

On Using and Computing the Kappa Statistic

Emmett lentilucci, 1-17-06

In this short report, we present the math, followed by an example, on how to use and interpret the Kappa Statistic. **The example error matrices come from (Congalton et. al. 1983).** The results and equations have been cross referenced with those from the journal paper. Additional information can be found in (Congalton and Green, 1999).

Congalton, R.G., Oderwald, R.G., Mead, R.A., "Assessing Landsat Classification Accuracy Using Discrete Multivariate Analysis Statistical Techniques", PERS, Vol. 49(12), pp. 1671-1678, 1983.

Congalton, R.G., Green, K., Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Lweis Publishers, 1999.

The Error Matrix

The example error matrices are the result of applying a non-supervised 10-cluster (method A) and 20-cluster (method B) classifier on LANDSAT data of the Ludwig Mountain area in Colorado. The four classes to be classified were conifer (C), deciduous (D), agriculture (A), and water (W). The following error matrices resulted.

$$\begin{array}{cccc}
 & \text{C} & \text{D} & \text{A} & \text{W} \\
 \text{A} := & \begin{bmatrix} 317 & 23 & 0 & 0 \\ 61 & 120 & 0 & 0 \\ 2 & 4 & 60 & 0 \\ 35 & 29 & 0 & 8 \end{bmatrix} & \begin{array}{l} \text{C} \\ \text{D} \\ \text{A} \\ \text{W} \end{array} & &
 \end{array}
 \qquad
 \begin{array}{cccc}
 & \text{C} & \text{D} & \text{A} & \text{W} \\
 \text{B} := & \begin{bmatrix} 377 & 79 & 0 & 0 \\ 2 & 72 & 0 & 0 \\ 33 & 5 & 60 & 0 \\ 3 & 20 & 0 & 8 \end{bmatrix} & \begin{array}{l} \text{C} \\ \text{D} \\ \text{A} \\ \text{W} \end{array} & &
 \end{array}$$

The columns represent the reference data while the rows represent that generated from classifying the remotely sensed data.

We now write some functions to compute the column and row totals that will be used in subsequent calculations. Therefore we have,

$$\text{CT}(\text{data}) := \begin{array}{l} \text{for } \text{col} \in 0.. \text{cols}(\text{data}) - 1 \\ \text{sumcol}_{\text{col}} \leftarrow \sum(\text{data})^{<\text{col}>} \\ \text{sumcol}^T \end{array}
 \qquad
 \text{RT}(\text{data}) := \begin{array}{l} \text{for } \text{row} \in 0.. \text{rows}(\text{data}) - 1 \\ \text{sumrow}_{\text{row}} \leftarrow \sum(\text{data}^T)^{<\text{row}>} \\ \text{sumrow}^T \end{array}$$

When these functions are applied to our example we have

<u>Method A</u>	<u>Method B</u>	
RT(A) = (340 181 66 72)	RT(B) = (456 74 98 31)	Row Totals
CT(A) = (415 176 60 8)	CT(B) = (415 176 60 8)	Column Totals

Since the classifiers were performed on the same data set, we expect the total number of classified pixels to be the same for error matrices A and B, though this is not a requirement for the analysis. From the row and column totals, we can compute the total number of pixels classified by simply summing up the marginals. That is

$$N_A := \sum RT(A) \quad N_B := \sum RT(B) \quad k := \text{rows}(A)$$

$$N_A = 659 \quad N_B = 659 \quad k = 4$$

A common measure of classification accuracy is to calculate the **overall accuracy** of the classifier. This simply measures the number of correctly classified units by summing up the diagonal and dividing by the total number of pixels. For our example we have

<u>Method A</u>	<u>Method B</u>	
$p_{oA} := \frac{\text{tr}(A)}{N_A}$	$p_{oB} := \frac{\text{tr}(B)}{N_B}$	
$p_{oA} = 0.7663$	$p_{oB} = 0.7845$	Overall Accuracy

Which says we have classified about $p_{oA} \cdot 100 = 77\%$ of the pixels correctly using method A and $p_{oB} \cdot 100 = 78\%$ of the pixels correctly using method B. A perfect classification would produce a value of one or 100%.

We can also calculate the proportion of pixels that were classified due to "chance agreement". That is,

Method A

$$p_{cA} := \frac{1}{N_A^2} \cdot \sum_{m=0}^{k-1} RT(A)^{\langle m \rangle} \cdot CT(A)^{\langle m \rangle}$$

$p_{cA} = 0.4087$

Method B

$$p_{cB} := \frac{1}{N_B^2} \cdot \sum_{m=0}^{k-1} RT(A)^{\langle m \rangle} \cdot CT(A)^{\langle m \rangle}$$

$p_{cB} = 0.47986$

This says that about $p_{cA} \cdot 100 = 41$ % of the pixels were classified due to random chance.

The Kappa Statistic

We now define the "Kappa statistic" which is based on the difference between the actual agreement in the error matrix and the chance agreement, which is indicated by the row and column totals. This would be similar to computing a percent error or difference. Therefore, a maximum likelihood estimate of Kappa is given by

Method A

$$k_{\text{hat}}_A := \frac{P_{oA} - P_{cA}}{1 - P_{cA}}$$

$k_{\text{hat}}_A = 0.605$

Method B

$$k_{\text{hat}}_B := \frac{P_{oB} - P_{cB}}{1 - P_{cB}}$$

$k_{\text{hat}}_B = 0.586$

Kappa values range from 0 to 1, though they can be negative and range from -1 to 1. However, since there should be a positive correlation between the remotely sensed classification and the reference data, positive Kappa values are expected. A perfect classification would produce a Kappa value of one.

Typically, values greater than 0.80 (i.e, 80%) represent strong agreement between the remotely sensed classification and the reference data while values between 0.4 and 0.8 represent moderate agreement. Anything below 0.4 is indicative of poor agreement.

We can additionally compute the **variance** and **standard deviation** related to the Kappa statistic as follows

$$a_{1A} := \sum \left[\frac{1}{N_A^2} \cdot \sum_{m=0}^{k-1} A_{m,m} \cdot (RT(A)^{\langle m \rangle} + CT(A)^{\langle m \rangle}) \right] \quad a_{1A} = 0.6686339$$

$$a_{1B} := \sum \left[\frac{1}{N_B^2} \cdot \sum_{m=0}^{k-1} B_{m,m} \cdot (RT(B)^{\langle m \rangle} + CT(B)^{\langle m \rangle}) \right] \quad a_{1B} = 0.8201119$$

$$a_{2A} := \sum \left[\frac{1}{N_A^3} \cdot \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} A_{i,j} \cdot (RT(A)^{\langle i \rangle} + CT(A)^{\langle j \rangle})^2 \right] \quad a_{2A} = 0.8231181$$

$$a_{2B} := \sum \left[\frac{1}{N_B^3} \cdot \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} B_{i,j} \cdot (RT(B)^{\langle i \rangle} + CT(B)^{\langle j \rangle})^2 \right] \quad a_{2B} = 1.1690231$$

$$\text{var_k}_A := \frac{1}{N_A} \cdot \left[\frac{p_{oA} \cdot (1 - p_{oA})}{(1 - p_{cA})^2} + \frac{2 \cdot (1 - p_{oA}) \cdot (2 \cdot p_{oA} \cdot p_{cA} - a_{1A})}{(1 - p_{cA})^3} + \frac{(1 - p_{oA})^2 \cdot (a_{2A} - 4 \cdot p_{cA}^2)}{(1 - p_{cA})^4} \right]$$

$$\text{var_k}_B := \frac{1}{N_B} \cdot \left[\frac{p_{oB} \cdot (1 - p_{oB})}{(1 - p_{cB})^2} + \frac{2 \cdot (1 - p_{oB}) \cdot (2 \cdot p_{oB} \cdot p_{cB} - a_{1B})}{(1 - p_{cB})^3} + \frac{(1 - p_{oB})^2 \cdot (a_{2B} - 4 \cdot p_{cB}^2)}{(1 - p_{cB})^4} \right]$$

Method A

$$\text{var_k}_A = 0.00073735$$

$$\text{stddev_k}_A := \sqrt{\text{var_k}_A}$$

Method B

$$\text{var_k}_B = 0.00087457$$

$$\text{stddev_k}_B := \sqrt{\text{var_k}_B}$$

A perfect classification would produce a variance and standard deviation of zero.

Significance Testing and Confidence Levels

We can also test to see if the classification which produced the Kappa (k_{hat}) statistic, is significantly better than a random result (*i.e.*, $k_{\text{hat}} = 0$ or no difference between p_c and p_o). That is we are testing

Ho: $k_{\text{hat}}=0$	<u>Not</u> better than random
H1 $k_{\text{hat}}\neq 0$	Better than random

The test statistic for testing the significance of a single error matrix is expressed by

<u>Method A</u>	<u>Method B</u>
$z_A := \frac{k_{\text{hat}}_A - 0}{\text{stddev}_k_A}$	$z_B := \frac{k_{\text{hat}}_B - 0}{\text{stddev}_k_B}$
$z_A = 22.272$	$z_B = 19.806$

where z is standardized (*i.e.*, z-score) and normally distributed (*i.e.*, standard normal deviate). The z-score (assuming normality) gives us a nice measure of how far, and in what direction, the item deviates from its distribution's mean, expressed in standard deviation units. We can then reject the null hypothesis (that the classifier producing k_{hat} is NOT better than a random result) if

$$z \geq Z_{\frac{\alpha}{2}} \quad \text{Criteria for rejecting null hypothesis}$$

where $Z_{\alpha/2}$ produces a standard normal critical value, z_c via the two-tailed Z-test. For small samples we would select from the Student's t distribution (*i.e.*, $N < 30$). Note: the Z- and t- distributions are about equal when $N > 200$).

If we choose a significance level of α (*i.e.*, a confidence level of $1 - \alpha$) then we must use a probability value, $p = \alpha/2$ because of the two-tail nature of the test. That is

$\alpha := 0.05$	CL := $1 - \alpha$	
CL = 0.95		
	$z_c := \left \text{qnorm}\left(\frac{\alpha}{2}, 0, 1\right) \right $	The Z-distribution is $N(0,1)$
	$z_c = 1.96$	

With a $CL \cdot 100 = 95\%$, we obtain a standard normal critical value $z_c = 1.96$, which is the number of standard deviation from the mean of the Z-distribution. Therefore, if the absolute value of the z-score is greater than z_c , the result is significant and we would conclude that the classification is better than random.

For methods A and B the results are:

Method A: z-score is $z_A = 22.272$ and $z_c = 1.96$

Method B: z-score is $z_B = 19.806$ and $z_c = 1.96$

Since the z-scores for BOTH methods are significantly greater than z_c , we can reject the NULL and say with 95% confidence that these classification methods are better than random.

Confidence Intervals

For large samples, the Kappa statistic is asymptotically normally distributed. Not only can we perform significance testing but this also allows us to compute confidence intervals (CI) around the Kappa value. With a $CL \cdot 100 = 95\%$, the confidence intervals are calculated as

Method A

$$\text{upper_k_hat}_A := k_hat_A + (z_c \cdot \text{stddev_k}_A)$$

$$\text{upper_k_hat}_A = 0.658$$

$$k_hat_A = 0.605$$

$$\text{lower_k_hat}_A := k_hat_A - (z_c \cdot \text{stddev_k}_A)$$

$$\text{lower_k_hat}_A = 0.552$$

Method B

$$\text{upper_k_hat}_B := k_hat_B + (z_c \cdot \text{stddev_k}_B)$$

$$\text{upper_k_hat}_B = 0.644$$

$$k_hat_B = 0.586$$

$$\text{lower_k_hat}_B := k_hat_B - (z_c \cdot \text{stddev_k}_B)$$

$$\text{lower_k_hat}_B = 0.528$$

Comparing Two Error Matrices

We can additionally compare two Kappa values generated from two independent error matrices to see if they are significantly different. This is useful when you are interested in whether different models, methodologies or interpreters produce significantly different results.

The hypothesis is now of the form

Ho: $(k_{\text{hat}}_A - k_{\text{hat}}_B) = 0$ Method A and Method B are not different

H1: $(k_{\text{hat}}_A - k_{\text{hat}}_B) \neq 0$ Method A and Method B are different

The test statistic for testing whether or not the two classification methods are different can be expressed as

$$z_{AB} := \frac{|k_{\text{hat}}_A - k_{\text{hat}}_B|}{\sqrt{\text{var}_k_A + \text{var}_k_B}}$$

$$z_{AB} = 0.475$$

where z_{AB} is standardized (*i.e.*, z-score) and normally distributed (*i.e.*, standard normal deviate) like before. We can then reject the null hypothesis (that method A is different than method B) if

$$z_{AB} \geq Z_{\frac{\alpha}{2}} \quad \text{Criteria for rejecting null hypothesis}$$

where $Z_{\alpha/2}$ produces a standard normal critical value, z_c via the two-tailed Z-test, as previously mentioned. Using the same criteria as previously defined we have,

$$\alpha = 0.05$$

$$CL = 0.95$$

Therefore, $z_{AB} = 0.475$ is NOT greater than $z_c = 1.96$ at a confidence level of $CL = 0.95$

We accept the null. There is no difference between methods A and B. This result indicates that there is no justification for spending the extra time to use the 20-cluster approach because the 10-cluster approach works just as well.