THE QUALITY CONTROL COLUMN SET: AN ALTERNATIVE TO THE CONFUSION MATRIX FOR THEMATIC ACCURACY QUALITY CONTROLS

Introduction

A classification may be considered accurate if it provides an unbiased representation of the reality (agrees with reality), or conforms to the "truth". Thematic accuracy is defined by ISO 19157 as the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships. Classification correctness is defined by the same standard as the comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference data). Classification correctness is a main concern in any remote sensed derived product (e.g. land cover, fire and drought incidence maps, etc.) and, in general, for any kind of spatial data (e.g. vector data such as cadastral parcels, road networks, topographic data bases, etc.). The main components for a thematic accuracy assessment are (Stehman and Czaplewski, 1998): i) the sampling design used to select the reference sample; ii) the response design used to obtain the reference land-cover classification for each sampling unit; and iii) the estimation and analysis procedures. But for a proper classification correctness assessment, a classification scheme is also needed. A classification scheme has two critical components (Congalton and Green, 2009): i) a set of labels, ii) a set of rules for assigning labels. From our point of view, the two previous aspects must be considered from a more general perspective of the production processes of spatial data, and from this perspective, the first thing to consider is a specification of the product (e.g. in the sense of ISO 19131). This specification should contain the classification scheme but also a specification of the level of quality required for each category (e.g. at least 90% of classification correctness for category A), and grade of confusion allowed between categories (e.g. at most 5% of confusion between categories A and B). These quality grades must be in accordance with the processes' voice (capacity to give some quality grade) and the user's voice (quality needs for a specific use case).

The confusion matrix is currently at the core of the accuracy assessment literature (Foody, 2002) and, as stated by Comber et al. (2012), the error matrix has been adopted as both the "*de facto*" and the "*de jure*" standard, the way to report on the thematic accuracy of any remotely sensed data product (e.g. image derived data). Of course, the same tool can be used for any kind of data directly originated in a vector form.

A confusion matrix and the indices derived from it are statistical tools for the analysis of paired observation. When the objective is to compare two classified data (by different processes, different operators, different times, or something similar), the observed frequencies in a confusion matrix are assumed to be modelled by a multinomial distribution (forming a vector after ordering by columns, for instance). The indexes derived, like overall accuracy, kappa, producer's and user's accuracies and so on, are based on this assumption (multinomial distribution) and they make sense due to the complete randomness of the elements inside the confusion matrix. However, this inherent randomness, that is the assumption of the underlying statistical model falls down when a true reference data is available. Suppose the reference data is located by column. If the reference data are considered as the truth, the total number of elements we know that belong to a particular category, can be correctly classified or confused with other categories, but always there will be located in the same column but never in other different column (category). This fact implies that inherent randomness of the multinomial is not possible now. However, we can deal with the available classification by considering a multinomial distribution for each category (column) instead of the initial multinomial distribution which involved all the elements in the matrix. For this reason, we will call this approach as Quality Control Column Sets (QCCS). Therefore, the goal of this paper is to present the basis of this new approach and to give an example of its application.

Quality control column set

A confusion matrix, or error matrix, is a contingency table, which is a statistical tool for the analysis of paired observations. The confusion matrix is proposed and defined as a standard quality-measure for spatial data (measure #62) by ISO 19157. For a given geographical space, the content of a confusion matrix is a set of values accounting for the degree of similarity between paired observations of k classes in a controlled data set (CDS), and the same k classes of a reference data set (RDS). Usually RDS and CDS are located by columns and by rows, respectively. So it is a $k \times k$ squared matrix. The diagonal elements of a confusion matrix contain the number of correctly classified items in each class or category, and the off-diagonal elements contain the number of confusions. So a confusion matrix is a type of similarity assessment mechanism used for thematic accuracy assessments.

CM(i,j) = [#items of class (j) of the RDS classified as class (i) in the CDS] (1)

A confusion matrix in not free of errors (Congalton and Green, 1993; Foody, 2002), and for this reason a quality assurance of intervening processes is needed; e.g. the proposal of Shehman and Czaplewski (1998) can be considered in this way (in order to apply a statistically rigorous accuracy assessment). As pointed out by Smits et al. (1999), obtaining a reliable confusion matrix is a weak link in the accuracy assessment chain. Here a key element is the RDS, denoted sometimes as "ground truth", which can be totally inappropriate and, in some cases, very misleading (Congalton and Green, 2009) and should be avoided. As pointed out by several studies, RDS often contain error and sometimes possibly more error than the CDS. Here, the mayor problem comes from the fact that classifications are often based on highly subjective interpretations. The problem of lack of quality in the reference data is still current (Congalton et al. 2014), and the thematic quality of products derived from remote sensing still presents problems. We understand that this situation is due to the fact that in most cases the RDS is simply another set of data (just another classification) and not a true reference (error free or of better quality).

The above mentioned situation does not occur in the quality assessment of other components of spatial data quality; in this way, compared to positional accuracy there is a clear lack of standardization. For example, in the case of positional accuracy, the ASPRS standard (ASPRS, 2015) establishes the following requirement: "The independent source of higher accuracy for checkpoints shall be at least three times more accurate than the required accuracy of the geospatial data set being tested". This situation is directly achievable when working with topographic and geodetic instruments, but it is not directly attainable when working with thematic categories because of the high subjectivity of interpretations. However, we believe that this situation should guide all processes for determining the RDS of an assessment of thematic accuracy.

In order to actually achieve greater accuracy for the RDS some quality assurance actions need to be deployed in order to reduce the subjectivity of the interpretations, for instance: i) using a group of selected operators, ii) designing a specific training procedure for the group of operators in each specific quality control (use case), iii) calibrating the work of the group of operators in a controlled area, iv) supplying the group with good written documentation of the product specifications and the quality control process, v) helping the group with good service support during the quality-control work and socializing the problems and the solutions and, finally vi) proceeding to the classification based on a multiple assignation process produced by the operators of the group, achieving agreements where needed. In this way Yang et al (2017) propose that validation sampling units be reviewed by 9 experts and to adopt a label requires a consensus of at least 6/9 among these experts. All these actions are quality assurance actions and must be deployed, paying special attention to improving trueness (reducing systematic differences between operators and reality), precision (increasing agreement between operators in each case) and uniformity (increasing the stability of operators' classifications under different scenarios).

If the RDS does not have the quality to be a reference, the confusion matrix can be understood as a complete multinomial. From this perspective, the analyses based on the confusion matrix are correct (e.g. overall accuracy, kappa, users' and producers' accuracies, and so on). But if the RDS does have the quality to be a reference, it is not correct to work with the complete confusion matrix because the inherent randomness in the matrix falls down. Now we can manage the data under a new approach: separating the matrix in columns (one for each category) and redefining a multinomial distribution for each category (column). Within this new approach we propose a category-wise control that allows the statement of our preferences of quality, category by category, but also the statement of misclassifications or confusions limited between classes. These preferences are expressed in terms of minimum percentages required in well-classified items and maximum percentage allowed in misclassifications between classes within each column.

In order to illustrate the application of the above with an example, Figure 1 shows a confusion matrix with results from the accuracy assessment of the classification of a synthetic data set with four categories. Now let us consider that the RDS used in this assessment does have the quality to be a reference. Therefore, the data from Figure 1 cannot be understood as a complete multinomial but rather a set of four multinomials, one for each category (column). Figure 2 illustrates this fact with locks that symbolize that the marginal of the columns are fixed and therefore the new structure "quality control column set" (QCCS) has to be considered instead of the classical method based on the confusion matrix.

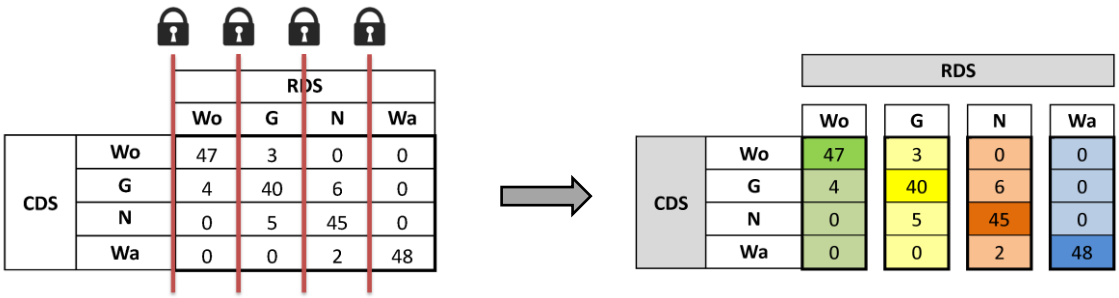

Figure 1. The new structure called "quality control column set" (QCCS) applied to data with the structure of a confusion matrix. The locks symbolize that the marginal of the columns are fixed. For clarity, each column is presented in a different colour, highlighting the number of correctly classified items. (Wo = Woodland, G = Grassland, N = Non-vegetated, Wa = Water)

Once the QCCS structure is considered our proposal allows us to consider a set of quality specifications in the following manner: for each category, a classification level could be stated but also misclassification levels with each other category (or group of them). In Table 1 we have summarized an example of quality specifications for the category Wo of Figure 1. We have indicated, the minimum percentage required for well-classified items, but also the maximum percentage allowed in misclassifications. This possibility of merging categories offers a more flexible quality control analysis. By this way, the quality specifications conform what we call quality control hypothesis set (QCHS). Each column of a QCHS allows the complete definition of a multinomial model for a category (e.g. Table 1). A QCCS supplies the observed data and a QCHS the specifications modelled by a set of multinomial, so a complete definition of a quality control has been performed and can be tested by means of an exact test based on the multinomial distribution function.

Table 1. Example of specifications: quality levels required for each category and the percentage of misclassifications allowed between classes within each category.

Category	Specification ID	Description
Woodland	SpWo#1	95% of minimum percentage required in well-classified items (≥95%)
	SpWo#2	4% of maximum percentage allowed in misclassifications with Grassland (≤4%)

	SpWo#3	1% of maximum percentage allowed in misclassifications with both Non-vegetated land and Water (≤1%)

Note: these specifications are only by way of example

Conclusions

A new approach for thematic accuracy quality control is presented. It is based on the assumption that the RDS is a reference, and this fact makes available a more powerful and complete method for thematic accuracy quality control than those based on a confusion matrix or on global indices. This method allows a class by class quality control, including some degree of misclassifications or confusions between classes. It is a very flexible procedure because it provides the possibility to merge classes, which means the possibility of varying the dimension of the underlying multinomial, and it also allows us to test simultaneously the quality levels for a set of categories.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness [grant number CMT2015-68276-R].

References

ASPRS, 2015. ASPRS Positional Accuracy Standards for Digital Geospatial Data. Photogrammetric Engineering & Remote Sensing 81 (3) : A1-A26. https://doi.org/10.14358/PERS.81.3.A1-A26

Comber, A., Fisher, P., Brunsdon, C., Khmag, A., 2012. Spatial analysis of remote sensing image classification accuracy. Remote Sensing of Environment 127, 237-246. https://doi.org/10.1016/j.rse.2012.09.005.

Congalton, R.; Green, K., 1993. A practical look at the sources of confusion in error matrix generation. Photogrammetric Engineering & Remote Sensing 59 (5), 641-644. https://www.asprs.org/wp-content/uploads/pers/1993journal/may/1993_may_641-644.pdf

Congalton, R.G., Green, K., 2009. Assessing the accuracy of remotely sensed data: Principles and practices. Lewis Publishers, Boca Raton, USA.

Congalton, R.G.; Gu, J.; Yadav, K.; Thenkabail, P.; Ozdogan, M. 2014. Global Land Cover Mapping: A Review and Uncertainty Analysis. Remote Sensing 2014 (6), 12070-12093. https://doi.org/10.3390/rs61212070

Foody G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80 (1), 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4.

Smits, P. C., Dellepiane, S. G., 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, 1461–1486. https://doi.org/10.1080/014311699212560.

Stehman, S.V., Czaplewski, R., 1998. Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. Remote Sens. Environ 64 (3), 331–344. https://doi.org/10.1016/S0034-4257(98)00010-8.

Yang, Y.; Xiao, P.; Feng, X.; Li, H. 2017. Accuracy assessment of seven global land cover datasets over China. ISPRS Journal of Photogrammetry and Remote Sensing 125, 156-173. https://doi.org/10.1016/j.isprsjprs.2017.01.016