PROBLEMS AND SOLUTIONS FOR KAPPA-BASED INDICES OF AGREEMENT

Robert Gilmore Pontius Jr[1] and Marco Millones[2]

[1]Clark University, Department of International Development, Community and Environment
[2]Clark University, Graduate School of Geography

Abstract

The kappa index of agreement (i.e. Kstandard) is one of the most commonly used parameters to measure the correspondence between two maps that share a categorical variable, such as land cover type. It is typically applied in situations to quantify map correspondence for accuracy assessment, model validation, and temporal change. Previous work by the first author of this paper and others show that Kstandard has severe conceptual problems that cause it to give extremely misleading information for some types of applications. Pontius (2000, 2002) explain some of these problems and propose a suite of additional indices that attempt to correct those problems. His proposed indices were Kappa for location (Klocation), Kappa for Quantity (Kquantity), and Kappa for no information (Kno). These three indices are now starting to be adopted by other scientists, who interpret them in a manner that their names imply. However, this conference paper shows that Klocation is not necessarily the best indication of agreement due to spatial allocation and Kquantity is not a good indication of agreement due to quantity. This paper exposes the problems with the entire suite of kappa-related parameters and offers a philosophy of map comparison that gives a simpler and more appropriate method to compare maps that share a categorical variable. Ultimately, we recommend that kappa-based indices be abandoned for common applications, and that the newly proposed approach be adopted for map comparisons.

Keywords

accuracy, agreement, confusion, error, change, matrix.

1 Introduction

The popular kappa index of agreement, denoted as Kstandard, is one of the most common summary statistics used to compare maps that share a categorical variable. Kstandard can be traced to Galton (1892) but it is usually attributed to Cohen (1960). We suspect its popularity for map comparison is linked to the fact that some software packages focus almost exclusively on Kstandard as a measure of accuracy assessment and Kstandard has been one of the few summary statistics highlighted in and recommended by frequently cited literature (Congalton et al. 1983, Monserud and Leemans 1992, Congalton and Green 1999, Smits et al. 1999, Wilkinson 2005). In spite of this, many researchers have been quite critical of Kstandard for various reasons (Foody 1992, 2002, in press, Ma and Redmond 1995, Fielding and Bell 1997, Stehman 1997, Stehman and Czaplewski 1998, Turk 2002). A main criticism from Pontius (2000) is that Kstandard is a one-dimensional index that fails to specify reasons for the disagreement between two maps, because Kstandard confounds information about the quantity of each category in the maps with information about the location of each category in the maps. His criticism is related to the concern about bias as a function of variation in prevalence of the categories (DiEugenio and Glass 2004, Allouche et al. 2006). To attempt to remedy this problem, Pontius (2000) proposed a suite of additional kappa based indices: Kquantity as an index of agreement associated with the quantity of each category in the maps, Klocation as an index of agreement associated with the location of each category in the maps, and Kno as an index of overall agreement between the maps relative to completely random agreement. Nearly simultaneously, Hagen (2002) derived Khisto, which is based exclusively on the association between the quantities of the categories in the maps, and showed how Kstandard can be separated into two factors: Khisto and Klocation. The scientific community paid attention, and these indices are starting to appear in the literature.

Now that we have had more experience with these various kappa-based indices in practice, we advise against using any of them for nearly all cases that we have seen. Upon reflection on additional years of experience, we have learned that there are more direct ways to express the important information that these indices attempt to describe. Therefore, we write this paper to answer the questions: What are the conceptual problems with these kappa-based indices of agreement, and what are better alternatives for map comparison? This paper's Methods section describes the concepts by dissecting a case study of that compares several maps that have four pixels and two-categories. The Results section exposes how these indices can be misleading or otherwise flawed. The Discussion section explains why Kstandard, Kno, Klocation, Kquantity, and Khisto are not useful for many common applications and how there are better alternatives. The paper concludes that it is frequently more helpful to compare two maps that share a categorical variable by reporting simply the quantity disagreement and location disagreement.

2 Methods

2.1 Eight comparison maps for a four-pixel two-category case study

We have designed figure 1 to expresses the most important concepts in the simplest possible terms. Figure 1 shows the reference map on the left and eight comparison maps to its right, each consisting of four pixels, where each pixel is labeled as exactly one category, 0 or 1. The comparison maps are named A, B, … , H in order from left to right and top to bottom.

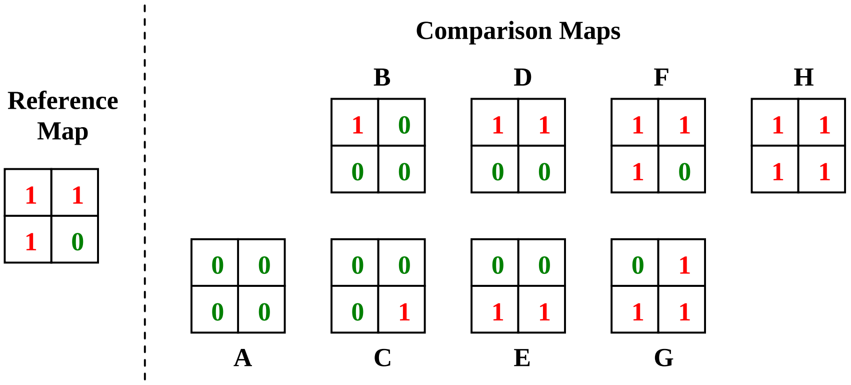

Figure 1. Reference map and eight comparison maps for the case study.

Figure 2 plots the observed agreement between the reference map and each comparison map versus the quantity of pixels of category 1 in the comparison map. The observed agreement is the percent of pixels that correspond when the reference map is compared to the comparison map. Each comparison map falls along the horizontal axis at 0%, 25%, 50%, 75%, or 100%, according to its quantity of 1s. For maps A and F, all the pixels are a single category, so there are not various possibilities for distinct spatial allocations of the pixels. At each of the other middle three levels of quantity along the horizontal axis, there are two comparison maps: the upper map which shows a spatial allocation of pixels that maximizes its agreement with the reference map, and the lower map which shows a spatial allocation of pixels that minimizes its agreement with the reference map. The diagonally-oriented rectangle defined by the thick solid lines in figure 2 shows the mathematically possible region in the space. The orientation of the rectangle is dictated by the composition of the reference map. The observed agreement between the reference map and any comparison map must be on or below the upper bound and on or above the lower bound. For each comparison map, the horizontal coordinate of its point in the space is determined by its quantity of 1s, while its vertical coordinate is determined by the spatial allocation of its 1s, therefore the orthogonal axes show conceptually independent components of information.

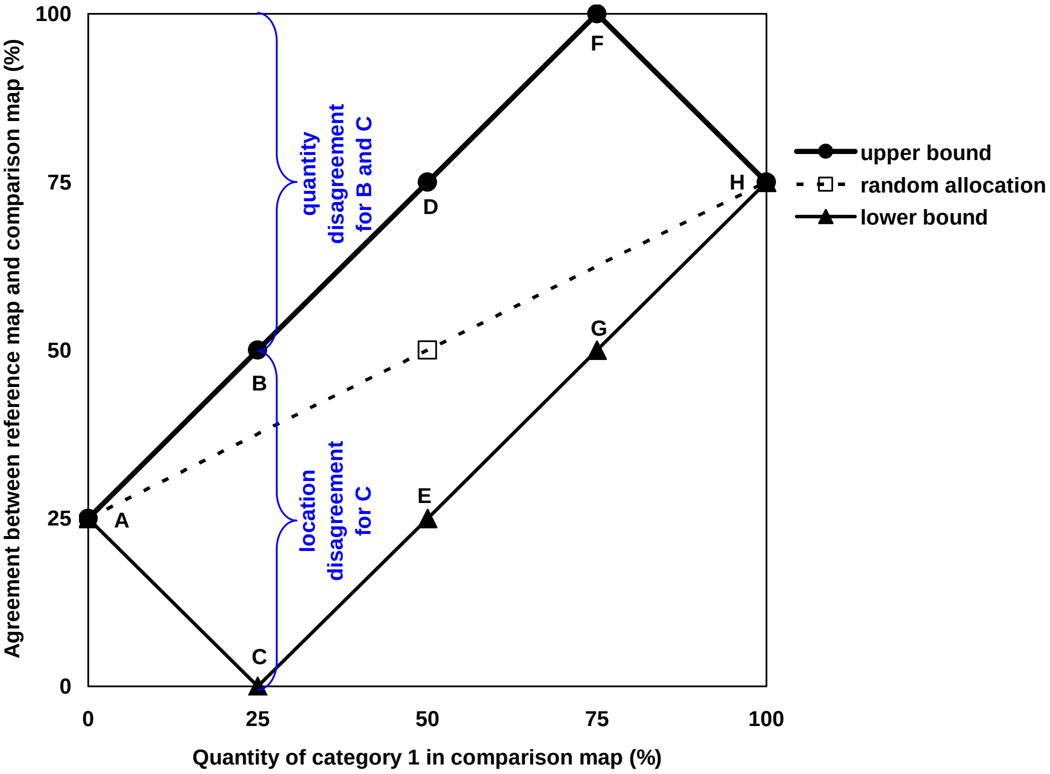

Figure 2. Computation of quantity disagreement and location disagrerement plotted in the space defined by the axis for percent agreement versus the axis for percent of category 1 for eight comparison maps.

The vertical coordinate for each comparison map gives its observed agreement with the reference map, so 100% minus its agreement is its overall disagreement. It is possible to separate the overall disagreement into two components. First, the quantity disagreement is the vertical distance between 100% agreement and the upper bound of the diagonally-oriented rectangle. It measures how much less than perfect the correspondence is between the reference map and the comparison map in terms of the quantity of each category in the maps. Second, the location disagreement is the vertical distance between that same upper bound of the rectangle and the observed agreement between the reference map and the comparison map. It measures how much less than optimal the correspondence is between the reference map and the comparison map in terms of the spatial allocation of each category in the maps, given the observed quantities of each category in the maps. Figure 2 illustrates how to compute the components for maps B and C. The observed agreement is 50% for map B and 0% for map C. The observed agreement plus the two components of disagreement always sum to 100 percent. The quantity disagreement is 50% for both maps B and C. The location disagreement is 0% for map B, while it is 50% for map C. Ultimately, this paper will conclude that it is usually most helpful to specify the two components of disagreement and to not proceed with more complex analysis that the kappa-indicators of agreement require.

2.2 Kappa-based indices

The kappa-based indices of agreement require additional elaboration on figure 2, because the kappa indices of agreement are designed to consider the agreement that would be expected if the pixels in the comparison map were to be rearranged according to random spatial allocation. The dashed line that bisects the rectangle in figure 2 gives this expected agreement due to random spatial allocation, for any fixed quantity along the horizontal axis.

The various kappa-based indices measure where the observed agreement lies with respect to the dashed and solid lines in figure 2. All of the indices are designed to yield 100% when the observed agreement between the reference map and the comparison map is perfect, and to yield 0% when the observed agreement is equal to the expected agreement due to some type of randomization in the comparison map. Kstandard indicates the observed agreement relative to both 100% and the random allocation line. Klocation indicates the observed agreement relative to both the upper bound and the random allocation line. Kno indicates the observed agreement relative to both 100% and the expected agreement for a map in which both the quantities and locations of the categories are randomized. The point (50%, 50%) in figure 2 shows this expected agreement under complete randomization for the case where there are two categories. In general, the coordinates are 100/*J* percent under complete randomization, where *J* is the number of categories. Khisto indicates the distance between the upper bound and the random allocation line relative to the distance between 100% and the random allocation line, at the place along the horizontal axis dictated by the comparison map. Kquantity attempts to indicate how well the quantities of the categories in the comparison map are specified compared to both perfect and random specification of the quantities, assuming that the observed Klocation is fixed. Kquantity is more complicated than the other indices because it attempts to address a more complicated question, which is motivated by an application to land change modeling. Pontius (2000) describes its derivation, which is more complex than figure 2 can show easily. Table 1 defines the indices in a manner that relates the concepts to figure 2.

Table 1. Conceptual description of indices with respect to figure 2.

PARAMETER	DESCRIPTION
Quantity Disagreement	100% minus upper bound
Location Disagreement	upper bound minus observed agreement
Kno for two categories	$\frac{\text{observed agreement minus 50\%}}{\text{100\% minus 50\%}}$
Kstandard	$\frac{\text{observed agreement minus random line}}{\text{100\% minus random line}}$
Klocation	$\frac{\text{observed agreement minus random line}}{\text{upper bound minus random line}}$
Khisto	$\frac{\text{upper bound minus random line}}{\text{100\% minus random line}}$
Kquantity	See Pontius (2000)

3 Results

Table 2 gives the results when the reference map is compared to each of the eight comparison maps in figure 1. For all of the cases, overall agreement, quantity disagreement, and location disagreement sum to 100% by design and have straight forward interpretations as the methods section describes. Table 2 reveals counter-intuitive characteristics or other types of flaws for each of the kappa-based indices of agreement. Specifically, Kno is identical for comparison maps A & E, but maps A & E have substantially different quantities of category 1. Kno produces a similar situation for map pairs B & G and D & H. Kstandard is positive for B and negative for G, in spite of the fact that the observed agreement the same for B and G. This is related to the fact that B has substantial quantity disagreement and G has no quantity disagreement. Klocation is undefined for A and H, which is the case wherever the upper bound meets the random line. Klocation is much smaller for case C than it is for case G, in spite of the fact that both cases show a spatial allocation that minimizes agreement, given the specified quantities of the categories. Klocation is substantially different for cases C, E and G, in spite of the fact that location disagreement is the same for C, E and G. Khisto is zero for both cases A and H, while the quantity disagreement is different for cases A and H, since case A has none of category 1 and H consists entirely of category 1. Kquantity is undefined for three of the eight cases. Kquantity is zero for case D and undefined for case E, in spite of the fact that the quantity is the same for cases D and E. For case C, quantity disagreement is equal to location disagreement, but Khisto is positive while Kquantity is negative, and both are greater than Klocation.

Table 2. Statistics expressed as percents for eight comparison maps in the case study. U means "undefined" due to zero in the denominator.

PARAMETER	COMPARISON MAP							
	A	B	C	D	E	F	G	H
Quantity Disagreement	75	50	50	25	25	0	0	25
Location Disagreement	0	0	50	0	50	0	50	0
Overall Agreement	25	50	0	75	25	100	50	75
Kno	-50	0	-100	50	-50	100	0	50
Kstandard	0	20	-60	50	-50	100	-33	0
Klocation	U	100	-300	100	-100	100	-33	U
Khisto	0	20	20	50	50	100	100	0
Kquantity	U	-100	-100	0	U	100	100	U

4 Discussion

Our major objection to all of the kappa-based indices of agreement is that they are distracting because they address questions that are frequently irrelevant for practical purposes and they do so in an unnecessarily complicated manner. The indices are designed to compare the observed agreement between a reference map and a comparison map vis-à-vis the expected agreement between the reference map and an altered comparison map that is generated by randomly rearranging the pixels of the original comparison map, while keeping constant the quantity of the pixels in each category. Most literature that we have seen attempts to explain this concept in an abbreviated manner. Consequently, it is typical that such literature claims that Kstandard gives the overall agreement corrected for chance agreement, which seems initially like it might be useful. However, this typical explanation is misleading because it is only partially true. A more complete description would be that Kstandard gives the overall agreement relative to the expected agreement under randomized spatial allocation, given the fixed quantity of each category in the maps. Consequently, Kstandard fails to penalize for large quantity disagreement and fails to reward for small quantity disagreement. If the literature were to report clearly a proper meaning of Kstandard, then readers would probably realize that Kstandard usually does not give information that is particularly useful for applied purposes.

Comparison to random spatial allocation is not particularly interesting because we usually already know that the patterns in the maps are not random, but we still want to know how they are different. The question “Is the agreement between two maps better than random?” is fundamentally different than the question “How are two maps different?” Kappa indices address the former question, while the components of disagreement address the latter question. Nevertheless some scientists are tempted to focus on comparison to randomness probably because comparison to randomness is a consistent theme in many university-level statistics courses. Randomness commonly serves as a baseline due to historical convention, but we have seen few cases in which randomness is a particularly relevant baseline, and we have seen no cases in which randomness is the single most important baseline. A helpful baseline is usually the agreement between the reference map and an alternative comparison map that is generated through some simpler method (Pontius et al. 2007). For example, if one is studying the accuracy of a map generated by a newly proposed classification algorithm, then an interesting baseline would be the accuracy of a map generated by an older well-established algorithm. In the case of predictive land change modeling through time, an appropriate baseline would be a prediction of complete persistence through time (Pontius et al. 2008). In general, an appropriate baseline map is the comparison map that a scientist would have generated with some existing conventional technique, and in no cases that we have seen is the conventional technique a method that assigns pixels randomly.

It is difficult for us to imagine many situations where it would be important to consider kappa-based indices. We ask readers to send us applications for which they think any of the kappa indices of agreement have been applied in a useful manner. Pontius et al. (2003) is the only case that we have seen where a kappa index seems to have been used appropriately for its intended purpose. In that case, Klocation was used, but Kstandard would have served just as well. In most of the other applications of kappa-based indices that we have seen, the statistics have been misinterpreted or simpler measurements would have been more useful. For example, Schneider and Pontius (2001) used Kquantity for its apparent intended purpose and consequently had the awkward challenge to interpret values greater than 100%, whereas it would have been clearer and more helpful simply to report and to interpret the components of disagreement.

Most scientists who have asked for our advice concerning map comparison want a clear answer to the direct question, “What are the differences between the reference map and the comparison map?” If this is the question, then it is frequently most helpful to examine how the reference map compares directly to the comparison map in terms of the quantities of the categories and the spatial allocation of the categories, which is what quantity disagreement and location disagreement measure. These two components of disagreement answer the question clearly, while the kappa indices do not, since the kappa indices rely on a third map that is generated through randomization. Even if it were somewhat interesting to use a kappa-based index for a particular application, proper interpretation of the kappa-based indices usually requires knowledge of the components of disagreement defined in the methods section and additional components of agreement defined in Pontius (2002) and Pontius et al. (2007). The components of disagreement show how the two maps disagree and the components of agreement show how much better the observed agreement is than the agreement that could be attributable to random chance. It is common for maps to show substantial disagreement while simultaneously showing agreement that is much larger than could be attributable to random chance, as illustrated by the Costa Rica case study in Pontius (2002).

While the components of disagreement and agreement can give helpful information directly, most kappa indices are ratios of those components. Consequently, kappa indices can suffer from the small denominator syndrome. The small denominator syndrome is the situation where the numerator is not meaningfully different from zero for practical applications, but the resulting ratio is large because the denominator is smaller than the numerator. This causes a serious problem in interpretation when the most relevant information is simply the size of the numerator. Worse yet, when the denominator is zero, the ratio is undefined regardless of the magnitude of the numerator, which can occur for cases that are important to analyze (Table 2).

5 Conclusions

We recommend that scientists use kappa-based indices of agreement only when the specific index answers a specific question for which the index has been designed. We have found very few cases that benefit from the insights offered from kappa-based measurements. We have found numerous cases where the indices offer misleading information. We recommend that scientists avoid kappa-based indices for purposes of general map comparison because: 1) they are designed to consider agreement relative to random agreement which is frequently not an important consideration, 2) they are very easily misinterpreted since Klocation does not necessarily indicate the match in terms of location and neither Kquantity nor Khisto necessarily indicate the match in terms of quantity, 3) they can vary wildly when the denominator is small, 4) they are undefined for potentially important cases, 5) there are simpler and more direct ways to express important components of information for general map comparison. We recommend that scientists report quantity disagreement and location disagreement for general map comparison because: 1) they reveal facts that are usually fundamentally important since they measure directly the differences between the relevant maps, 2) they are much simpler mathematically and conceptually than the kappa indices, and 3) it is necessary to know them for proper interpretation of the kappa indices for cases where kappa-based indices might be helpful.

Acknowledgements

We thank the bachelors, masters, and doctoral students in GIS classes at Clark University for helping us to think through these ideas and to communicate them in the simplest possible manner. Clark Labs facilitated this work by creating the GIS software Idrisi®, which computes the useful measurements that this paper discusses.

Literature

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.

Cohen, J. (1960) A coeffient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37-46.

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 K. Green. (1999) *Assessing the accuracy of remotely sensed data: principles and practices*, Lewis, Boca Raton FL.

Di Eugenio, B. and Glass, M. (2004) The kappa statistic: a second look, *Computational Linguistics*, 30, pp. 95-101.

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(1) p. 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(1), pp. 185-201.

Foody, G.M. (in press) Harshness in image classification accuracy assessment, *International Journal of Remote Sensing*.

Galton, F. (1892) *Finger Prints*, Macmillan, London.

Hagen, A. (2002) Multi-method assessment of map similarity. Presented at the 5th Conference on Geographic information Science, Palma (Mallorca, Spain) April 25th-27th.

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 and Remote Sensing* 66(8), 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 and Remote Sensing*, 68(10), 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(3), pp. 253-273.

Pontius Jr, R.G., Walker, R., 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(4), pp. 677-695.

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. (2008) Comparing the input, output, and validation maps for several models of land change. *Annals of Regional Science*, 42(1) pp.11-47.

Schneider, L. and Pontius Jr, R.G. (2001). Modeling land-use change in the Ipswich watershed, Massachusetts, USA, *Agriculture, Ecosystems & Environment*, 85(1-3), pp. 83-94.

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. 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.

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.