# COMPLETE CONTROL OF AN OBSERVED CONFUSION MATRIX

[a]*Ariza-López F.J.*, [b]*Rodríguez-Avi J.*, [b]*Alba-Fernández M.V.*

[a]Universidad de Jaén, Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría
[b]Universidad de Jaén, Departamento de Estadística e Investigación Operativa

## ABSTRACT

The error matrix has been adopted as a standard way to report on the thematic accuracy of any remotely sensed data product. A very usual way to perform the thematic accuracy analysis of an error matrix is by means of global indexes (e.g. overall accuracy, Kappa coefficient). But global indices do not allow for a category-wise control. This work proposes a new method for accuracy control of thematic classification based on an application of the chi-square goodness of fit test. By this way we can stablish our preferences of accuracy for each category but also we can consider some degree of misclassification between some categories. A practical example is provided for a 4x4 matrix (16 class combinations). In this example, a total of 13 quality levels (specifications) are imposed for certain class combinations. In this way, much more is controlled than by means of global indexes. In addition, the method allows to be more demanding or flexible in the quality levels of certain classes, according to convenience.

***Index Terms***— Thematic accuracy, quality control, confusion matrix, statistical test

## 1. INTRODUCTION

A confusion matrix is a common tool for assessing thematic accuracy of many remote sensed derived products (e.g. land cover classifications) and, in general, of spatial data [8]. As stated by Comber et al. [4], the confusion matrix has been adopted as both the "*de facto*" and the "*de jure*" standard, the way to report on the thematic accuracy. In this way, the International Standard 19157 [8] includes the confusion matrix as measure for thematic quality assessment. A confusion matrix (error matrix or contingency table) [1], is a statistical tool for the analysis of paired observations. The content of a confusion matrix is a set of values accounting for the degree of similarity between paired observations between a controlled data set (CDS), a product, and a reference data set (RDS) (e.g. ground truth). Thus, a confusion matrix is created from a control sampling on the product (CDS) and a response design (protocol for determining a reference) [13]. So, it is a M×M squared matrix where M denotes the number of classes under consideration. A confusion matrix gives us a complete vision of the allocation (Ec.1) of errors, where the diagonal elements of a confusion matrix contain the correctly classified items, and the off diagonal elements contain the number of confusions, the errors due to omissions or commissions [10].

$$CM(i,j) = [\#\text{items of class (j) of the RDS classified as class (i) of the CDS}] \quad (1)$$

For example, Table 1 presents the confusion matrix for a proof developed by the Construction Engineering Research Lab (U.S. ARMY) [12].

Table 1. Example of confusion matrix (source: [12])

| | | Reference data | | | |
|---|---|---|---|---|---|
| | | Wo | G | N | Wa |
| Data classification | Wo | 47 | 3 | 0 | 0 |
| | G | 4 | 40 | 6 | 0 |
| | N | 0 | 5 | 45 | 0 |
| | Wa | 0 | 0 | 2 | 48 |
| W=Woodland, G=Grassland, N=Non-vegetated, Wa=Water | | | | | |

When controlling differences between remote sensed derived products (e.g. between two independent works, between products derived from two different sources, dates or methods, etc.), quantitative comparisons can be achieved using statistical hypothesis testing. However, a complete confusion matrix is difficult to manage in a simple way. For this reason, different indices (accuracy parameters) are derived to summarize this information by means of a value, or a reduced set of values ([5], [13]). Two widely adopted indices are the overall accuracy (OA, see Ec.2) and the Kappa coefficient ($K$, see Ec.3). The OA is simple to compute, easy to understand and helpful to interpret [11]. Congalton and Green [5] state that 'Kappa analysis has become a standard component of most every accuracy assessment and is considered a required component of most image analysis software packages that include accuracy assessment procedures'.

Both global indexes can be used in statistical z-tests. For instance, Ec.4 presents the z-test statistic for testing an observed value of OA versus a threshold or hypothesized value $OA_T$ ($H_0$: $OA=OA_T$). The same Z-test statistics can be applied for testing the same hypothesis using the Kappa

coefficient. See [5] for more details on these basic analysis techniques.

$$OA = \frac{1}{n}\sum_{i=1}^{M} n_{ii} = \sum_{i=1}^{M} p_{ii} \quad (2)$$

$$K = \frac{OA - Ps}{1 - Ps} \quad (3)$$

Where:

$$Ps = \frac{1}{N^2}\sum_{i=1}^{M} n_{i+} n_{+i}$$

$$Z = \frac{OA - OA_T}{\sqrt{OA(1-OA)/N}} \quad (4)$$

Both indices (OA and Kappa), are widely adopted, but exists some criticism about their use in relation to the over and under estimations induced by them [15], [9], and several authors have revealed some concerns about the use of these former measures for global thematic accuracy assessments [9], [7], specially for the Kappa coefficient [11].

But an accuracy control performed as described above is very general, the accuracy of the overall classification can be good enough, but particular categories, and cases of the confusion matrix, can be no so well classified. This is not a category-wise control and minority categories can be masked by more majority categories.

The goal of this work is to propose the use of a statistical hypothesis test in such way that allows us the statement of our preferences of quality levels for each cell of the confusion matrix.

## 2. STATISTICAL APPROACH

Table 2 shows a confusion matrix, that can be seen as $k \times k$ matrix, where $k$ is the number of classes, or categories, under consideration. We consider that columns show the true value (or reference) $(A_1, \dots, A_k)$, obtained under a method of greater accuracy (e.g. ground truth), and rows are the observed value $(a_1, \dots, a_k)$ which correspond to the product under control.

Table 2. Confusion matrix

| | True | | | | |
|---|---|---|---|---|---|
| Observed | $A_1$ | | $A_j$ | | $A_k$ |
| $a_1$ | $x_{11}$ | | $x_{1j}$ | | $x_{1k}$ |
| | | | | | |
| $a_i$ | $x_{i1}$ | | $x_{ij}$ | | $x_{ik}$ |
| | | | | | |
| $a_k$ | $x_{k1}$ | | $x_{kj}$ | | $x_{kk}$ |
| Totals | $T_{.1}$ | | $T_{.j}$ | | $T_{.k}$ |

where:

- $x_{jj}$: Counting of elements where Category $a_j$ is observed when the correct category is $A_j$: Correct.
- $x_{ij}, i \neq j$: Counting of elements where Category $a_i$ is observed when the correct category is $A_j$: Error.
- $T_{.j} = \sum x_{ij}$: Sum of elements that truly belongs to category $A_j$

Our goal is to decide if elements of a confusion matrix are distributed according to a (previously) specified model. In the case of quality controls this model is established by the quality level specifications for each category and mix of categories. For instance, let consider the 4x4 confusion matrix presented by Table 1, and now let consider our preferences about quality levels for each cell of the matrix: diagonal cells (pure categories) and out-of-diagonal cells (mix between categories). Because not all errors (mistakes or confusions between classes) are equally important and have the same probability of occurrence [14], [2], we can fix a quality value (proportion in each category, denoted by $\pi^0_{ij}$) for each cell inside the confusion matrix, determining the minimum (for diagonal cells) or maximum (for out-of-diagonal cells) quality (proportion) allowed under the condition that the sum of proportions by column sum 100%. As indicated, these proportional values have different meanings in the elements of the diagonal and in the elements outside the diagonal. For the elements of the diagonal, the minimum allowed is indicated, since we would like the real value to be higher (increasing direction). For the elements outside the diagonal, the maximum allowed is indicated, since we would like the real value to be lower (decreasing direction). The matrix of values established in such way is the matrix of assumed minimum values for quality. This will be the null hypothesis for the statistical test we will performed. In this way, Table 3 shows some arbitrary conditions we have established for this example. These conditions actuate as thematic accuracy quality level specifications for the product under control. The example shows three different cases, i) columns labeled G and N present a complete situation where all M alternatives are possible, ii) in the column labeled with Wo two classes have been aggregated or collapsed (N + Wa), iii) column labeled with Wa, where a binomial situation is presented (Water versus all the others). As example, the specifications for the case of Woodland category (first column of Table 3) can be understood in this way:

- At least 95% of classification accuracy ($\geq 95\%$) is required for the woodland category.
- Woodland category can be somewhat confused with grassland, but not more than 4% ($\leq 4\%$).
- Woodland category cannot be confused with non-vegetated or with water ($\leq 1\%$).

Table 3. Example of assumed minimum values for quality for the example of Table 1

| | | Reference data | | | |
|---|---|---|---|---|---|
| | | Wo | G | N | Wa |
| Data classification | Wo | 0.95 | 0.04 | 0.01 | 0.01 |
| | G | 0.04 | 0.85 | 0.08 | |
| | N | | 0.10 | 0.90 | |
| | Wa | 0.01 | 0.01 | 0.01 | 0.99 |
| | Total | 1.00 | 1.00 | 1.00 | 1.00 |
| W=Woodland, G=Grassland, N=Non-vegetated, W=Water | | | | | |

Following the above mentioned considerations, we can apply $k$ goodness-of-fit tests based on the chi-square distribution [3]. For this, we calculate the standardized residual for each cell in respect with the expected value according to the corresponding null hypothesis in the column.

$$R_{ij} = \frac{x_{ij} - \Pi_{ij}^{0} T_{.j}}{\sqrt{\Pi_{ij}^{0} T_{.j}}} \qquad (5)$$

Each residual is calculated as the difference between the truly observed and the expected value under the null hypothesis. In consequence, a positive sign shows that we have observed more elements than we expected under our specifications. This play an important role in order to interpret the sense of the differences, but only when the null hypothesis is rejected. In this sense, positive values out the diagonal say that the real distribution is worse than the expected one, whereas positive values in the diagonal, and negative out of the diagonal shows that the true distribution is better than the expected one.

For the column $j$ a goodness-of-fit test is calculated by Ec.6.

$$\chi_j = \sum R_{ij}^2 = \sum_{i=1}^{k} \frac{(x_{ij} - \pi_{ij}^{0} T_{.j})^2}{\pi_{ij}^{0} T_{.j}} \qquad (6)$$

For each column, this statistic is distributed according to a $\chi^2$ distribution with *(k-1)* degrees of freedom, due that we do not have to estimate any parameter and we calculate a p-value for each column. In consequence, we have $k$ statistical contrasts, one for each column, and we can take an overall decision applying the Bonferroni criterion [6].

This is a bilateral test. In consequence, if we reject the null hypothesis, we must decide if the observed case is worse or better than the null hypothesis. Therefore, in this case we must examine the resulting value $R_{ij}$ for each column separately, in order to determine why the null hypothesis is rejected. The sign of $R_{ij}$ indicates the rejection sense: negative in the diagonal element and/or positive for the elements outside the diagonal. These signs will indicate that the matrix is better/worse than the hypothesis established for each column.

This test allows the possibility that not all the categories need to be specified. In several circumstances, some error categories may be collapsed. This collapse of cells in the matrix means that the existence of confusion between the categories that are grouped is accepted. The product quality specifications will establish the maximum level of occurrence (%) that is accepted in each confusion. Table 3 is an example of several cases (e.g. N and Wa for the Woodland category). This test is also applicable in this case, with the only difference in the degree of freedom. So if $k'<k$ categories are determined, the degrees of freedom for this column has to be *(k'-1)*.

## 3. EXAMPLE OF APPLICATION

We are going to apply the proposed method to the observed data presented in Table 1, and considering Table 3 as the matrix of assumed values for quality levels. Table 4 presents, cell by cell, the standardized residuals, based on the difference between observed and hypothesized values. This value, for each cell is obtained using Ec. 5 and indicates the value of the residuals in relation with the expected frequency under the null hypothesis, as well as the sign of this deviation.

Table 4. Standardized Residuals for the example

| | | Reference data | | | |
|---|---|---|---|---|---|
| | | Wo | G | N | Wa |
| Data classification | Wo | -0.21 | 0.78 | -0.73 | -0.69 |
| | G | 1.37 | -0.12 | 0.85 | |
| | N | | 0.09 | -0.39 | |
| | Wa | 0.71 | -0.69 | 2.02 | 0.07 |
| W=Woodland, G=Grassland, N=Non-vegetated, W=Water | | | | | |

The tests are performed in Table 5. This table shows chi square residuals ($R_{ij}^2$, that is to say, the addends of Ec. 3). The last three rows ("Analysis") show respectively the value of the test for each column (labeled by $\chi_j$), the degrees of freedom in each case (df) and the corresponding p-value.

Table 5. Tests values for the example

| | | Reference data | | | |
|---|---|---|---|---|---|
| | | Wo | G | N | Wa |
| Data classification | Wo | 0.04 | 0.61 | 0.53 | 0.48 |
| | G | 1.88 | 0.02 | 0.73 | |
| | N | 0.51 | 0.01 | 0.15 | |
| | Wa | | 0.48 | 4.08 | 0.00 |
| Analysis | $\chi_j$ | 2.43 | 1.12 | 5.49 | 1.00 |
| | d.f. | 2 | 3 | 3 | 1 |
| | p-value | 0.29 | 0.77 | 0.14 | 0.49 |
| W=Woodland, G=Grassland, N=Non-vegetated, W=Water | | | | | |

In this case we have made four contrasts, one for each column. In consequence, and in order to keep an eye on the global Type I error level, first we have to apply Bonferroni criteria, that says that in the case of multiple tests, the final decision is rejected the null hypothesis if, at least one of the p-values is lesser than the fixed α divided by the number of tests. In this case, assuming α=0.05, we would reject the

overall null hypothesis if at least one of the p-values was lesser than 0.05/4=0.0125. In this case, as we can see in Table 4, all p-values are greater than 0.0125, and we conclude that for the data of Table 1 we cannot reject the null hypothesis given the specifications shown in Table 3.

## 4. CONCLUSIONS

Confusion matrices appear in a natural way in many control processes of thematic products derived from remote sensing. Nevertheless, the evaluation of quality through a confusion matrix is not easy and many times is approached by an overall quality index. In this work, we propose a procedure for statistical testing the quality of a remote-sensed derived product through its entire confusion matrix in respect with a set of class-by-class requirements that have been previously established.

The proposed approach is based on a very well-known statistic. We have adopted a perspective of goodness of fit tests where the null hypothesis is totally specified. For each class (column of the confusion matrix) we can propose a chi-square test and accept or reject the behavior of quality for this class. When we reject the null hypothesis, the signs of the Standardized residuals tell us the meaning of the actual situation with respect to the cause of rejection. A overall chi-square test has been proposed based on the k separate chi-square tests (one per class) applying the Bonferroni's correction in order to preserve the global type I error level previously fixed.

This method allows for a very versatile control because allows: i) the use of a complete situation (multinomial) (the class and all the possible confusions), ii) the aggregation/collapse of some confused cases, iii) a binomial situation (pure class versus all the confusions).

## 5. AKNOWLEDGEMENT

This work has been partially supported by the grant: CMT2015-68276-R of the Spanish Ministry of Economy and Competitiveness with funds of the European Regional Development Fund (ERDF).

## 6. REFERENCES

[1] Ariza-López, FJ, *Fundamentos de Evaluación de la Calidad de la Información Geográfica*. Ser. Publicaciones de la Universidad de Jaén. Jaén (Spain), 2013.

[2] B Poulter, MacBean N, Hartley A, Khlystova I, Arino O, Betts R, Bontemps S, Boettcher M, Brockmann C, Defourny P, Hagemann S, Herold M, Kirches G, Lamarche C, Lederer D, Ottlé C, Peters M, and P Peylin, "Plant functional type classification for earth system models: results from the European Space Agency's Land Cover Climate Change Initiative," *Geosci. Model Dev.*, 8:2315–2328, 2015. doi:10.5194/gmd-8-2315-2015.

[3] Cai Y and K Krishnamoorty, "Exact size and Power Properties of five tests for Multinomial Proportions," *Communications in Statistics: Simulation and Computation*. 35:149-160, 2006. doi:10.1080/03610910500415993.

[4] Comber A, Fisher P, Brunsdon C and A Khmag, "Spatial analysis of remote sensing image classification accuracy," *Remote Sensing of Environment*, 127:237-246, 2012.

[5] Congalton RG and K Green, *Assessing the Accuracy of Remotely Sensed Data—Principles and Practices* (Second edition), CRC Press, Taylor & Francis Group, Boca Raton FL (USA), 2009.

[6] Dmitrienko A and R D'Agostino, "Traditional multiplicity adjustment methods in clinical trials," *Statistics in Medicine*, 32:5172-5218, 2013. doi:10.1002/sim.5990.

[7] Foody G, "Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy," *Photogrammetric Engineering & Remote Sensing*, 70(5):627–633, 2004.

[8] ISO. *ISO 19157:2013 Geographic Information – Data Quality*. International Organization for Standardization, Geneva, Switzerland, 2013.

[9] Nishii R and S Tanaka, "Accuracy and inaccuracy assessments in land-cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):491-498, 1999.

[10] Pinilla C, *Elementos de Teledetección*. Ed. Ra-Ma, Madrid (Spain), 1995.

[11] Pontius RG and M Millones, "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment," *International Journal of Remote Sensing*, 32(15):4407-4429, 2011.

[12] Senseman G, Bagley CF and SA Tweddale, *Accuracy Assessment of the Discrete Classification of Remotely-Sensed Digital Data for Landcover Mapping.* USACERL Technical Report EN-95-04, 1995.

[13] Stehman S and RL Czaplewski, "Design and analysis for Thematic Map Accuracy Assessment: Fundamental Principles," *Remote Sensing of the Environment*, 64:331-334, 1998.

[14] Tsendbazar NE, de Bruin, S Mora, B Schouten and M Herold, "Comparative assessment of thematic accuracy of GLC maps for specific applications using existing reference data," *International Journal of Applied Earth Observation and Geoinformation*, 44:124-135, 2016.

[15] Veregin H, T*axonomy of error in spatial databases*. Technical papers of the National Center for Geographic Information and Analysis. University of California, Santa Barbara (USA), 1989.