{2\min\left[\left(\sum_{i=1}^{J} p_{ig}\right)-p_{gg},\ \left(\sum_{j=1}^{J} p_{gj}\right)-p_{gg}\right]\right\}}{2}$$

$$=\frac{\sum_{g=1}^{J} q_g}{2}+\frac{\sum_{g=1}^{J} a_g}{2}=Q+A$$

equation A4

Literature

AICKIN, M. 1990, Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*, **46**, pp. 293-302.

ALLOUCHE, O., TSOAR, A. and KADMON, R. 2006, Assessing the accuracy of species distribution models: prevalence, Kappa and true skill statistic (TSS). *Journal of Applied Ecology*, **43**, pp. 1223-1232.

ALO, C. and PONTIUS JR, R.G. 2008, Identifying systematic land cover transitions using remote sensing and GIS: The fate of forests inside and outside protected areas of Southwestern Ghana. *Environment and Planning B*, **435**, pp. 280-295.

BRENNAN, R., and PREDIGER, D., 1981, Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, **41**, pp. 687-699.

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

CONGALTON, R.G., 1981, The use of discrete multivariate analysis for the assessment of Landsat classification accuracy. MS Thesis, Virginia Polytechnic Institute and State University, Blacksburg ,VA. pp. 111.

CONGALTON, R.G., 1991, A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, pp. 35-46.

CONGALTON, R.G., ODERWALD, R.G. and MEAD, R.A., 1983, Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, **49**, pp. 1671-1678.

CONGALTON, R.G. and GREEN, K., 1999, Assessing the accuracy of remotely sensed data: principles and practices, pp. 137 (Boca Raton, FL: Lewis Publishers)

CONGALTON, R.G. and GREEN, K., 2009, Assessing the accuracy of remotely sensed data: principles and practices, second edition, pp. 183 (Boca Raton, FL: CRC Press)

DI EUGENIO, B. and GLASS, M., 2004, The Kappa statistic: a second look, *Computational Linguistics*, **30**, pp. 95-101.

EASTMAN, J.R., 2009, *IDRISI taiga Tutorial*. Accessed in IDRISI 16.05. pp. 333 (Worcester, MA: Clark University)

ERDAS INC, 2008, *Field Guide*. Volume 2 August. (Norcross, GA: Erdas, Inc.)

FIELDING A. H. and BELL, J.F., 1997, A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation,* **24**, pp. 38-49.

FOODY, G.M., 1992, On the compensation for chance agreement in image classification accuracy assessment, *Photogrammetric Engineering and Remote Sensing*, **58**, pp. 1459-1460.

FOODY, G.M., 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, pp. 185-201.

FOODY, G.M., 2004, Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing,* **70**, pp. 627-633.

FOODY, G.M., 2008, Harshness in image classification accuracy assessment, *International Journal of Remote Sensing*, **29**, pp. 3137-3158.

GERGEL, S.E. and TURNER, M.G. (eds.), 2002, *Learning Landscape Ecology: a practical guide to concepts and techniques*. pp. 316 (New York, NY: Springer-Verlag)

GALTON, F., 1892, *Finger Prints*, pp. 216 (London: Macmillan)

GOODMAN, L.A. and KRUSKAL, W.H., 1954, Measures of Association for cross classification. *Journal of the American Statistical Association*, 49, pp. 732-764.

HAGEN, A., 2002, Multi-method assessment of map similarity. In *5th Conference on Geographic information Science*, 25-27 April 2002, Palma de Mallorca, Spain.

HUDSON, W. and RAMM, C., 1987, Correct formulation of the kappa coefficient of agreement, *Photogrammetric Engineering and Remote Sensing*, **53**, pp. 421-422.

JUNG, H-W., 2003, Evaluating interrater agreement in SPICE-based assessments, *Computer Standards and Interfaces*, **25**, pp. 477-499.

LANDIS, J. and KOCH, G., 1977, The measurement of observer agreement for categorical data, *Biometrics*, **33**, pp. 159-174.

LIU, C., FRAZIERB, P. and KUMA, L., 2007, Comparative assessment of the measures of thematic classification accuracy, *Remote Sensing of Environment*, **107**, pp. 606-616.

MA, Z. and REDMOND, R.L., 1995, Tau coefficients for accuracy assessment of classification of remote sensing data, *Photogrammetric Engineering and Remote Sensing*, **61**, pp. 435-439.

MONSERUD, R.A. and LEEMANS, R., 1992, Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*, **62**, pp. 275-293.

PONTIUS JR, R.G., 2000, Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, **66**, pp. 1011-1016.

PONTIUS JR, R.G., 2002, Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering & Remote Sensing,* **68**, pp. 1041-1049.

PONTIUS JR, R.G., AGRAWAL, A. and HUFFAKER, D., 2003, Estimating the uncertainty of land-cover extrapolations while constructing a raster map from tabular data. *Journal of Geographical Systems*, **5**, pp. 253-273.

PONTIUS JR, R.G., BOERSMA, W., CASTELLA, J.-C., CLARKE, K., DE NIJS, T., DIETZEL, C., DUAN, Z., FOTSING, E., GOLDSTEIN, N., KOK, K., KOOMEN, E., LIPPITT, C.D., McCONNELL, W., MOHD SOOD, A., PIJANOWSKI, B., PITHADIA, S., SWEENEY, S., Trung, T.N., VELDKAMP, A.T. and Verburg, P.H., 2008a, Comparing the input, output, and validation maps for several models of land change. *The Annals of Regional Science*, **42**, pp. 11-47.

PONTIUS JR, R.G., SHUSAS, E. and MCEACHERN, M., 2004, Detecting important categorical land changes while accounting for persistence. *Agriculture, Ecosystems & Environment,* **101**, pp. 251-268.

PONTIUS JR. R.G. and SPENCER, J., 2005, Uncertainty in extrapolations of predictive land change models. *Environment and Planning B: Planning and Design*, **32**, pp. 211-230.

PONTIUS JR. R.G. and SUEDMEYER, B., 2004, Components of Agreement between categorical maps at multiple resolutions. In *Remote Sensing and GIS Accuracy Assessment*, Lunetta, R.S. and Lyon, J.G. (Eds.), pp. 233-251 (Boca Raton, FL: CRC Press).

PONTIUS JR. R.G., THONTTEH, O. and CHEN, H., 2008b, Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics*, **15**, pp. 111-142.

PONTIUS JR, R.G., WALKER, R.T., YAO-KUMAH, R., ARIMA, E., ALDRICH, S., CALDAS, M., and VERGARA, D., 2007, Accuracy assessment for a simulation model of Amazonian deforestation. *Annals of the Association of American Geographers* **97**, pp. 677-695.

REMMEL, T.K., 2009, Investigating Global and Local Categorical Map Configuration Comparisons Based on Coincidence Matrices. *Geographical Analysis*, **41**, pp. 113-126.

ROSENFIELD, G. and FITZPATRICK-LINS, K., 1986, A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*, **52**, pp. 223-227.

RUELLAND, D., DEZETTER, A., PUECH, C. and ARDOIN-BARDIN, S., 2008, Long-term monitoring of land cover changes based on Landsat imagery to improve hydrological modelling in West Africa. *International Journal of Remote Sensing*, **29**, pp. 3533-3551.

SCHNEIDER, L. and PONTIUS JR, R.G., 2001, Modeling land-use change in the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment*, **85**, pp. 83-94.

SCOTT, W.A., 1955, Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, **19**, pp. 321-325.

SMITS, P.C., DELLEPIANE, S.G. and SCHOWENGERDT, R.A., 1999, Quality assessment of image classification algorithms for land-cover mapping: a review and proposal for a cost-based approach. *International Journal of Remote Sensing*, **20**, pp. 1461-1486.

STEHMAN, S.V., 1997, Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62**, pp. 77-89.

STEHMAN, S.V., 2009, Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, **30**, pp. 5243-5272.

STEHMAN, S.V. and CZAPLEWSKI, R.L., 1998, Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, **64**, pp. 331-344.

TURK, G., 2002, Map evaluation and ‘chance correction’. *Photogrammetric Engineering and Remote Sensing,* **68**, pp. 123-133.

VISSER, H, and DE NIJS, T., 2006, The map comparison kit. *Environmental Modeling & Software*, **21**, pp. 346-358.

WILKINSON, G.G., 2005, Results and implications of a study of fifteen years of satellite image classification experiments, *IEEE Transactions on Geoscience and Remote Sensing*, **43**, pp. 433-440.

WU, S., XIAOMIN, Q, USERY, E.L., and WANG, L., 2009, Using geometrical, textural, and contextual information of land parcels for classifying detailed urban land use. *Annals of the Association of American Geographers*, **99**, pp. 76-98.

WUNDRAM, D. and LÖFFER, J., 2008, High-resolution spatial analysis of mountain landscapes using a low-altitude remote sensing approach. *International Journal of Remote Sensing*, **29**, pp. 961-974.

Tables

Table 1. Format for observed sample matrix.

		Reference					
		j=1	j=2	…	j=J	Sample Total	Population Total
Comparison	i=1	n_{11}	n_{12}		n_{1J}	$\sum_{j=1}^{J} n_{1j}$	N_1
	i=2	n_{21}	n_{22}		n_{2J}	$\sum_{j=1}^{J} n_{2j}$	N_2
	…						
	i=J	n_{J1}	n_{J2}		n_{JJ}	$\sum_{j=1}^{J} n_{Jj}$	N_J

Table 2. Format for estimated population matrix.

		Reference				
		j=1	j=2	…	j=J	Comparison Total
Comparison	i=1	p_{11}	p_{12}		p_{1J}	$\sum_{j=1}^{J} p_{1j}$
	i=2	p_{21}	p_{22}		p_{2J}	$\sum_{j=1}^{J} p_{2j}$
	…					
	i=J	p_{J1}	p_{J2}		p_{JJ}	$\sum_{j=1}^{J} p_{Jj}$
	Reference Total	$\sum_{i=1}^{J} p_{i1}$	$\sum_{i=1}^{J} p_{i2}$		$\sum_{i=1}^{J} p_{iJ}$	1

Table 3. Parameter descriptions in terms of the disagreement space in figures 2-9.

Parameter	Description
Q	lower bound
A	total disagreement minus quantity disagreement
κ_{no}	(one half minus total disagreement) / one half
κ_{standard}	(random line minus total disagreement) / random line
$\kappa_{\text{allocation}}$	(random line minus total disagreement) / (random line minus quantity disagreement)
κ_{histo}	(random line minus quantity disagreement) / random line
κ_{quantity}	See Pontius (2000)

Figures

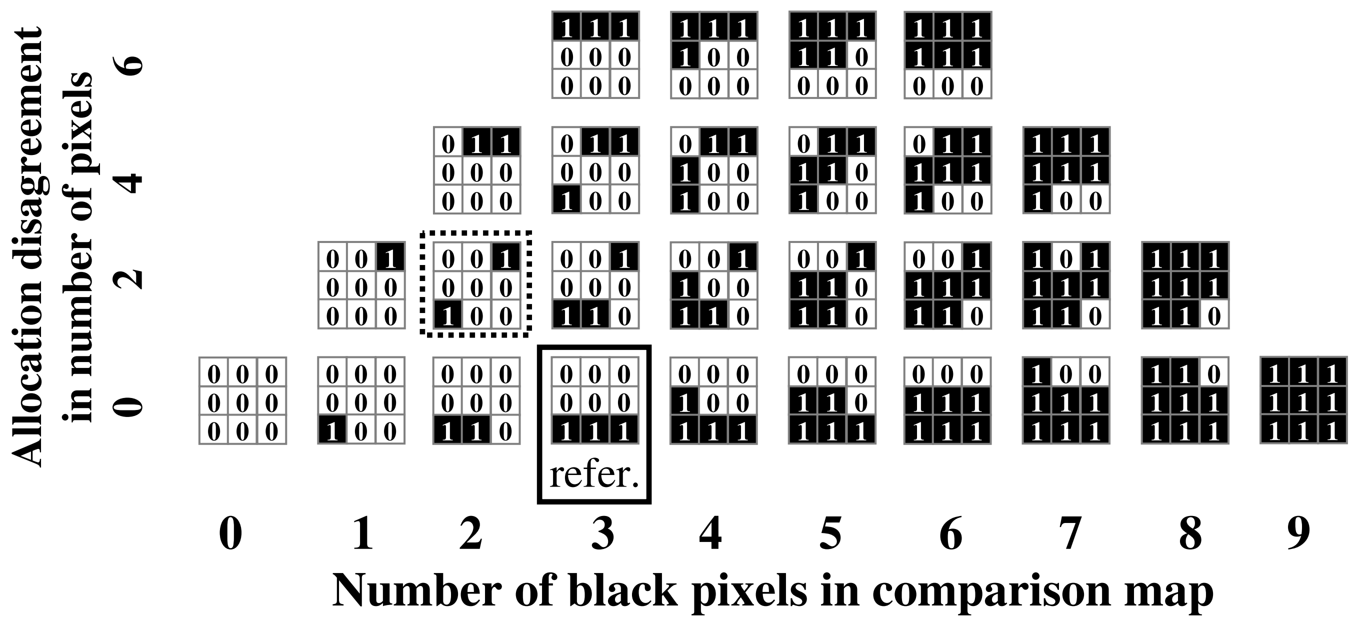

Figure 1. Reference (refer.) map and comparison maps that show all possible combinations of quantity disagreement and allocation disagreement. The dotted box highlights a comparison map that has three pixels of disagreement, where the two pixels of disagreement at the bottom are omission disagreement for the black category and the one pixel in the upper right is comission disagreement for the black category. This implies that the comparison map in the dotted box has one pixel of quantity disgreement and two pixels of allocation disagreement, since two pixels in the comparison map could be reallocated in a manner that would increase agreement with the reference map.

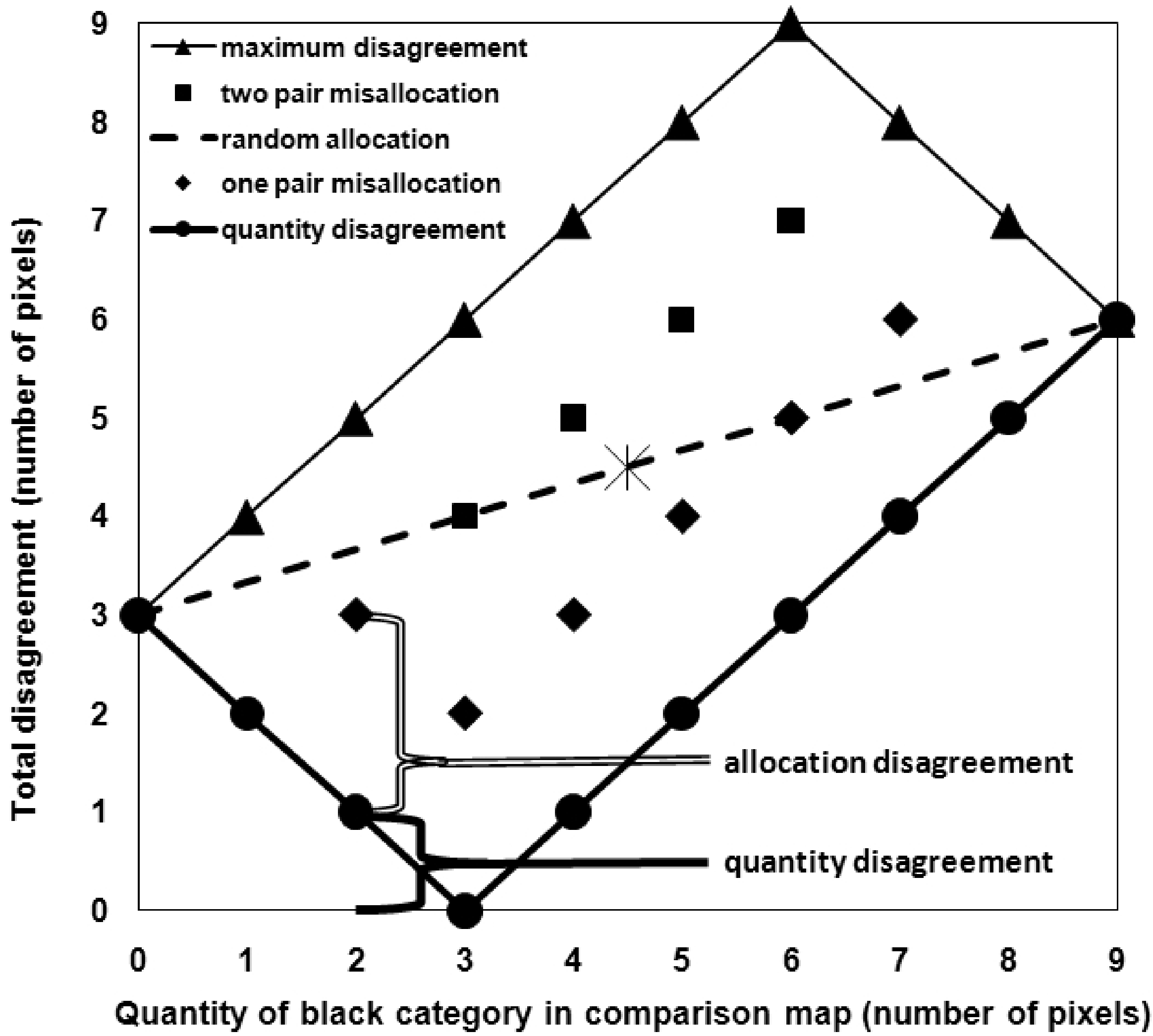

Figure 2. Disagreement space for all comparison maps, showing quantity disagreement and allocation disagreement for the highlighted comparison map in figure 1.

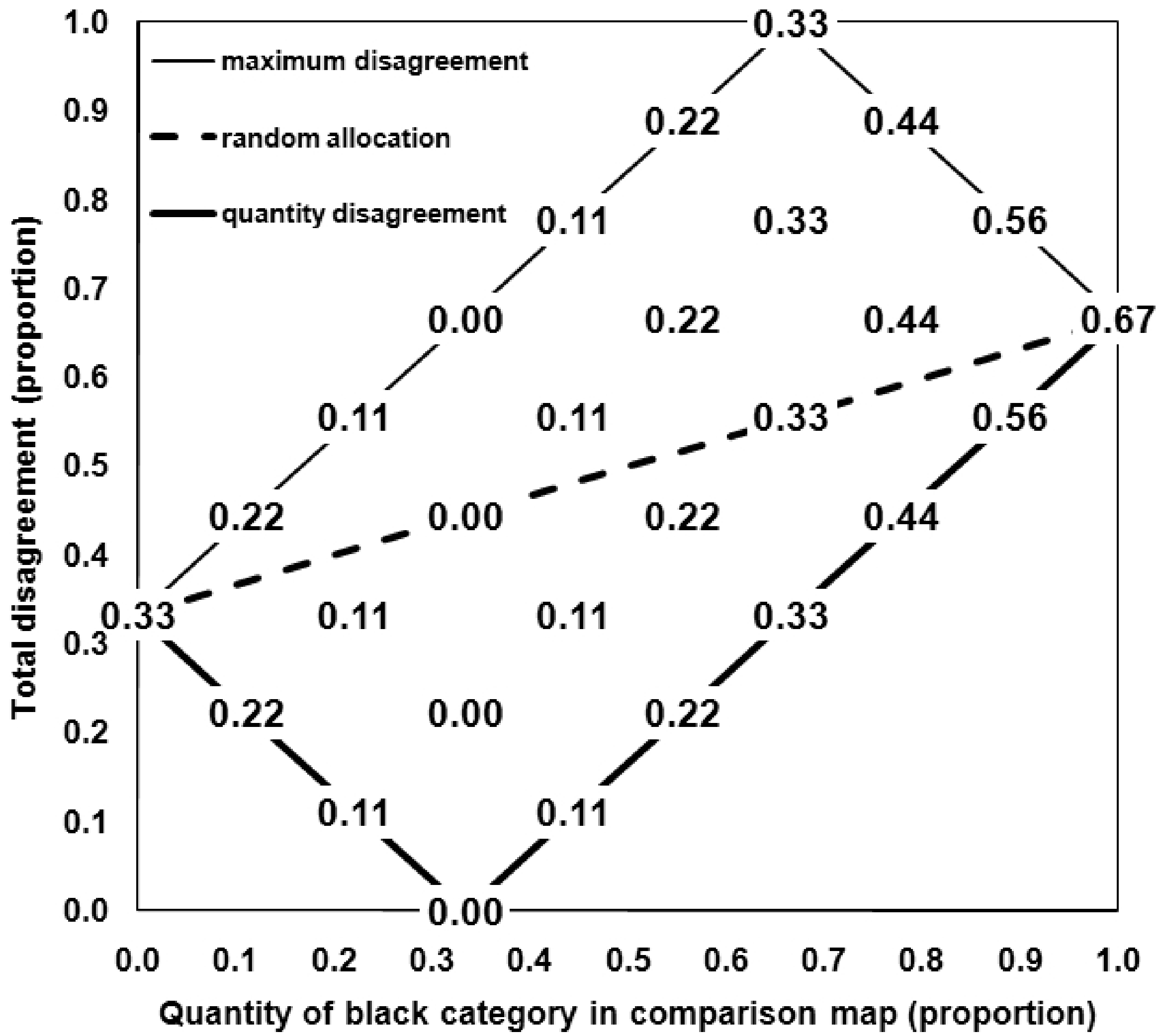

Figure 3. Quantity disagreement *Q* shown by the values plotted in the space.

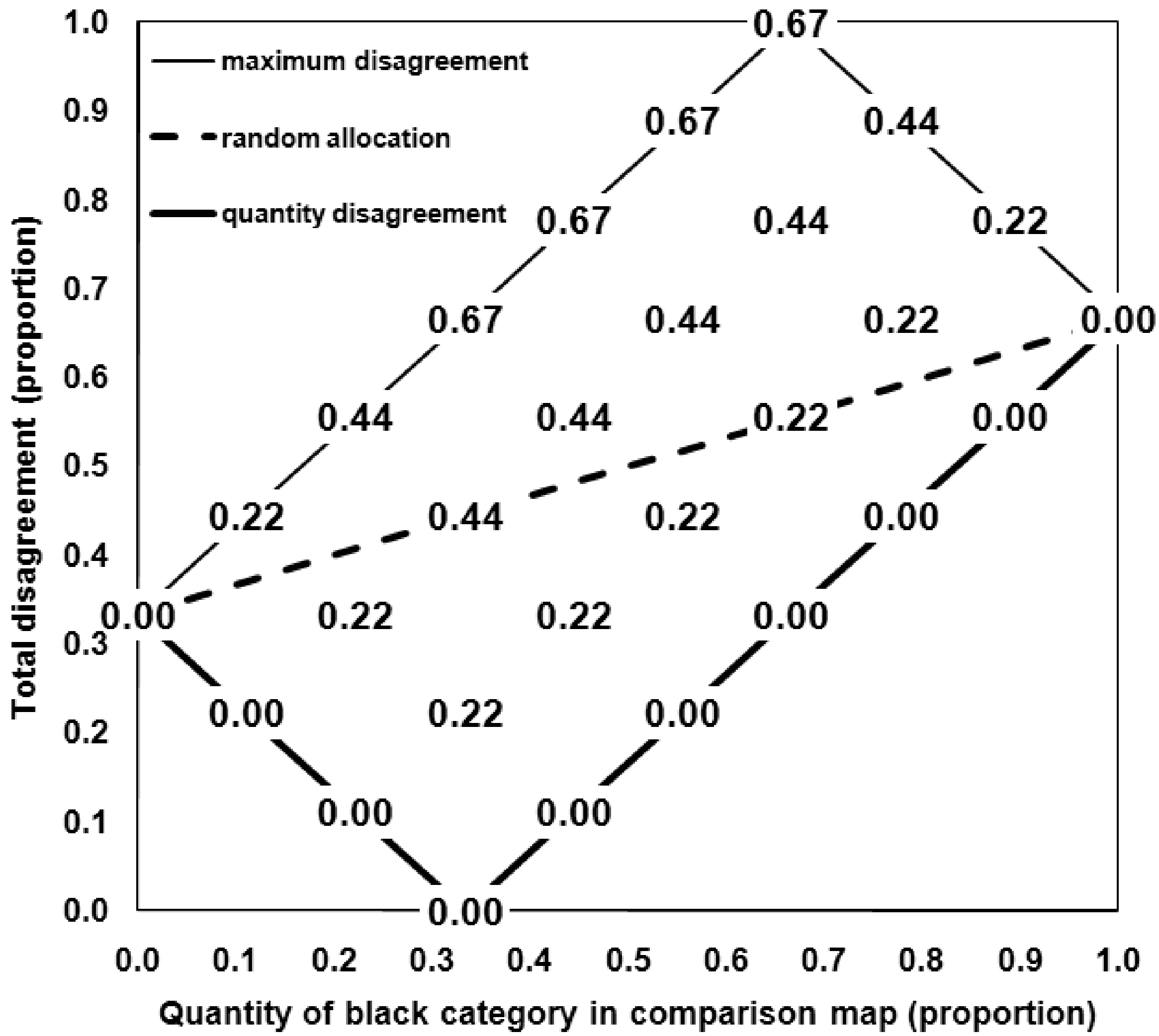

Figure 4. Allocation disagreement *A* shown by the values plotted in the space.

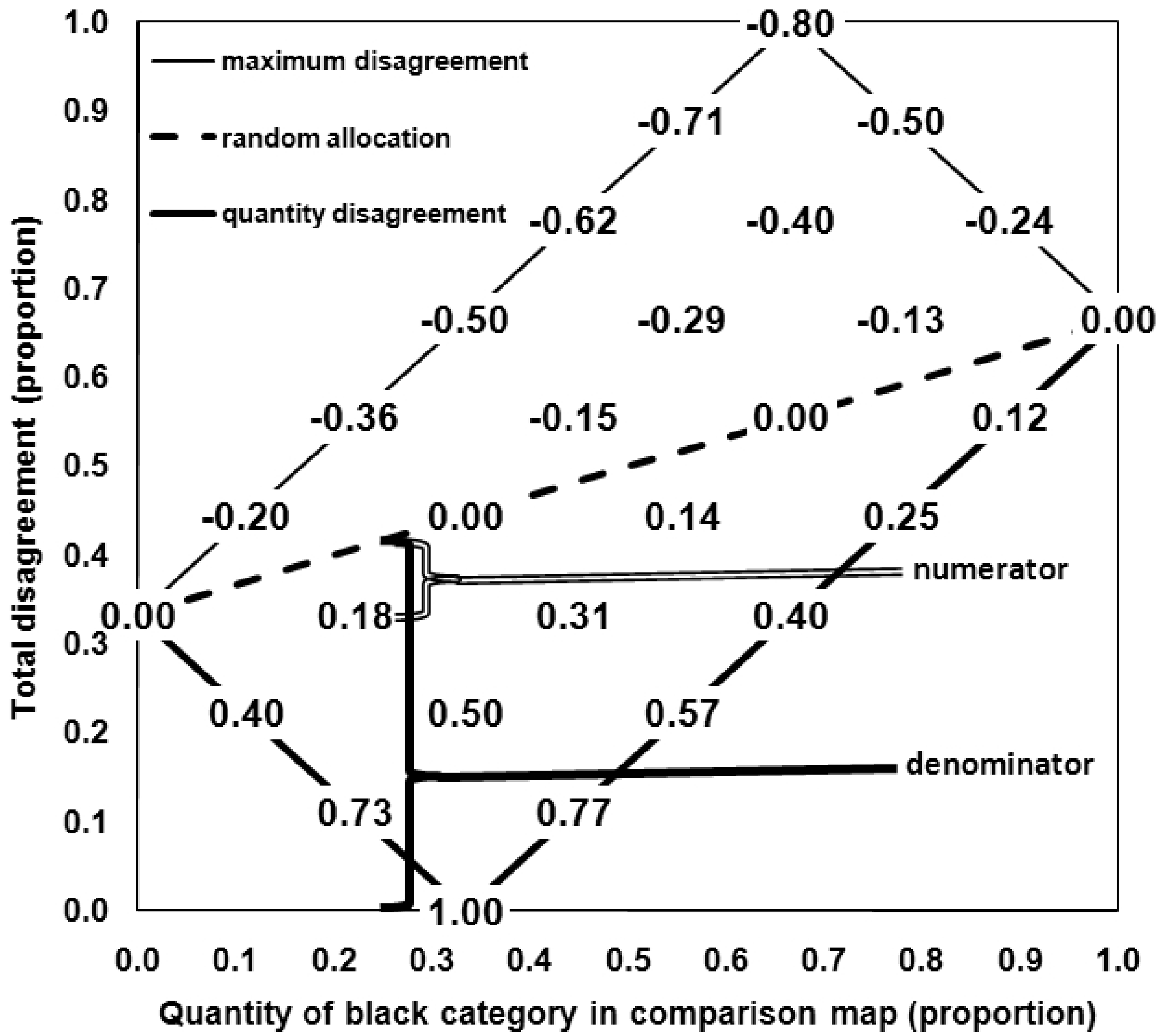

Figure 5. Standard Kappa κ_{standard} shown by the values plotted in the space, where the braces show the numerator and denominator for the highlighted comparison map in figure 1.

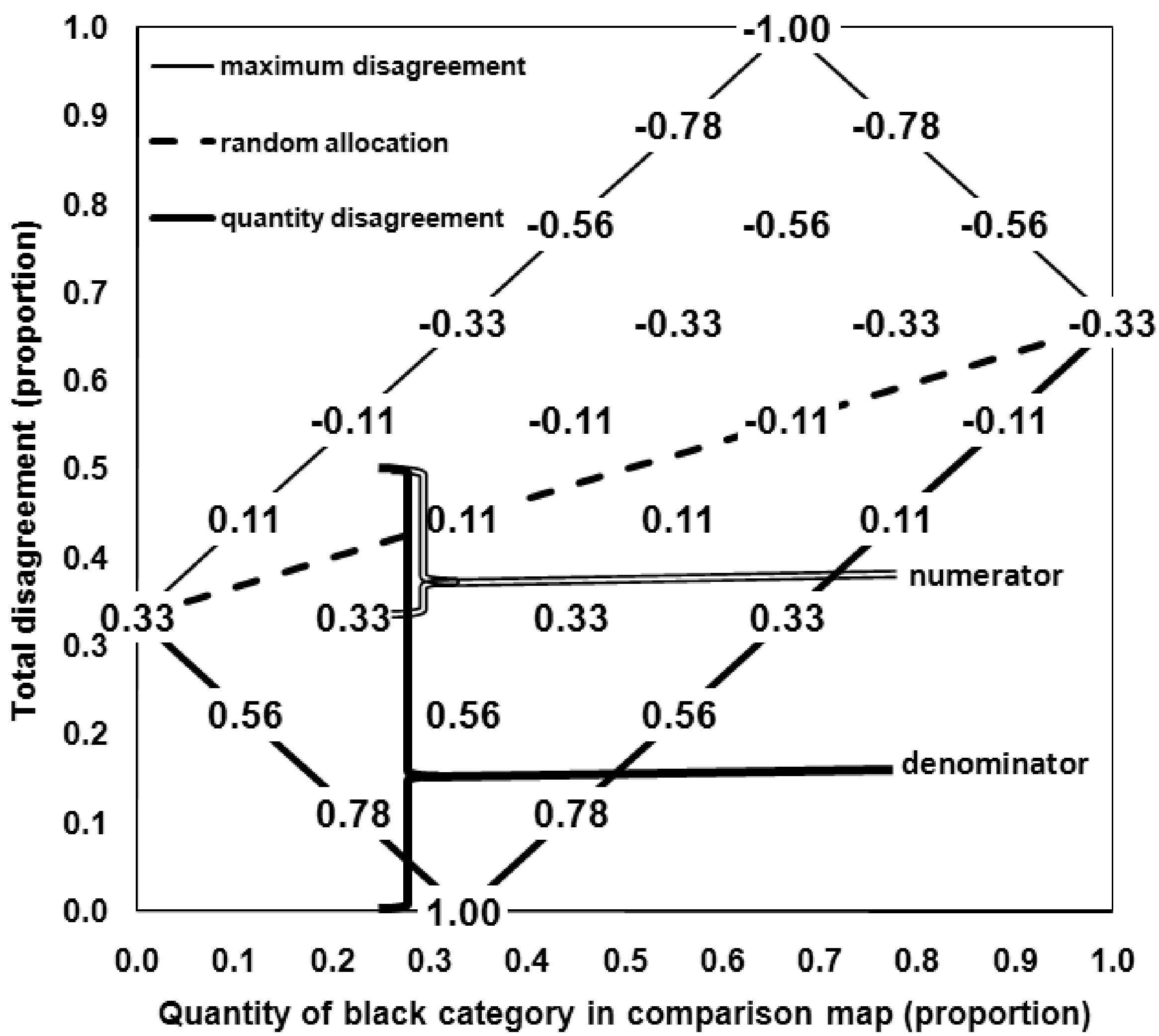

Figure 6. Kappa for no information κ_{no} shown by the values plotted in the space, where the braces show the numerator and denominator for the highlighted comparison map in figure 1.

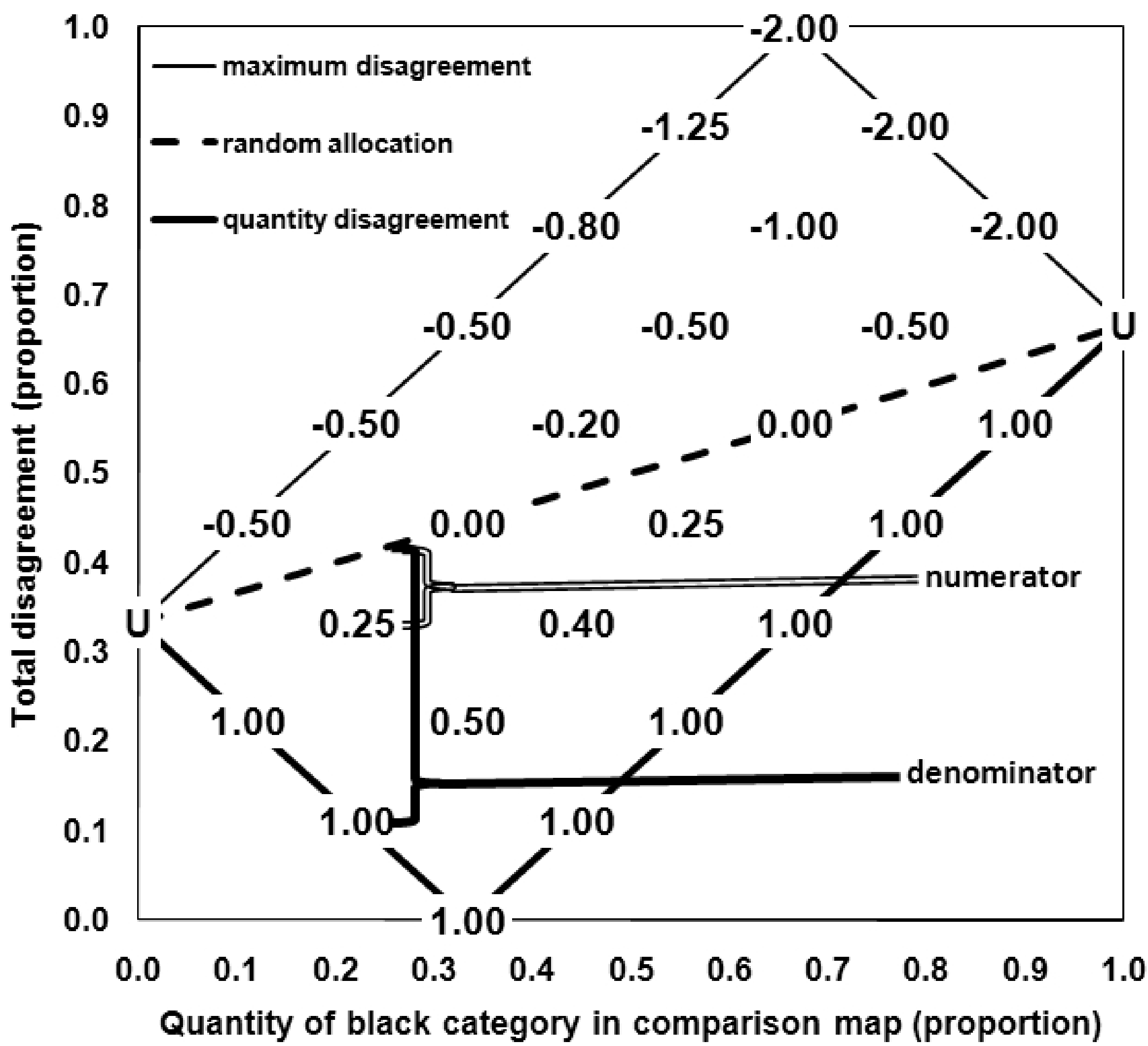

Figure 7. Kappa for allocation $\kappa_{\text{allocation}}$ shown by the values plotted in the space, where the braces show the numerator and denominator for the highlighted comparison map in figure 1. U means undefined.

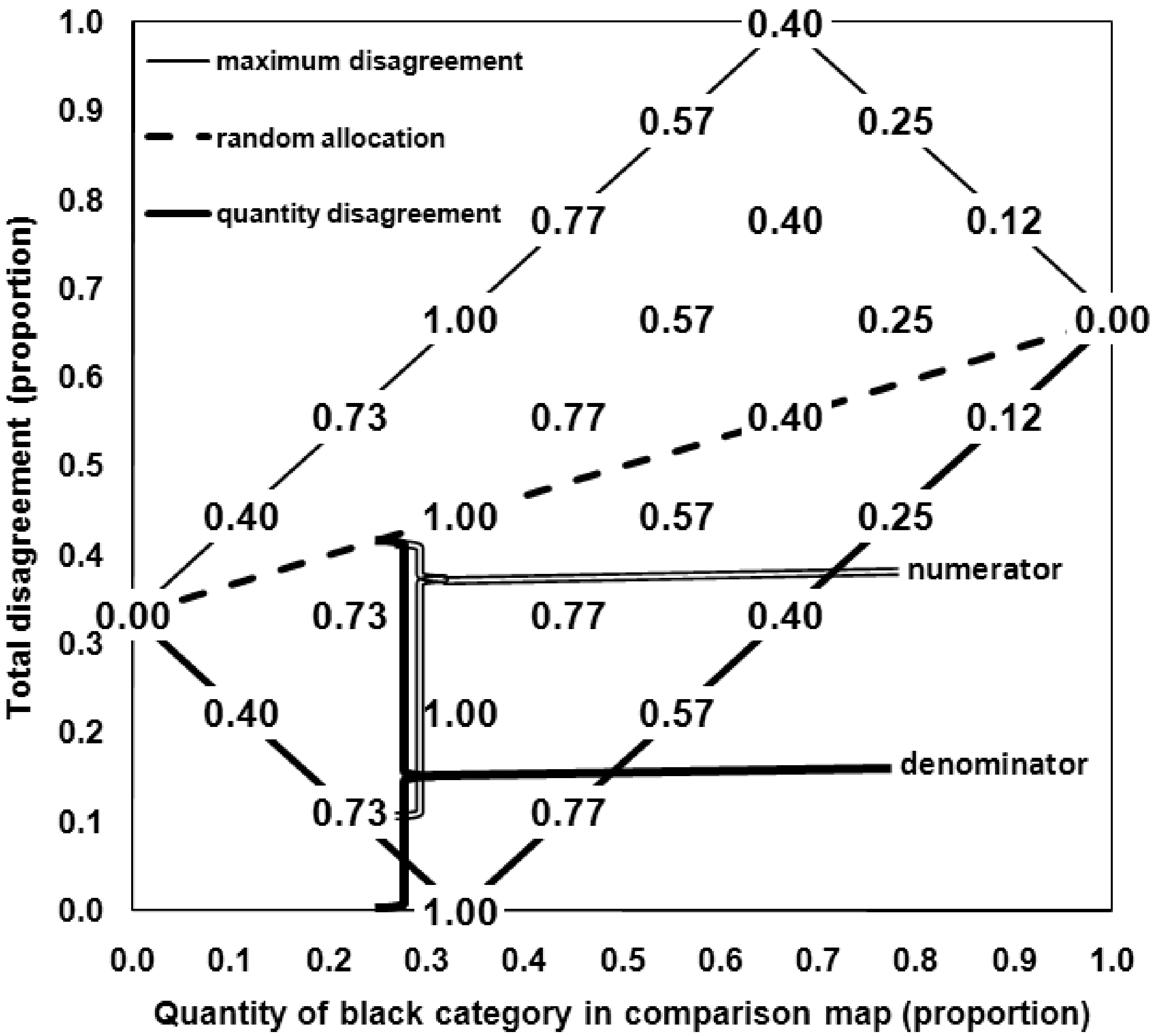

Figure 8. Kappa for histogram κ_{histo} shown by the values plotted in the space, where the braces show the numerator and denominator for the highlighted comparison map in figure 1.

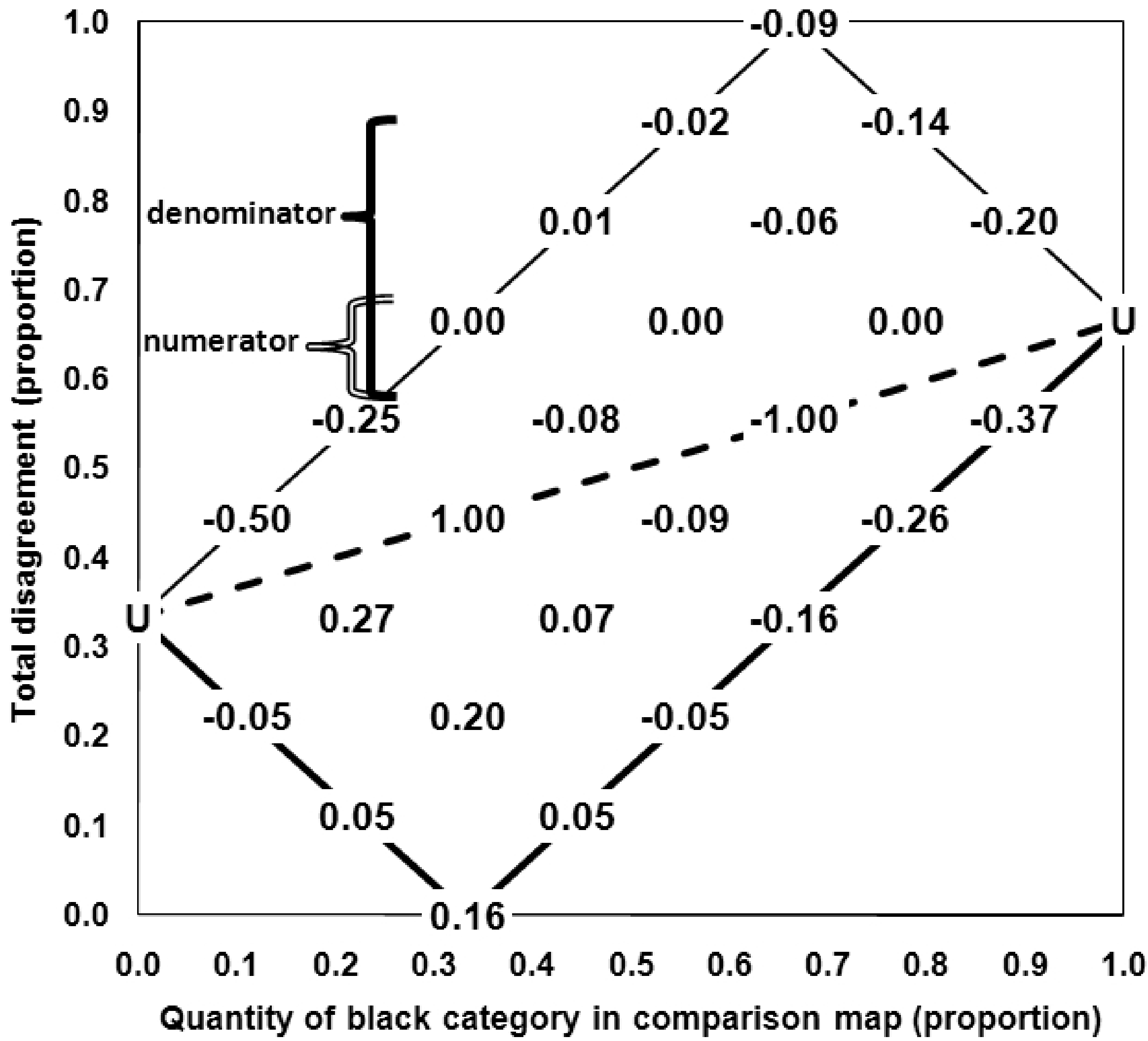

Figure 9. Kappa for quantity $\kappa_{quantity}$ shown by the values plotted in the space, where the braces show the numerator and denominator for the highlighted comparison map in figure 1. U means undefined.

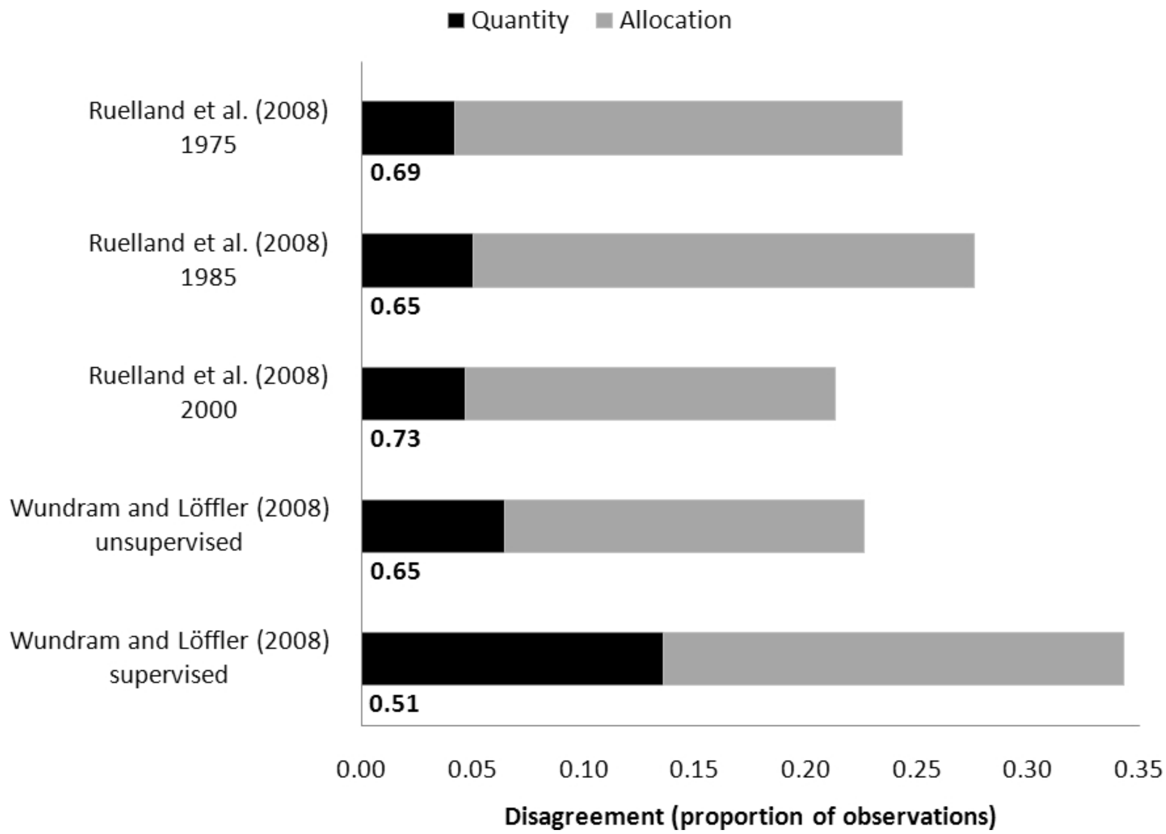

Figure 10. Quantity disagreement, allocation disagreement, and $\kappa_{standard}$ below each bar for five matrices published in International Journal of Remote Sensing.