Malta, January 2020

3rd INTERNATIONAL WORKSHOP ON SPATIAL DATA QUALITY

The Quality Control Column Set: an alternative to the Confusion Matrix for Thematic Accuracy Quality Controls

José Rodríguez-Avi, Francisco Javier Ariza-López, Mª Virtudes Alba-Fernández, José Luis García-Balboa

[jravi, fjariza, mvalba, jlbalboa]@ujaen.es;

Universidad de Jaén; Paraje de las Lagunillas S/N, E-23.071-Jaén

Funded by the Project: MEMCALIG (Modelización Estadística de Matrices de Confusión en Calidad de la Información Geográfica)

GOBIERNO DE ESPAÑA

MINISTERIO DE ECONOMIA Y COMPETITIVIDAD

PROYECTO COFINANCIADO POR LA UNIÓN EUROPEA

FONDO EUROPEO DE DESARROLLO REGIONAL

Objectives

Our goals:

- A new reformulation of a confusion matrix.
- The definition of a QCCS (Quality Control Column set)
- Some proposal about the statistical analysis of a QCCS.

Contents

- Introduction
- A limitation of Confusion Matrices
- QCCS definition
- Why a QCCS is proposed?
- Further analysis of a QCCS
- Example
- Conclusion

Introduction

Confusion Matrix (CM)

Definition:

- **It is a contingency table, which is a statistical tool for the analysis of paired observations (between equals sources).**
- **The content of a CM is a set of cell 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)**
- **A multinomial approach could be taken into account**

$$CM \sim \mathcal{M}(n, p_{11}, \dots, p_{ij}, \dots, p_{kk})$$

→ Overall Accuracy
Kappa index
Etc.

- **It is proposed and defined as a standard quality measure for spatial data (measure #62) by ISO 19157.**

Introduction

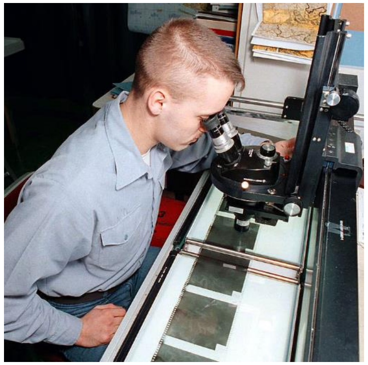
Confusion Matrix (CM)

- It is a $k \times k$ squared matrix with the same categories by row and columns and in the same order
- Diagonal elements count number of correctly classified items
- Off-diagonal elements count the number of confusions
- For convenience, we set RDS by columns and CDS by rows.

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

Introduction

Confusion Matrix (CM)



	Asphalt	Concrete	Grass	Tree	Building	Total
Asphalt	2385	4	0	1	4	2394
Concrete	0	332	0	0	1	333
Grass	0	1	908	8	0	917
Tree	0	0	0	1084	9	1093
Building	12	0	0	6	2053	2071
Total	2397	337	908	1099	2067	6808

A limitation of Confusion Matrices

- This situation does not occur in the quality assessment of other components of spatial data quality
- *"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".* (ASPRS, 2015)

The CM is not valid

The multinomial approach is not valid

$$CM \sim \mathcal{M}(n, p_{11}, \ldots, p_{ij}, \ldots, p_{kk})$$

~~Overall Accuracy~~
~~Kappa index~~
~~Etc.~~

A limitation of Confusion Matrices

- In qualitative aspects the highest accuracy of the RDS is achieved through assurance and multiple assignment. This implies some requirements:
  - 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,
  - vi) proceeding to the classification based on a multiple assignation process produced by the operators of the group, achieving agreements where needed.

A limitation of Confusion Matrices

- All these actions are quality assurance actions and must be deployed, paying special attention to
  i. Improving trueness (reducing systematic differences between operators and reality),
  ii. Precision (increasing agreement between operators in each case),
  iii. Uniformity (increasing the stability of operators' classifications under different scenarios).

- This is a more complex and expensive procedure, but multiple advantages are obtained in order to assure the quality.

QCCS definition

Consequences

- In the later case for a RDS, the CM cannot be seen as a complete multinomial:

    i. the inherent randomness in the complete matrix falls down.
    ii. The number of diagonal elements cannot superate the corresponding column size.
    iii. The analyses based on the CM (overall accuracy, kappa, users' and producers' accuracies, and so on) are incorrect.

A new approach is needed

QCCS definition

New approach

- Consists on:
  i. separate the matrix in columns (one for each category) and
  ii. redefining a multinomial distribution for each category (column).

- We propose:
  i. A category-wise control that allows the statement of our preferences of quality, category by category, but also
  ii. the statement of misclassifications or confusions limited between classes.
  iii. 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

QCCS definition

New approach

		RDS			
		Wo	G	N	Wa
CDS	Wo	47	3	0	0
	G	4	40	6	0
	N	0	5	45	0
	Wa	0	0	2	48

		RDS			
		Wo	G	N	Wa
CDS	Wo	47	3	0	0
	G	4	40	6	0
	N	0	5	45	0
	Wa	0	0	2	48

		RDS			
		Wo	G	N	Wa
CDS	Wo	47	3	0	0
	G	4	40	6	0
	N	0	5	45	0
	Wa	0	0	2	48

Why a QCCS is proposed

Once a QCCS is considered

- We can determine a set of quality specifications (one for each category)
- For each category a classification level could be stated but also misclassification levels with each other category (or group of them)
- Classification levels are independent among them: each column has its own specification
- Classifications may differ in respect with
  1. the percentage of well-classified elements
  2. The percentage of errors allowed between the true category and others categories
  3. The number of total specifications
- A whole decision about quality can be obtained, as well as partial decision for each column (or subset of columns)

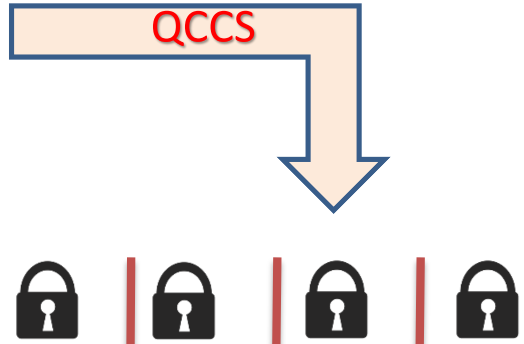
The Quality Control Column Set: an alternative to the Confusion Matrix for Thematic Accuracy Quality Controls

Example

Category	Specification ID	Description
Woodland	SPWo#1	90% of minimum percentage required in well-classified items (≥90%)
	SpWo#2	7% of maximum percentage allowed in misclassifications with Grassland (≤7%)
	SpWo#3	3% of maximum percentage allowed in misclassifications with both Non-vegetated and Water (≤3%)
Grassland	SpG#1	80% of minimum percentage required in well-classified items (≥80%)
	SpG#2	15% of maximum percentage allowed in misclassifications with Non-vegetated (≤15%)
	SpG#3	5% of maximum percentage allowed in misclassifications with both Woodland and Water (≤5%)
Non-vegetated	SpN#1	85% of minimum percentage required in well-classified items (≥85%)
	SpN#2	10% of maximum percentage allowed in misclassifications with Grassland (≤10%)
	SpN#3	5% of maximum percentage allowed in misclassifications with both Woodland and Water (≤5%)
Water	SpWa#1	95% of minimum percentage required in well-classified items (≥95%)
	SpWa#2	5% of maximum percentage allowed in misclassifications with the rest of categories (≤5%)
Note: these specifications are only by way of example		

Example

		Reference data			
		Wo	G	N	Wa
Data classification	Wo	80	10	10	2
	G	15	36	15	5
	N	5	5	66	0
	Wa	0	3	5	83
Wo=Woodland, G=Grassland, N=Non-vegetated, Wa=Water					



		RDS			
		Wo	**G**	**N**	**Wa**
CDS	**Wo**	80	10	10	2
	G	15	36	15	5
	N	5	5	66	0
	Wa	0	3	5	83

Example

		RDS
		Wo
CDS	Wo	≥ 0.90
	G	≤ 0.07
	N/Wa*	≤ 0.03

* N and Wa grouped because of SpWo#3

		RDS
		G
CDS	G	≥ 0.80
	N	≤ 0.15
	Wo/Wa*	≤ 0.05

* Wo and Wa grouped because of SpG#3

		RDS
		N
CDS	N	≥ 0.85
	G	≤ 0.15
	Wo/Wa*	≤ 0.05

* Wo and Wa grouped attending to SpN#3

		RDS
		Wa
CDS	Wa	≥ 0.95
	G/N/Wo*	≤ 0.05

* G, N and Wo grouped attending to SpWa#2

Values in assumed order

		RDS
		Wo
CDS	Wo	80
	G	15
	N/Wa*	5

* N and Wa grouped because of SpWo#3

		RDS
		G
CDS	G	36
	N	5
	Wo/Wa*	13

* Wo and Wa grouped because of SpG#3

		RDS
		N
CDS	N	66
	G	15
	Wo/Wa*	15

* Wo and Wa grouped attending to SpN#3

		RDS
		Wa
CDS	Wa	83
	G/N/Wo*	7

* G, N and Wo grouped attending to SpWa#2

Values in assumed order

Example

Case: Information about the "whole" matrix: *multinomial*

Woodland:

$\mathbb{H}_0$: $\pi_{Wo} = 0.90$; $\pi_{Wo,G} = 0.07$; $\pi_{Wo,others} = 0.03$

$\mathbb{H}_1$: $\pi_{Wo} < 0.90$ *or* $\pi_{Wo} = 0.90$ *and* $\pi_{Wo,G} < 0.07$

$T_{Wo} = (80, 15, 5)$

$\mathcal{M}(100; 0.90, 0.07; 0.03)$.

p-value=0.001

Grassland:

$\mathbb{H}_0$: $\pi_G = 0.80$; $\pi_{G,N} = 0.15$; $\pi_{G,others} = 0.05$

$\mathbb{H}_1$: $\pi_G < 0.80$ *or* $\pi_G = 0.80$ *and* $\pi_{G,N} < 0.15$

$T_G = (36,5,13)$

$\mathcal{M}(54; 0.80, 0.15; 0.05)$

p-value=0.007

Non vegetated:

$\mathbb{H}_0$: $\pi_N = 0.85$; $\pi_{N,G} = 0.10$; $\pi_{N,others} = 0.05$

$\mathbb{H}_1$: $\pi_N < 0.85$ *or* $\pi_N = 0.85$ *and* $\pi_G < 0.10$

$T_N = (66,15,15)$

$\mathcal{M}(96; 0.80, 0.15; 0.05)$.

p-value=0.000

Water:

$\mathbb{H}_0$: $\pi_{Wa} = 0.95$

$\mathbb{H}_1$: $\pi_{Wa} < 0.95$

$T_{Wa} = 48$

$B(90, 0.95)$

p-value= 0.164

Conclusions

- A new approach of a confusion matrix has been presented.
- It is based on the assumption that the RDS is a reference (ground truth)
- This give 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
- It also allows us to test simultaneously the quality levels for a set of categories
- An example has been provided

THANKS FOR YOUR ATTENTION!!!