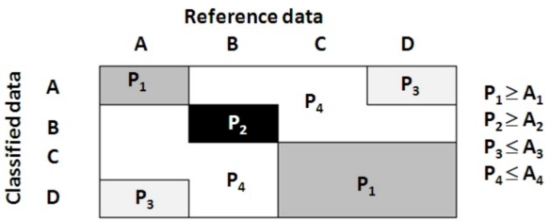
Malta, February 2018

2nd INTERNATIONAL WORKSHOP ON SPATIAL DATA QUALITY



Invited session on Thematic accuracy assessment by means of confusion matrices

Francisco Javier Ariza-López[1] , José Rodríguez-Avi[1] ,Virtudes Alba-Fernández[1], José Luis García-Balboa[1] and Inmaculada Barranco-Chamorro[2]

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

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

Funded by: MEMCALIG: Modelización Estadística de Matrices de Confusión en Calidad de la Información Geográfica

GOBIERNO DE ESPANA
MINISTERIO DE ECONOMIA Y COMPETITIVIDAD

PROYECTO COFINANCIADO POR LA UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL

http://coello.ujaen.es/investigacion/web_giic/SubWeb_MEMCALIG/index.htm

Observer and Promoter Entities of the MEMCALIG project:

National Mapping agencies

INSTITUTO GEOGRÁFICO NACIONAL cnig
GOBIERNO DE ESPAÑA MINISTERIO DE FOMENTO
INEGI INSTITUTO NACIONAL DE ESTADISTICA GEOGRAFIA E INFORMATICA
EXERCITO BRASILEIRO SERVIÇO GEOGRÁFICO

Regional Mapping agencies

ICGC Institut Cartogràfic i Geològic de Catalunya
JUNTA DE ANDALUCIA Instituto de Estadística y Cartografía de Andalucía CONSEJERÍA DE ECONOMÍA, INNOVACIÓN, CIENCIA Y EMPLEO

Companies:

tracasa

Objectives

- **Thematic accuracy.**
- **Explain what a confusion matrix is.**
- **Explain the differences between confusion matrices of vector and raster data.**
- **Explain how to create a confusion matrix in a rigorous way** (the main components for a thematic accuracy assessment).
- **Present some traditional quality control techniques based on confusion matrices** (user's and producer's risk, overall agreement and Kappa)**.**
- **Present some new quality control techniques based on confusion matrices and developed within MEMCALIG project** (e.g. bias, similarity, several quality levels, etc.)**.**

Contents

- **Introduction** (a reminder about thematic accuracy).
- **What is a confusion matrix** (applied and statistical points of view).
- **Confusion matrices in raster and vector data.**
- **How to create a confusion matrix** (main components of a thematic accuracy assessment).
- **Traditional quality control based on confusion matrices** (User's and producer's risk, Overall Agreement, and so on).
- **Frame #1: Control respect stablished product's specifications** (hypothesized values)
- **Frame #2: Control between observed values.**
- **Other advances of the MEMCALIG project.**
- **Conclusions.**

ARIZA

AVI

ALBA

Introduction

Thematic accuracy is….

7.4.5 Thematic accuracy

Thematic accuracy is defined as the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships. It consists of three data quality elements:

— classification correctness – comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference data);

— non-quantitative attribute correctness – measure of whether a non-quantitative attribute is correct or incorrect;

— quantitative attribute accuracy – closeness of the value of a quantitative attribute to a value accepted as or known to be true.

ISO 19157

Introduction

Universe of discourse

UNIVERSE OF DISCOURSE (UoD)

View of the real or hypothetical world that includes everything of interest → **Categories**

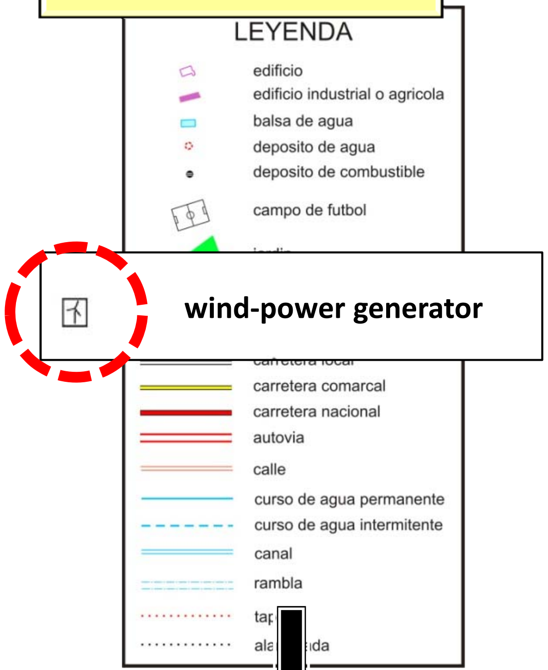
NOMINAL TERRAIN:

Materialization of the Universe of discourse in the field, as a selection of elements from the real world.

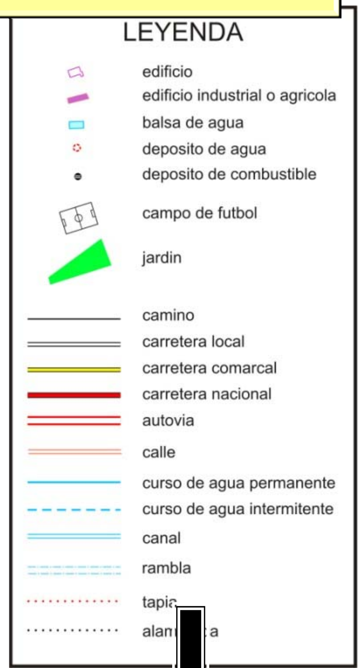
"Ideal" data set with which a product is compared.

ISO 19131

UoD 1



UoD 2



Real world

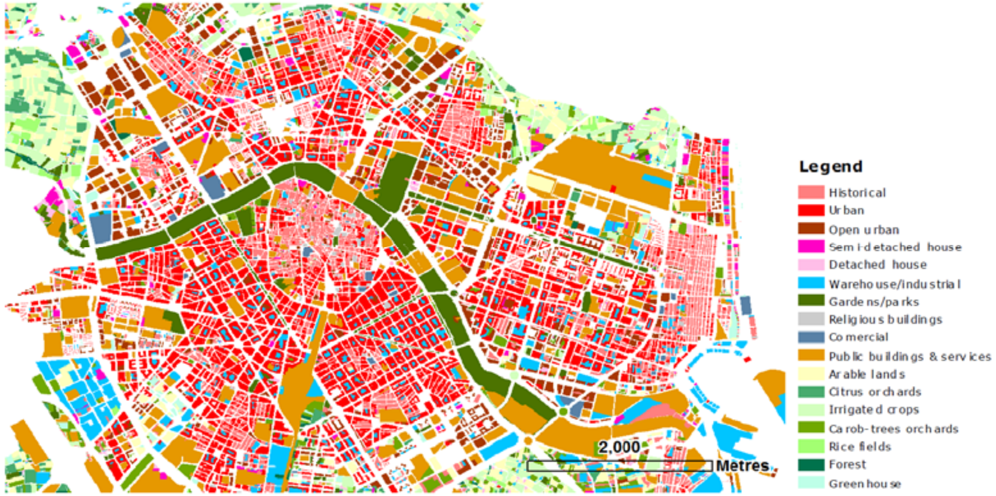
Nominal ground 1

Nominal ground 2

Francisco Javier Ariza-López

Introduction

Raster data



- **Raster Model: a model of reality based on the tiling of space.**
- **An exhaustive representation is made.**
- **Model oriented to manage the occupation of space, not its objects.**
- **There are no gaps.**

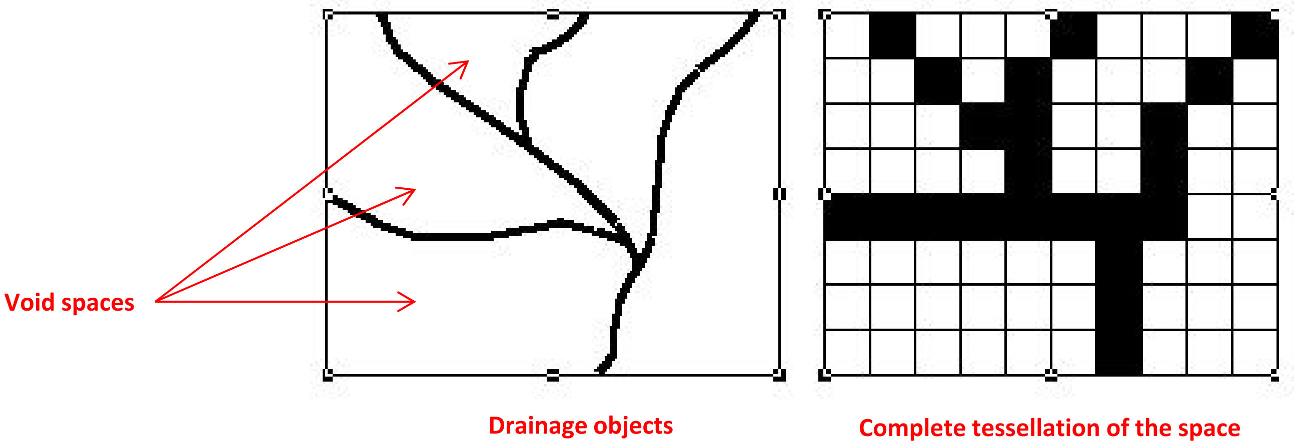
  The latter indicates that there can be neither pure omissions nor pure commissions, only confusions can exist

Introduction

Vector data

- Vector model: a model of reality based on the representation of objects.
- There is no exhaustive representation of the space, only the objects of interest.
- Model oriented to manage objects.
- There may be gaps between the objects.

The latter indicates that there may be pure omissions and pure commissions, not only may there be confusion

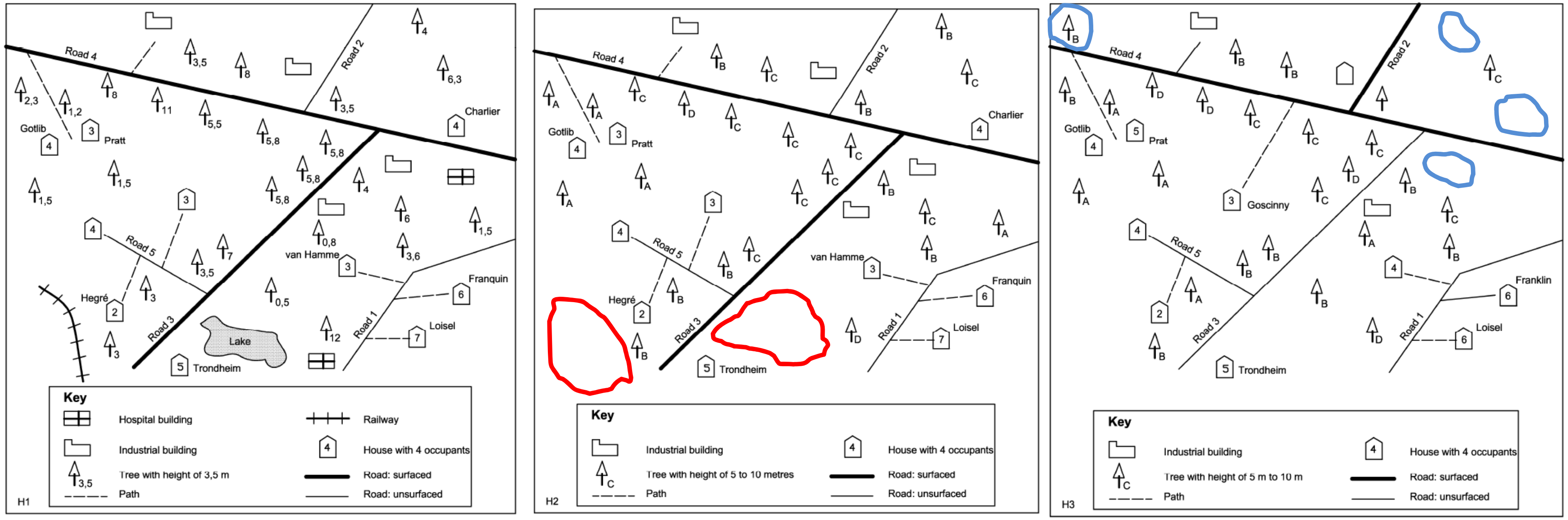


Francisco Javier Ariza-López

Introduction

Vectorial data

Real world →specifications (ISO 19131)→ Universe of discourse →Data Capture (ISO 19131)→ Data set



Confusion matrix

ISO 19157

	Path	Road	Tree	Ibuilding	House	Commisions	
Path	4	0	0	0	0		4
Road	2	5	0	0	0		7
Tree	0	0	23	0	0		23
Ibuilding	0	0	0	2	0		2
House	0	0	0	1	9		10
Ommisions							
Sum	6	5	23	3	9		

For vector data

Introduction

Examples

Overall global confusion matrix

SIRS \ IFI	110	113	121	122	123	124	131	133	134	141	142	200	300	500
110	22786	61	649	301	3	1	0	40	30	314	132	249	74	4
113	61	89	18	1	0	0	0	0	0	2	2	41	11	0
121	1055	15	10919	367	63	123	47	91	70	266	127	224	59	19
122	52	0	62	1357	0	4	1	3	2	12	3	12	3	2
123	2	0	96	9	572	0	3	2	2	2	0	3	0	1
124	1	0	31	5	0	1451	0	1	0	2	6	20	9	0
131	3	0	34	7	0	2	960	20	23	2	2	128	10	5
133	20	0	75	30	5	5	12	348	39	8	4	45	3	0
134	65	1	61	21	4	1	1	6	217	62	12	96	2	0
141	228	0	61	79	3	1	2	4	43	2844	136	357	320	24
142	193	3	137	50	10	22	0	7	6	85	3354	214	15	4
200	501	20	308	269	18	38	63	36	311	294	123	33424	545	69
300	115	4	30	60	1	4	4	3	6	736	76	940	12105	13
500	4	0	6	6	10	0	13	1	0	16	9	16	3	2288
Total	25086	193	12487	2562	689	1652	1106	562	749	4645	3986	35769	13159	2429
Producer accuracy	90,8%	46,1%	87,4%	53,0%	0,0%	87,8%	86,8%	3,6%	3,1%	61,2%	84,1%	93,4%	92,0%	0,2%
Omission error	9,2%	53,9%	12,6%	47,0%	100,0%	12,2%	13,2%	96,4%	96,9%	38,8%	15,9%	6,6%	8,0%	99,8%

Overall accuracy = 88,2%

Urban Atlas (EC, 2011)

Wrong interp Val-Cod2

CLC 2000	111	112	121	122	123	124	131	132	133	141	142	211	212	213	221	222	223	231	241	242	243	244	311	312	313	321	322	323	324	331	332	333	334	335	411	412	421	422	423	511	512	521	522	Total
111	0	3							1																																			4
112		0	1								1				1	1		3		1	1																								9
121			0						1																																				1
122				0																																								
123					0																																							
124						0																																						
131							1											1					1																					3
132								0																																					
133		1							0		1							1																										3
141										0	1																																	1
142											0							3																			1							4
211		4										1			2	4	3	69		46	19	4	1		3	9			2			1													168
212		1											0			1		1		2	1																								6
213														0																															
221												1			0	1	1			8	1																								12
222												2			1	0	1			6	3				1																						14
223												2					0	1		6	2								1																		12
231												29						0		44	14		1		1	4	2		2						3									100
241															1	1		0	1	3									1																7
242		1									1	40	2		4	6	3	41	1	0	12	1	1		1	1			3																118
243												15				1		26		14	0	1	8	3	4	2	3	7	13			1													98
244																						0	1					2	1																4
311		1						1											1			6	3	2	7	32		1	9	10				1							1				75
312																		3		3	4		7	2	60			6	26			1				3					1			116
313												1						3			1		11	24	0		2	4	4				1												51
321			1									2						7			3	2	2		1	0	10	6	10		1	1			1										47
322																			7		1	3		1	1		10	0	1	4	1	2	2				1									34
323												1			1		1	2		1	9	2	3	2		14		0	5		2	2														45
324				1								1						3			6	1	9	9	26	1	3	11	0			1				6					1			79
331																													0																
332																											3			1		0	3														7
333												1								1			1			4	2	3			1	0														13
334																																	0											
335																																		0										
411																		2					1						3							0											6
412																							1			5	2		4			3				0	1									16
421																												1											0									1
422																																						0							
423																																							0						
511																																								0					
512												1													1												1						0			3
521																																										0		
522																																											0	
Total		11	2	1			1	1	1	1	4	97	2		9	15	10	174	1	134	88	14	51	49	129	53	25	51	89	1	6	15	2		5	10	2			1	2			1,057

Corine Land Cover (EEA, 2006)

Confusion Matrix

Definition

- **Alternative names**: error matrix, miss-classification matrix, contingency table.
- **Provides**:
  - A general view of the assignments, both correct ones (elements of the diagonal) and incorrect (elements outside the diagonal).
  - Omission errors in a category (producer's accuracy) NOT PURE OMISSIONS.
  - Commission errors in a category (user's accuracy) NOT PURE COMISSIONS.
- **Uses**: quality assessment in photointerpretation, remote sensing, thematic products, thematic accuracy of topographic maps.
- **Size**: great variability.

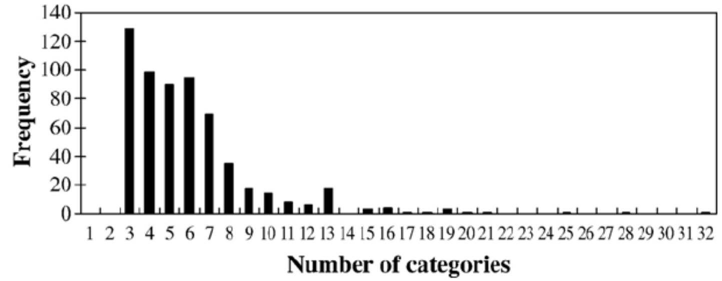


Fig. 1. The frequencies of the number of categories for the error matrices used in this study.

Francisco Javier Ariza-López

Confusion Matrix

Definition (ISO 19157)

Table D.65 — Misclassification matrix

Line	Component	Description
1	Name	misclassification matrix
2	Alias	confusion matrix
3	Element name	classification correctness
4	Basic measure	-
5	Definition	matrix that indicates the number of items of class (i) classified as class (j)
6	Description	The misclassification matrix (MCM) is a quadratic matrix with n columns and n rows. n denotes the number of classes under consideration. MCM (i,j) = [# items of class (i) classified as class (j)] The diagonal elements of the misclassification matrix contain the correctly classified items, and the off diagonal elements contain the number of misclassification errors.
7	Parameter	Name: n Definition: number of classes under consideration Value Type: Integer
8	Value type	Integer
9	Value structure	Matrix ($n \times n$)
10	Source reference	-

		Dataset class			
True class		A	B	C	Count
	A	7	2	1	10
	B	1	2	2	5
	C	1	1	3	5
	Count	9	5	6	20

Line	Component	Description
11	Example	(table above)
12	Identifier	62

Table D.66 — Relative misclassification matrix

Line	Component	Description
1	Name	relative misclassification matrix
2	Alias	-
3	Element name	classification correctness
4	Basic measure	-
5	Definition	matrix that indicates the number of items of class (i) classified as class (j) divided by the number of items of class (i)
6	Description	The relative misclassification matrix (RMCM) is a quadratic matrix with n columns and n rows. n denotes the number of classes under consideration. RMCM (i,j) = [# items of class (i) classified as class (j)] / (# items of class (i)] × 100 %
7	Parameter	Name: n Definition: number of classes under consideration Value Type: Integer
8	Value type	Real
9	Value structure	Matrix ($n \times n$)
10	Source reference	-
11	Example	-
12	Identifier	63

ISO 19157

Confusion Matrix

Some conditions are needed :

1st Over the categories:

- Categories totally exhaustive → every piece of UoD will be labeled.
- Categories are mutually exclusive → no one piece of the UoD can receive more than one label.
- Categories with a hierarchical structure → classes can be lumped into more general groups.

2nd Statistical:

- Sample size enough → in order to assure the representativeness of the sample error matrix.
- No autocorrelation is introduced → use of sampling methods that avoid autocorrelation.

Confusion Matrix

Content and nomenclature

Content: square array of numbers set out in rows and columns that expresses the number of sample units assigned to a particular category in one classification relative to the number of sample units assigned to a particular category in another classification.

Organization: Columns content the reference values and rows observed values.

Nomenclature: k is the number of classes, the matrix is $k \times k$

columns are $(A_1, \ldots, A_k)$, rows are $(a_1, \ldots, a_k)$

Observed	True				
	A_1		A_j		A_k
a_1	n_{11}		n_{1j}		n_{1k}
a_i	n_{i1}		n_{ij}		n_{ik}
a_k	n_{k1}		n_{kj}		n_{kk}
Total	n_{+1}		n_{+j}		n_{+k}

n_{ii} -> number of sample elements observed in Category a_j when the correct category is A_j: Correct.

$n_{ij}, i \neq j$ -> number of sample elements observed in Category a_i when the correct category is A_j: Error

$$N = \sum_{j=1}^{k} n_{+j}$$

Quality controls based on confusion matrices

Considerations for QC

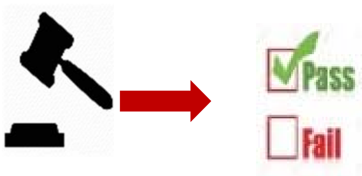
- We assume that, in a confusion matrix, a data product is compared against a source of better accuracy, which may be considered as the true population value.
- The truth/reference are in columns → This implies that the actual number of elements in each column must be equal to the total of the column.
- An element can never change between the diagonal.
- An element only belongs to a one column.
- In this way, the model can be presented as multinomial random variables by columns, one multinomial for each column.

Quality controls based on confusion matrices

Considerations for QC

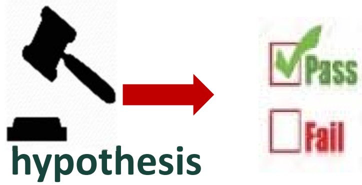
- We use a confusion matrix as a tool for an analysis with a user-risk perspective in a framework of quality control (ISO 19157).
- For this, a desired quality level is proposed (ISO 19131), a Type I error level is assumed and we want to test if our product is according to this quality level, or, on the contrary, it must be rejected.

Statistical hypothesis testing is USED¡¡¡



Quality controls based on confusion matrices

Quality Control / Hypothesis testing



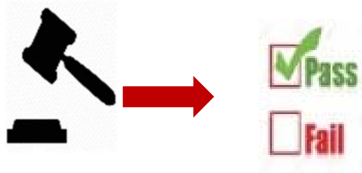
- In QC we use an statistical tool to make a decision in respect with a population hypothesis
- The hypothesis that would be contrasted is fixed in advance and it is called "null hypothesis", ($\mathbb{H}_0$).
- The decision we adopt if we consider that $\mathbb{H}_0$ is false is called "Alternative hypothesis" ($\mathbb{H}_1$)

- $\mathbb{H}_0$ can be true or not, and we can accept or reject it. So, four situations may be allowable:

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$ ☺	Type II Error β ☹
	The Alternative Hypothesis is True	Type I Error α ☹	Accurate $1 - \beta$ ☺

Quality controls based on confusion matrices

Quality Control / Hypothesis testing



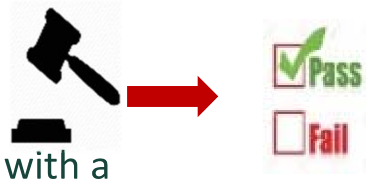
Procedure

- State the null $\mathbb{H}_0$, and the alternative hypotheses $\mathbb{H}_1$.
- Determine the appropriate test statistic used to decide about the trueness of the null hypothesis, T.
- Derive the probability distribution of the test statistic under the null hypothesis.
- Select a value of α, the probability threshold below which the null hypothesis will be rejected.
- Compute from the observations the observed value of the test statistic, $T_{obs.}$
- Calculate the probability that T exceeds T_{obs}, called "p-value", p.
- Decide to reject the null hypothesis in favor of the alternative if $p < \alpha$, or not reject it if $p > \alpha$.

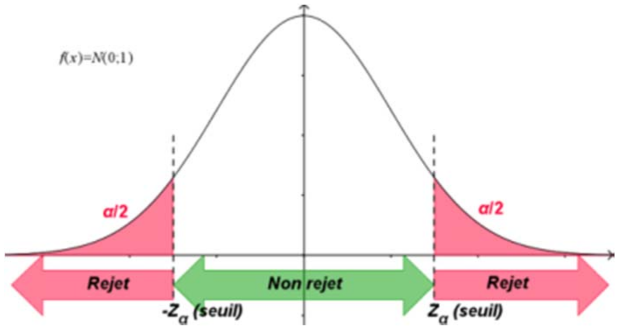
Francisco Javier Ariza-López

Quality controls based on confusion matrices

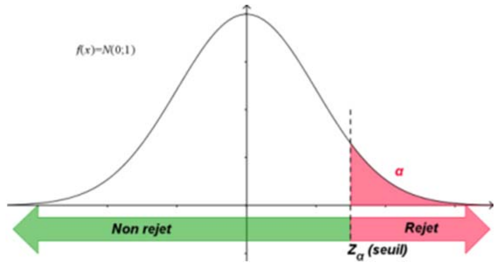
Quality Control / Contrast of hypothesis

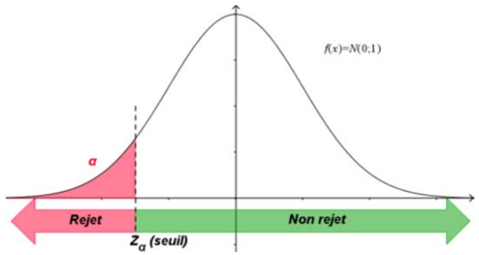


- **Bilateral test**: the null hypothesis consists in testing the equality of the test value with a given value. Actually, the reject of the hypothesis is decided whether the test value is significantly different, even it is inferior (link reject area) or superior (right reject area).
- **Unilateral test**: The null hypothesis checks if a value is superior or equal to the test value (left unilateral) or inferior or equal to this value (right unilateral).



programmed obsolescence





Main components for a thematic accuracy assessment

Phases

The main components for a thematic map accuracy assessment

1st Specification of the product

2nd The reference

3rd The response design

4th The analysis and estimations

5th The control of the control

Main components for a thematic accuracy assessment

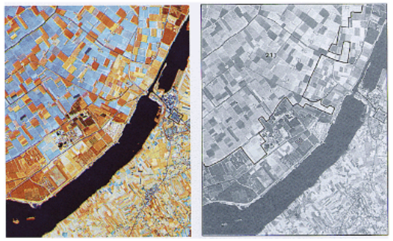
1st The data product specification

General specifications:

- A complete definition of all the elements composing the universe of discourse (categories)
- An appropriate classification system of categories (totally exhaustive, mutually exclusive, hierarchical structure, etc.)

2.1.1. Non-irrigated arable land

Cereals, legumes, fodder crops, root crops and fallow land. Includes flowers and tree (nurseries cultivation and vegetables, whether open field or under plastic or glass (includes market gardening). Includes aromatic, medicinal and culinary plants. Does not include permanent pasture.



EEA (1997)

2.1.1. Netherlands/Area: Hardenwijk Interpretation

Landsat TM 4.5.3. 1:100 000, May 1989

The arable land shown here is made up of a large number of recently ploughed parcels (blue tone on the image).

Ploughed land with no productive vegetal cover on the date of data acquisition belongs under this category.

Care must be taken not to confuse item 2.1.1 with other agricultural cover (2.1.2, 2.3.1, 2.4.1 and 2.4.4). To remove all doubt it may be necessary to consult ancillary data (aerial photographs, agricultural calendars, statistics, multidate data).

Temporary and artificial pasture (fodder crops) under rotation are to be included under 2. 1. 1.

a.Thematic similarity

LC classes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	.86	1											
3	.86	.71	1										
4	.71	.86	.86	1									
5	.86	.86	.86	.86	1								
6	.71	.71	.71	.71	.86	1							
7	.61	.75	.68	.82	.75	.79	1						
8	.39	.46	.54	.61	.57	.54	.64	1					
9	.57	.71	.57	.71	.68	.75	.79	.57	1				
10	.14	.14	.14	.14	.14	.14	.14	.21	.07	1			
11	.07	.07	.07	.07	.07	.07	.07	.14	.14	.64	1		
12	.29	.29	.29	.29	.29	.29	.14	.14	.21	.71	.71	1	
13	.07	.07	.07	.07	.07	.07	.07	.14	.14	.64	.86	.64	1

b. Agriculture assessments

LC classes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	1	1											
3	1	1	1										
4	1	1	1	1									
5	1	1	1	1	1								
6	1	1	1	1	1	1							
7	1	1	1	1	1	1	1						
8	0	0	0	0	0	0	0	1					
9	1	1	1	1	1	1	1	0	1				
10	1	1	1	1	1	1	1	0	1	1			
11	1	1	1	1	1	1	1	0	1	1	1		
12	1	1	1	1	1	1	1	0	1	1	1	1	
13	1	1	1	1	1	1	1	0	1	1	1	1	1

c. General Circulation Models (GCM)

LC classes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	.64	1											
3	.53	.50	1										
4	.34	.56	.60	1									
5	.43	.48	.68	.74	1								
6	.14	.20	.42	.47	.48	1							
7	.10	.17	.30	.27	.34	.70	1						
8	.13	.31	.35	.61	.55	.77	.56	1					
9	.22	.29	.32	.38	.40	.56	.75	.65	1				
10	.08	.14	.16	.20	.29	.31	.51	.28	.49	1			
11	.06	.00	.08	.06	.16	.23	.44	.26	.44	.67	1		
12	.09	.02	.11	.06	.19	.24	.51	.26	.47	.64	.72	1	
13	.08	.00	.10	.06	.16	.23	.44	.29	.43	.67	.89	.72	1

d. Dynamic Global Vegetation Models (DGVM)

LC classes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	.53	1											
3	.83	.53	1										
4	.59	.58	.59	1									
5	.66	.41	.64	.48	1								
6	.59	.46	.59	.60	.47	1							
7	.61	.55	.60	.61	.51	.62	1						
8	.71	.58	.71	.61	.52	.54	.69	1					
9	.58	.47	.53	.55	.57	.46	.54	.52	1				
10	.58	.61	.57	.72	.41	.60	.70	.62	.55	1			
11	.53	.61	.53	.67	.38	.53	.63	.61	.52	.72	1		
12	.57	.64	.57	.67	.41	.57	.75	.66	.53	.82	.77	1	
13	.53	.65	.53	.67	.38	.53	.63	.61	.60	.73	.83	.77	1

Quality levels:

- Overall expected quality level for the product (e.g. OA≥ 95%)
- Specific quality levels for each category (PA(class1) ≥ 90%, PA(class3) ≥ 80%, etc.).
- Specific confusion restrictions, or allowances, between categories.

Francisco Javier Ariza-López

Main components for a thematic accuracy assessment

2nd The reference

Independence from the product.

Higher accuracy: Qualitative attributes are more problematic than quantitative, except in cases where the qualitative attribute is derived from a quantitative attribute by means of a reclassification. If no instrument is involved:

- **An operator makes a judgment → It is very conditioned by the operators.**
- **In this case, the method affects a lot. → The method and degree of specification/definition are essential.**

We must ensure:

Accuracy (trueness, precision) and uniformity (in the sense of ISO 19157)

Main components for a thematic accuracy assessment

2nd The reference

The reference source: Existing data, new data, new field survey…

When: before, at the same time, after…

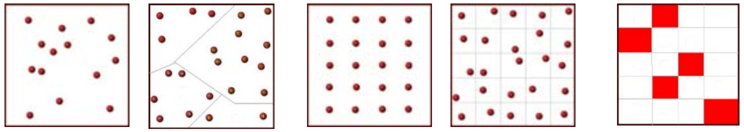
How much temporary distance can exist between the product and the reference?

Which sampling unit: 0.1 ha pixel, 10 ha polygon, 1000 ha circular plot…

The frame. List frames: a list of all sampling units (e.g. all pixels or polygons in a region). Area frames: a map or description of the population boundaries.

The probability sampling design

The sampling scheme



Main components for a thematic accuracy assessment

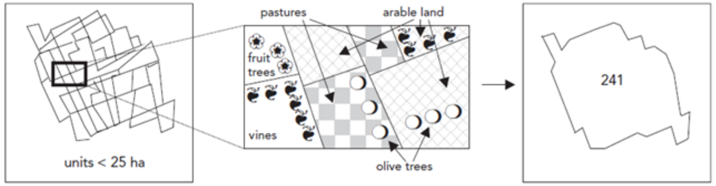
3rd The response design

The response design is the protocol for determining the reference of a sampling unit.
The response design includes procedures to collect information pertaining to the reference determination, and rules for assigning one or more reference classifications to each sampling unit.
The response design is applied to each sampling unit to obtain the reference classification, and the comparison of the map and reference classifications is conducted on the scale of a sampling Unit.

Evaluation protocol.

Information and training for operators.

Labeling protocol. Single classes and mixtures, etc.



Single  **or Multiple**  **imputation method**

Use of graphic records 

Main components for a thematic accuracy assessment

4th The analysis and estimations

Which accuracy parameters should be used?

Are they suitable for the product specifications?

Are they congruent with the decisions of the reference sample?

Are their statistical hypotheses fulfilled?

Main components for a thematic accuracy assessment

5th The control of the control

The control of the control **IS POSSIBLE**.

It is important to have same degree of knowledge about the quality of quality assessment.

This information can be used for metaquality information purposes.

Same cautions as for the field work are needed → operators, method, etc.

Accuracy parameters

Three main perspectives of accuracy parameters derived from a confusion matrix:

- **Per case:** Oriented to inform about an unique cell of a matrix.
- **Per class:** Oriented to inform about a class (complete) or about the diagonal cell of a class.
- **Global:** Oriented to inform about the complete matrix (or about all diagonal cells).

Accuracy parameters

Accuracy parameters (category-level)

Category-level accuracy measures (derived from the error matrix in Table 1)

No.	Name	Formula	References
1.	User's accuracy	$ua_i = p_{ii}/p_{i+}$	Story and Congalton (1986)
2.	Producer's accuracy	$pa_i = p_{ii}/p_{+i}$	Story and Congalton (1986)
3.	Average of user's and producer's accuracy	$aup_i = (ua_i + pa_i)/2$	
4.	Individual classification success index	$ICSI_i = ua_i + pa_i - 1$	Koukoulas and Blackburn (2001), Türk (2002)
5.	Hellden's mean accuracy	$mah_i = 2/(1/ua_i + 1/pa_i) = 2p_{ii}/(p_{i+} + p_{+i})$	Hellden (1980), Rosenfield and Fitzpatrick-Lins (1986)
6.	Short's mapping accuracy	$mas_i = p_{ii}/(p_{i+} + p_{+i} - p_{ii})$	Short (1982), Rosenfield and Fitzpatrick-Lins (1986)
7.	Conditional *kappa* (user's)	$cku_i = (ua_i - p_{+i})/(1 - p_{+i}) = (p_{ii} - p_{i+}p_{+i})/(p_{i+} - p_{i+}p_{+i})$	Rosenfield and Fitzpatrick-Lins (1986)
8.	Conditional *kappa* (producer's)	$ckp_i = (pa_i - p_{i+})/(1 - p_{i+}) = (p_{ii} - p_{i+}p_{+i})/(p_{+i} - p_{i+}p_{+i})$	Rosenfield and Fitzpatrick-Lins (1986)
9.	Modified conditional *kappa* (user's)	$mcku_i = (ua_i - 1/m)/(1 - 1/m)$	Stehman (1997)
10.	Modified conditional *kappa* (producer's)	$mckp_i = (pa_i - 1/m)/(1 - 1/m)$	Stehman (1997)
11.	Category-level normalized accuracy	$cnma_i$ (Normalizing error matrix by forcing the marginal total of individual category to 1)	Congalton (1991)
12.	Ground truth index	$GT_i = (p_{ii} - R_i)/(1 - R_i)$ where R_i is lucky guesses (chance agreement), which can be calculated using Türk's (1979) algorithm.	Türk (1979), Rosenfield and Fitzpatrick-Lins (1986)
13.	Relative change of entropy given a category on map	$ecnu_i = (H(A) - H(A\|b_i))/H(A)$ where, $H(A) = -\sum_{j=1}^{m} p_{+j}\log(p_{+j}), H(A\|b_i) = -\sum_{j=1}^{m} \frac{p_{ij}}{p_{i+}}\log\left(\frac{p_{ij}}{p_{i+}}\right)$	Finn (1993)
14.	Relative change of entropy given a category on ground truthing	$ecnp_j = (H(B) - H(B\|a_j))/H(B)$ where, $H(B) = -\sum_{i=1}^{m} p_{i+}\log(p_{i+}), H(B\|a_j) = -\sum_{i=1}^{m} \frac{p_{ij}}{p_{+j}}\log\left(\frac{p_{ij}}{p_{+j}}\right)$	Finn (1993)

Accuracy parameters

Accuracy parameters (category-level)

	Agr	U	I	A	Total
Agr	65	4	22	24	115
U	6	81	5	8	100
I	0	11	85	19	115
A	4	7	3	90	104
Total	75	103	115	141	434

Producer's accuracy

Easy to calculate

Easy to interpret

Intuitive

- It is based on a binomial distribution in each column, $x_{jj} \rightarrow B(n_{+j}, \pi_{jj}), j = 1, \dots, k$

  Estimate: $\hat{\pi}_{jj} = \frac{x_{jj}}{n_{+j}}$ -> proportion of sample units correctly classified within category j. Example Agr-> 65/75=86.666 -> 87%
- For inference purpose it is used that the Binomial distribution is approximated by a Normal distribution

  IC (1-α)x100% → $\left(\hat{\pi}_{jj} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}_{jj}(1-\hat{\pi}_{jj})}{n_{+j}}}, \hat{\pi}_{jj} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}_{jj}(1-\hat{\pi}_{jj})}{n_{+j}}}\right)$, $Z_{1-\alpha/2}$ is the upper pertentile 1-α/2 of N(0,1).

Accuracy parameters

Accuracy parameters (category-level)

	Agr	U	I	A	Total
Agr	65	4	22	24	115
U	6	81	5	8	100
I	0	11	85	19	115
A	4	7	3	90	104
Total	75	103	115	141	434

User's accuracy

Easy to calculate

Easy to interpret

Intuitive

- It is based on a binomial distribution in each row, $x_{ii} \rightarrow B(n_{i+}, \pi_{ii}), i = 1, \ldots, k$

  Estimate: $\hat{\pi}_{ii} = \frac{x_{ii}}{n_{i+}}$ -> proportion of sample units in category i correctly classified

  Example Agr 65/115=0.5652 -> 57%

- For inference purpose it is used that the Binomial distribution is approximated by a Normal distribution

  IC (1-α)x100% → $\left(\hat{\pi}_{ii} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}_{ii}(1-\hat{\pi}_{ii})}{n_{i+}}}, \hat{\pi}_{jj} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}_{ii}(1-\hat{\pi}_{ii})}{n_{i+}}}\right), Z_{1-\alpha/2}$ is the upper pertentile 1-α/2 of N(0,1).

Accuracy parameters

Accuracy parameters (category-level)

User's and producer's accuracy

	Agr	U	I	A	Total
Agr	65	4	22	24	115
U	6	81	5	8	100
I	0	11	85	19	115
A	4	7	3	90	104
Total	75	103	115	141	434

Example

Class	Producer's Ac.		CI 95%
Agr	65/75	0.867	(0.790, 0.944)
U	81/103	0.786	(0.707, 0.866)
I	85/115	0.739	(0.659, 0.819)
A	90/141	0.638	(0.559, 0.718)

Class	User's Ac.		CI 95%
Agr	65/115	0.565	(0,475, 0.656)
U	81/100	0.810	(0.733, 0.887)
I	85/115	0.739	(0.659, 0.819)
A	90/104	0.865	(0.80, 0.932)

Accuracy parameters

Accuracy parameters (global-level)

Table 3
Map-level accuracy measures (derived from the error matrix in Table 1)

No.	Name	Formula	References
1	Overall accuracy	$oa = \sum_{i=1}^{m} p_{ii}$	Story and Congalton (1986)
2	Average accuracy from user's perspective	$aau = \frac{1}{m}\sum_{i=1}^{m} ua_i = \frac{1}{m}\sum_{i=1}^{m} \frac{p_{ii}}{p_{i+}}$	Fung and LeDrew (1988)
3	Average accuracy from producer's perspective	$aap = \frac{1}{m}\sum_{i=1}^{m} pa_i = \frac{1}{m}\sum_{i=1}^{m} \frac{p_{ii}}{p_{+i}}$	Fung and LeDrew (1988)
4	Double average of user's and producer's accuracy	$daup = (aau + aap)/2 = \frac{1}{m}\sum_{i=1}^{m} \frac{ua_i + pa_i}{2}$	
5	Classification success index	$CSI = aau + aap - 1 = \frac{1}{m}\sum_{i=1}^{m} (ua_i + pa_i) - 1$	Koukoulas and Blackburn (2001)
6	Average of Hellden's mean accuracy index	$amah = \frac{1}{m}\sum_{i=1}^{m} mah_i = \frac{1}{m}\sum_{i=1}^{m} \frac{2p_{ii}}{p_{i+} + p_{+i}}$	
7	Average of Short's mapping accuracy index	$amas = \frac{1}{m}\sum_{i=1}^{m} mas_i = \frac{1}{m}\sum_{i=1}^{m} \frac{p_{ii}}{p_{i+} + p_{+i} - p_{ii}}$	
8	Combined accuracy from user's perspective	$cau = (oa + aau)/2$	Fung and LeDrew (1988)
9	Combined accuracy from producer's perspective	$cap = (oa + aap)/2$	Fung and LeDrew (1988)
10	Combined accuracy from both user's and producer's perspectives	$caup = (oa + amah)/2$	
11	*kappa*	$kappa = (oa - ea)/(1 - ea)$ where $ea = \sum_{i=1}^{m} p_{i+}p_{+i}$	Cohen (1960), Rosenfield and Fitzpatrick-Lins (1986)
12	Modified *kappa*	$mkp = (oa - 1/m)/(1 - 1/m)$	
13	*kappa*-like statistic *alfa*	$alfa = (oa - ca)/(1 - ca)$ with $p_r(i) = n_{i+}/(N(1 - alfa + alfa \cdot p_c(i)/ca))$ and $p_c(i) = n_{+i}/(N(1 - alfa + alfa \cdot p_r(i)/ca))$ where N is the total number of cases. The statistic *alfa* is estimated by iteration. *Kappa* and observed marginal totals are used as initial estimates for *alfa*, $p_c(i)$ and $p_r(i)$. *ca* is a parameter estimated in the iteration.	Aickin (1990), Foody (1992)
14	Map-level normalized accuracy	*mnma* (normalizing error matrix by forcing the marginal total of individual category to 1 and dividing the resulted diagonal total by *m*)	Congalton (1991)

Accuracy parameters

Accuracy parameters (global-level)

15	Average mutual information	$ami = \sum_{i,j=1}^{m} p_{ij} \log\left(\frac{p_{ij}}{p_{i+}p_{+j}}\right)$	Finn (1993)
16	Normalized mutual information using the entropy on map	$nmiu = ami/H(B)$ where, $H(B) = -\sum_{i=1}^{m} p_{i+} \log(p_{i+})$	Finn (1993)
17	Normalized mutual information using the entropy on ground truthing	$nmip = ami/H(A)$ where, $H(A) = -\sum_{j=1}^{m} p_{+j} \log(p_{+j})$	Finn (1993)
18	Normalized mutual information using the arithmetic mean of the entropies on map and on ground truthing	$nmiam = 2\ ami/(H(A)+H(B))$ where, $H(A)$ and $H(B)$ are as above.	Strehl and Ghosh (2002)
19	Normalized mutual information using the geometric mean of the entropies on map and on ground truthing	$nmigm = ami/\sqrt{H(A)H(B)}$ where, $H(A)$ and $H(B)$ are as above.	Ghosh et al. (2002)
20	Normalized mutual information using the arithmetic mean of the maximum entropies on map and on ground truthing	$nmimx = 2\ ami/(\max(H(A)) + \max(H(B))) = ami/\log m$	Strehl (2002)

Accuracy parameters

Accuracy parameters (global-level)

OA: Overall accuracy (% of agreement).

Easy to calculate and intuitive.
Only uses the elements of the main diagonal.
It can be considered as the probability of being well classified, so it can be assumed distributed according to a binomial distribution.

When N is large, it can be considered that it is distributed according to a normal.

$$OA = \frac{1}{N}\sum_{i=1}^{k} n_{ii} = \sum_{i=1}^{k} p_{ii}$$

Example OA:

	Agr	U	I	A	Total
Agr	65	4	22	24	115
U	6	81	5	8	100
I	0	11	85	19	115
A	4	7	3	90	104
Total	75	103	115	141	434

$$\textstyle\sum_{i=1}^{M} n_{i,i} \rightarrow B(N,\pi), \quad \pi = \sum_{i=1}^{M} \pi_{ii}$$

OA = (65+81+85+90)/434 = 0.739

95 % confidence limits -> (0.6972, 0.7821)

QC of confusion matrices

Frames

- **Frame #1**:
    - Observed values (matrix, indices, etc.) are compared to fixed values established by the specifications.
    - Fixed values stablish the $\mathbb{H}_0$ hypothesis for the statistical test.
    - Hypothesized values stablish a set of specifications. Hypothesized values are considered error-free values (fixed reference).
    - E.g. A tolerance value is proposed for the OA or Kappa index, a complete matrix is proposed as $\mathbb{H}_0$ for a matrix-wise comparison.
- **Frame #2:**
    - Two observed cases (two matrices, two indices, etc.) are compared in-between.
    - There are no previous specifications.
    - Both cases sources have the same value (there is no fixed reference).

QC of confusion matrices (Frame 1)

Frame #1 → Comparison with a set of specifications

- In literature there are several existing options.
- In MEMCALIG project we have systematized these options and developed new ones.
- The appropriate option for each case depends on the OBJECTIVE and on the AMOUNT of available information.
- All conditions stablished for the use of confusion matrices in QC are considered here.

José Rodríguez-Avi

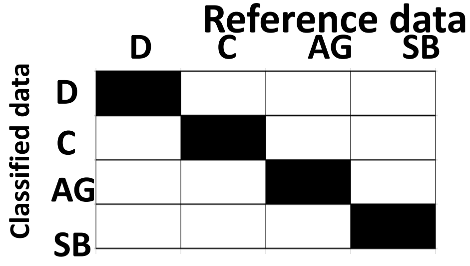
QC of confusion matrices (Frame 1)

MEMCALIG

Frame #1 → Comparison with a set of given specifications

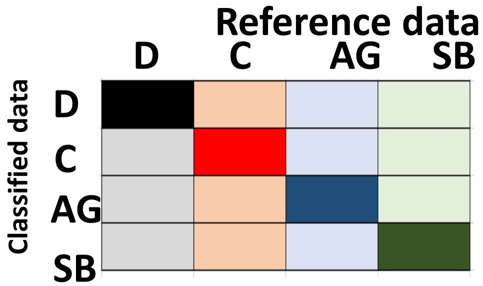
Case 1: An overall hypothesis/specification

- 1.1 One binomial



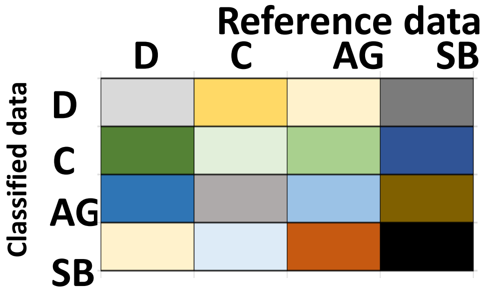
Case 2: One hypothesis/specification per class

- 2.1 k binomials
- 2.2 A chi-square overall test



Case 3: Complete set of hypothesis/specifications for each class

- 3.1 k chi-square tests
- 3.2 An overall chi square test
- 3.3 k exact multinomial tests

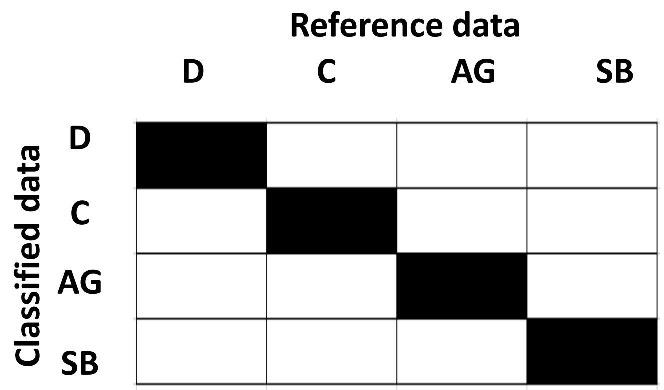


QC of confusion matrices (Frame 1)

Frame #1: Case 1.1 → A (single) binomial distribution test

- This is the OA case.
- The number of sample units correctly classified is distributed according to a binomial distribution with parameters N, and π^0 the proportion of well-classified cases,

$$T = \sum_{j=1}^{k} x_{jj} \underset{\mathbb{H}_0}{\rightarrow} B(N, \pi^0)$$



Pros:

- It is unilateral, so we reject only if the observed values are worse than the expected ones.
- It is very well known.
- It is very easy to implement.

Cons:

- It is a waste of information, because of the conversion of a matrix into one only number.
- The decision is highly influenced with the categories' size: A high number of agreements in the category that has the greatest size conceals severe discrepancies in categories with smaller number of elements.

QC of confusion matrices (Frame 1)

Frame #1: Case 1.1

Example 1

$$T = \sum_{i=1}^{k} x_{jj} \underset{\mathbb{H}_0}{\rightarrow} B(N, \pi^0)$$

Observed matrix

	Woodland	Grassland	Non vegetated	Water
Woodland	80	10	10	2
Grassland	15	36	15	5
Non vegetated	5	5	66	0
Water	0	3	5	83
Column marginal	100	54	96	90

Senseman et al. (1995)

$\mathbb{H}_0$: The whole proportion of well-classified elements is $\geq 80\%$ $(OA \geq 80\%)$
$\mathbb{H}_1$: The whole proportion of well-classified elements is $< 80\%$ $(OA < 80\%)$

From product specifications

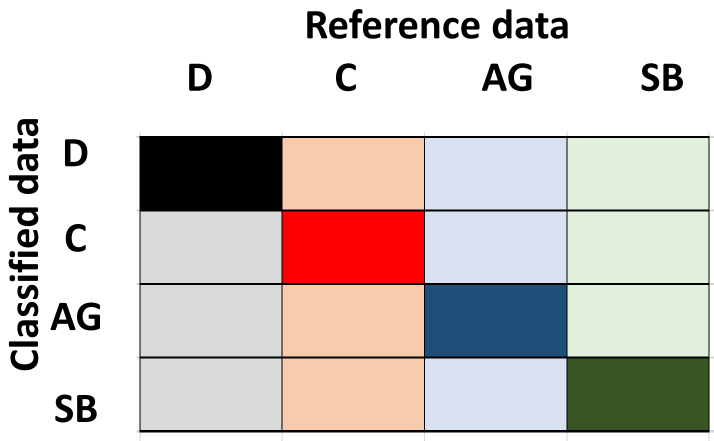
- Test statistic: $T = 80 + 36 + 66 + 83 = 265$
- Distribution under $\mathbb{H}_0$: $\mathrm{B}(340,\ 0.80)$
- $p = P[X \leq 265 | X \ \rightarrow \mathrm{B}(340,\ 0.80)] = 0.1882$

Do not reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

Frame #1: Case 2.1 → k binomial distributions test

- A binomial distribution in each column.
- k binomial contrasts are proposed.
- A set of k different proportion of well-classified elements, attending to the category peculiarities are tested.
- k p-values are obtained.
- Rejection under Bonferroni's rule.



Pros:

- It is unilateral, so we reject only if the observed values are worse than the expected ones.
- It takes into account the different sizes between columns, and all of them are equally important
- If we reject $\mathbb{H}_0$ we can know why it is rejected, analyzing by columns.
- Different assumed probabilities for each column are allowed.

Cons:

- It needs Bonferroni's correction because of the number of test involved, in order to control the overall type I error level.

QC of confusion matrices (Frame 1)

Frame #1: Case 2.1

Example 2

$$T_j = x_{jj} \underset{\mathbb{H}_0}{\rightarrow} B(n_{+j}, \pi^0)$$

Observed matrix

	Woodland	Grassland	Non vegetated	Water
Woodland	80	10	10	2
Grassland	15	36	15	5
Non vegetated	5	5	66	0
Water	0	3	5	83
Column marginal	100	54	96	90

Senseman et al. (1995)

$\mathbb{H}_0$: The whole proportion of well-classified elements is $\geq 80\%$ ***IN EACH CATEGORY***
$\mathbb{H}_1$: The whole proportion of well-classified elements is $< 80\%$ ***IN AT LEAST ONE CATEGORY***

From product specifications

- Test statistics: $\boldsymbol{T} = (80,\ 36,\ 66,\ 83)$; Bonferroni's alpha= $0.0125 = \alpha/4$
- Distributions under $\mathbb{H}_0$: $\mathrm{B}(100, 0.8)$; $\mathrm{B}(54, 0.8)$; $\mathrm{B}(96, 0.8)$; $\mathrm{B}(90, 0.8)$
- $\boldsymbol{p} = (0.5398;\ 0.0149;\ 0.0060;\ 0.9996)$
- At least one element in $\boldsymbol{p}$ is less than 0.0125. In consequence, we reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

Frame #1: Case 2.1

Example 3

$$T_j = x_{jj} \underset{\mathbb{H}_0}{\rightarrow} B(n_{+j}, \pi^0)$$

Observed matrix

	Woodland	Grassland	Non vegetated	Water
Woodland	80	10	10	2
Grassland	15	36	15	5
Non vegetated	5	5	66	0
Water	0	3	5	83
Column marginal	100	54	96	90

Senseman y col (1995)

$\mathbb{H}_0$: The proportion of well-classified elements is $\geq (80\%, 75\%, 70\%, 90\%)$ ***RESPECTIVELY***
$\mathbb{H}_1$: The proportion of well-classified elements is less than $(80\%, 75\%, 70\%, 90\%)$ ***respectively IN AT LEAST ONE CATEGORY***

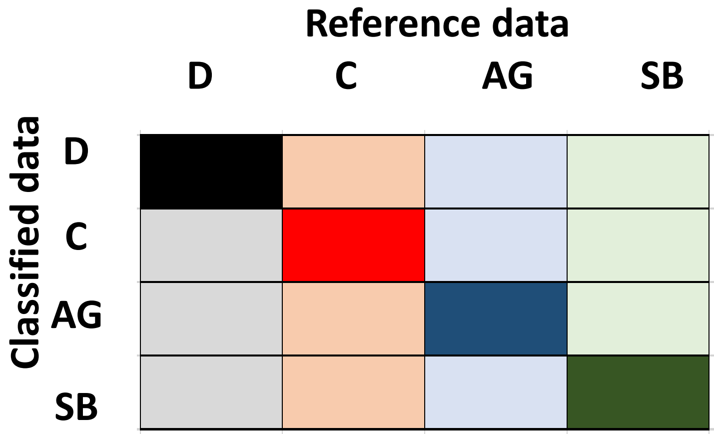
From product specifications

- Test statistics: $\boldsymbol{T} = (80,\ 36,\ 66,\ 83)$; Bonferroni's alpha= 0,0125
- Distributions under $\mathbb{H}_0$: $\mathrm{B}(100, 0.8); \mathrm{B}(54, 0.75); \mathrm{B}(96, 0.7); \mathrm{B}(90, 0.9)$
- $\boldsymbol{p} = (0.5398; 0.1065; 0.4323; 0.8075)$
- All elements in $\boldsymbol{p}$ are greater than 0.0125. In consequence, we cannot reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

Frame #1: Case 2.2 → Chi-square (overall) test

- A χ^2 overall test based on the sum of squares of normal approximations for each binomial.
- It is a bilateral test.
- The distribution is a chi-square with k degrees of freedom.



Pros:

- It is an overall test. In consequence the p-value is directly compared with α.
- It takes into account the different sample sizes between columns, and all of them are equally important.
- If we reject $\mathbb{H}_0$ we can know why it is rejected, analyzing by columns .
- Different assumed probabilities for each column are allowed.

Cons:

- It is a bilateral test. In consequence, $\mathbb{H}_0$ will be rejected if the observed values are worse or better than those specified by $\mathbb{H}_0$.

QC of confusion matrices (Frame 1)

Frame #1: Case 2.2

Example 4

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(x_{jj} - n_{+j} * \pi_j^0\right)^2}{n_{+j} * \pi_j^0 * (1 - \pi_j^0)} \xrightarrow[\mathbb{H}_0]{} \chi_k^2$$

Observed matrix

	Woodland	Grassland	Non vegetated	Water
Woodland	80	10	10	2
Grassland	15	36	15	5
Non vegetated	5	5	66	0
Water	0	3	5	83
Column marginal	100	54	96	90

Senseman et al. (1995)

$\mathbb{H}_0$: The whole proportion of well-classified elements is equal to 80% ***IN EACH CATEGORY***
$\mathbb{H}_1$: The whole proportion of well-classified elements is $\neq$ 80% ***IN AT LEAST ONE CATEGORY***

From product specifications

Test statistics

$$Z_1 = \frac{80-100*0.8}{\sqrt{100*0.8*0.2}};\ Z_2 = \frac{36-54*0.8}{\sqrt{54*0.8*0.2}};\ Z_3 = \frac{66-96*0.8}{\sqrt{96*0.8*0.2}};\ Z_4 = \frac{83-90*0.8}{\sqrt{90*0.8*0.2}}$$

$$\chi^2 = \sum_{i=1}^{4} Z_i^2 = 21.9965 \hookrightarrow \chi_4^2$$

QC of confusion matrices (Frame 1)

Frame #1: Case 2.2

Example 4

- $\boldsymbol{p} = \mathrm{P}[\chi_4^2 > 21.9965] = 0.0002 < 0.05$

- In consequence, we reject $\mathbb{H}_0$.

Once $\mathbb{H}_0$ is rejected, we analyze each column separately

Test statistics

$Z_1 = 0.0000$; $Z_2 = -2.4495$; $Z_3 = -2.7557$; $Z_4 = 2.8988$

$Z_1^2 = 0.0000$; $\boxed{Z_2^2 = \mathbf{6.0000};}$ Worse; $\boxed{Z_3^2 = \mathbf{7.5930};}$ Worse; $\boxed{Z_4^2 = 8.4028}$ Better

QC of confusion matrices (Frame 1)

Frame #1: Case 2.2

Example 5

$\mathbb{H}_0$: The whole proportion of well-classified elements is equal to $(80\%, 75\%, 70\%, 90\%)$ ***RESPECTIVELY***
$\mathbb{H}_1$: The whole proportion of well-classified elements is different from $(80\%, 75\%, 70\%, 90\%)$ ***respectively IN AT LEAST ONE CATEGORY***

From product specifications

Test statistics

$$Z_1 = \frac{80-100*0.8}{\sqrt{100*0.8*0.2}}; Z_2 = \frac{36-54*0.75}{\sqrt{54*0.75*0.25}}; \quad Z_3 = \frac{66-96*0.7}{\sqrt{96*0.7*0.3}}; \quad Z_4 = \frac{83-90*0.9}{\sqrt{90*0.9*0.1}}$$

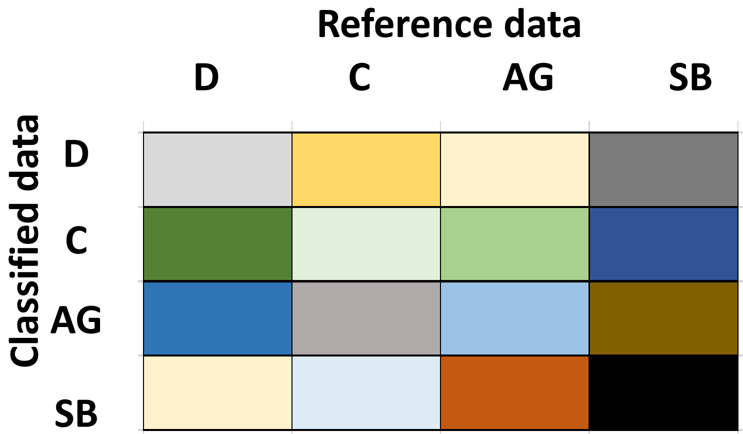
$$\chi^2 = \sum_{i=1}^{4} Z_i^2 = 2.5652 \hookrightarrow \chi_4^2$$

- $\boldsymbol{p} = \mathrm{P}[\chi_4^2 > 2.5652] = 0.6330 > 0.05$
- $\boldsymbol{p}$ is greater than 0.05. In consequence, we cannot reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

Frame #1: Case 3.1 → Multiple goodness of fit test

- We have more information about error specifications.
- Comparing each column against a (assumed) table of probabilities for each column.
- k Chi-square distribution with $(k_j - 1)$ degrees of freedom each.
- Decision under Bonferroni's criteria



Pros:

- A very detailed specifications about errors may be provides.
- It takes into account different sizes and/or specifications between columns, and all of them are equally important.
- If we reject $\mathbb{H}_0$ we can know why it is rejected, analyzing by columns.
- It is flexible: The number of specifications may vary between columns.

Cons:

- It is a bilateral test: $\mathbb{H}_0$ will be rejected if the observed values are worse or better than the specified by $\mathbb{H}_0$.
- It needs Bonferroni's correction because of the number of test involved, in order to control the overall type I error level.

QC of confusion matrices (Frame 1)

Frame #1: Case 3.1

Example 6

Woodland

- *At least 80% of classification accuracy* ($\geq 95\%$)
- *Can be somewhat confused with grassland, but not more than 14%* ($\leq 14\%$)
- *Cannot be confused with non-vegetated (≤3%) or with water (≤3%)*

Grassland

- *Al least 75% of classification accuracy* ($\geq 75\%$)
- *Can be somewhat confused with Woodland but not more than 15%*
- *Can be somewhat confused with non-vegetated but not more than (≤5%)*
- *Can be somewhat confused with water, but not more than ≤5%*

Non-vegetated

- *Al least 70% of classification accuracy* ($\geq 70\%$)
- *Can be somewhat confused with Woodland but not more than 15%*
- *Can be somewhat confused with grassland but not more than 10%*
- *Cannot be confused with water (at most, 5% of misclassified points)*

Water

- *Must be well classified (at least, 90%)*
- *Cannot be confused with Non vegetated (at most, 5% of misclassified points)*
- *Cannot be confused with Grassland (at most, 3% of misclassified points)*
- *Cannot be confused with Woodland (at most, 2% of misclassified points)*

From product specifications

QC of confusion matrices (Frame 1)

MEMCALIG

Frame #1: Case 3.1

Example 6

The product specifications stablished in $\mathbb{H}_0$ can be written in the form of a matrix → MH_o

$\mathbb{H}_0$: The proportion of elements in each cell in the model is given by MH_o

	Woodland	Grassland	Non vegetated	Water
Woodland	0.80	0.15	0.15	0.02
Grassland	0.14	0.75	0.10	0.03
Non vegetated	0.03	0.05	0.70	0.05
Water	0.03	0.05	0.05	0.90

MH_0
From product specifications

Observed matrix A

	W	G	NV	Wa
W	80	10	10	2
G	15	36	15	5
NV	5	5	66	0
Wa	0	3	5	83
T	100	54	96	90

Test statistics

$$R_{ij} = \frac{A[i,j] - MH_0[i,j] * n_{+j}}{\sqrt{MH_0[i,j] * n_{+j}}} \qquad \chi_j = \sum_i R_{ij}^2 \underset{\mathbb{H}_0}{\rightarrow} \chi_k^2$$

$$\chi = (4.4047, 2.9382, 4.4117, 6.5308)$$

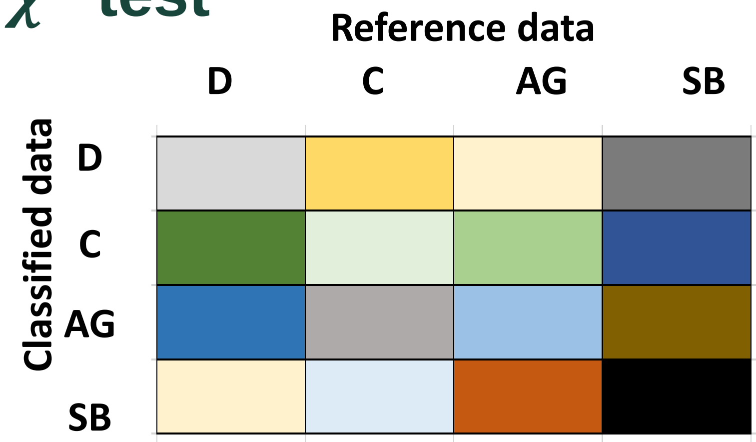
$$\boldsymbol{p} = (0.2209, 0.4012, 0.2203, 0.088)$$

$\boldsymbol{p}$ is greater than 0.05. In consequence we cannot reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

Frame #1: Case 3.2 → A global χ^2 test

- Same operational framework as Case 4th.
- Comparing each column against a (assumed) table of probabilities (specifications).
- Chi-square distribution with $k(k_j - 1)$ degrees of freedom.



Pros:

- It is an overall test. In consequence the p-value is directly compared with α.
- It takes into account different sizes and/or specifications between columns, and all of them are equally important.
- If we reject $\mathbb{H}_0$ we can know why it is rejected, analyzing by columns.
- It is flexible: The number of specifications may vary between columns.

Cons:

- It is a bilateral test. In consequence, $\mathbb{H}_0$ will be rejected if the observed values are worse or better than the specified by $\mathbb{H}_0$.

QC of confusion matrices (Frame 1)

MEMCALIG

Frame #1: Case 3.2

Example 6

The product specifications stablished in $\mathbb{H}_0$ can be written in the form of a matrix → MH_o

$\mathbb{H}_0$: The proportion of elements in each cell in the model is given by MH_o:

	Woodland	Grassland	Non vegetated	Water
Woodland	0.80	0.15	0.15	0.02
Grassland	0.14	0.75	0.10	0.03
Non vegetated	0.03	0.05	0.70	0.05
Water	0.03	0.05	0.05	0.90

MH_0
From product specifications

Observed matrix

	W	G	NV	Wa
W	80	10	10	2
G	15	36	15	5
NV	5	5	66	0
Wa	0	3	5	83
T	100	54	96	90

Tests statistics

$$R_{ij} = \frac{A[i,j] - MH_0[i,j] * n_{+j}}{\sqrt{MH_0[i,j] * n_{+j}}} \qquad \chi^2 = \sum_{ij} R_{ij}^2 \hookrightarrow \chi^2_{12}$$

$$\chi^2 = 18.2856 \qquad \boldsymbol{p} = 0.1073$$

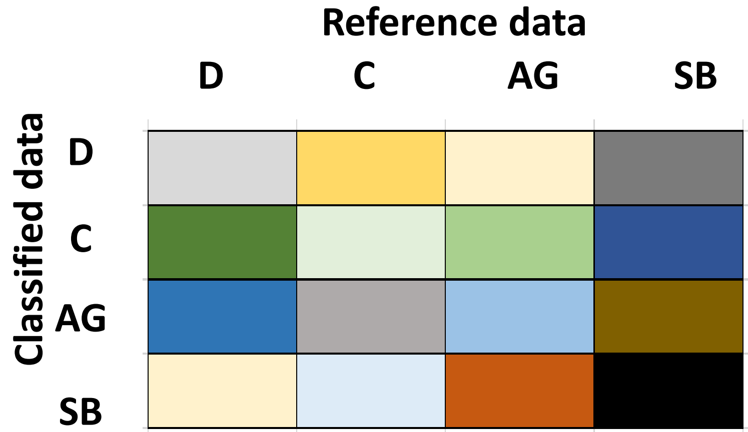
$\boldsymbol{p}$ is greater than 0.05. In consequence we cannot reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

MEMCALIG

Frame #1: Case 3.3 → k (exact) multinomial tests

- Same operational framework as Cases 4th and 5th
- Assuming a multinomial distribution with parameters n_{+j} and π_j^0. Columns are arranged in decreasing order of probabilities (from the most to the less)
- k exact multinomial tests are performed.
- Decision is taken using Bonferroni's criteria.



Pros:

- It is an exact and unilateral test, so the p-value is exactly calculated and we reject only if the observed values are worse than the expected ones.
- It takes into account different sizes and/or specifications between columns, and all of them are equally important.
- If we reject $\mathbb{H}_0$ we can know why it is rejected, analyzing by columns.
- It is flexible: The number of specifications may vary between columns.

Cons:

- It needs Bonferroni's correction because of the number of test involved, in order to control the overall type I error level.
- Calculations may be more complicated than the precedent ones.

QC of confusion matrices (Frame 1)

Frame #1: Case 3.3

- A reordering of cases is needed in order to express our preferences about confusion errors between classes.
- This order has to be in decreasing order (C is moss preferred than E1, this is most preferred than E2 and so on).
- An exact multinomial test is carried out for each column. All p-values are compared against α/k and the null hypothesis will be rejected if at least one p value is less than α/k.

Reference data

Classified data	D	C	AG	SB
D	C	E1	E2	E3
C	E1	C	E1	E2
AG	E2	E2	C	E1
SB	E3	E3	E3	C

D	C	AG	SB
C	C	C	C
E1	E1	E1	E1
E2	E2	E2	E2
E3	E3	E3	E3

QC of confusion matrices (Frame 1)

MEMCALIG

Frame #1: Case 3.3

Example 7

The product specifications stablished in $\mathbb{H}_0$ can be written in the form of a matrix → MH_o

$\mathbb{H}_0$: The proportion of elements in each cell in the model is given by MH_o:

Observed matrix (reordered by column)

	W	G	NV	Wa
C	80	36	66	83
E1	15	10	10	0
E2	5	5	15	5
E3	0	3	5	2
Total	*100*	*54*	*96*	*90*

First statistic

	W	G	NV	Wa
C	0.80	0.75	0.70	0.90
E1	0.14	0.15	0.15	0.05
E2	0.03	0.05	0.10	0.03
E3	0.03	0.05	0.05	0.02

MHo
From product specifications (considering reordering)

First multinomial parameters

Observed matrix

	W	G	NV	Wa
W	80	10	10	2
G	15	36	15	5
NV	5	5	66	0
Wa	0	3	5	83
T	100	54	96	90

$$p = (0.5162, 0.7279, 0.1268, 0.6953)$$

$\boldsymbol{p}$ is greater than 0.05. In consequence, we cannot reject $\mathbb{H}_0$

QC of confusion matrices (Frame 1)

Frame #1: Summary

Information	*Test*	*Remarks*
We are interested only in the global proportion of well-classified points without differences between categories	One binomial	- Is unilateral
We are interested in the proportion of well-classified points with different probabilities for each category	k-binomials	- Is unilateral - Bonferroni's correction is required - If $\mathbb{H}_0$ is rejected we can search for categories that not verify the specifications
	Overall chi-square	- Is bilateral. In consequence $\mathbb{H}_0$ may also be rejected if proportions are better than specifications - We obtain a single p-value - If $\mathbb{H}_0$ is rejected we can search for categories that not fits well with $\mathbb{H}_0$
We are interested not only in the proportion of well-classified points in each category but also with probability error levels for different categories	Multiple goodness-of fit test	- Is bilateral. In consequence $\mathbb{H}_0$ may also be rejected if proportions are better than specifications - Bonferroni's correction is required - If $\mathbb{H}_0$ is rejected we can search for categories that not fits well with $\mathbb{H}_0$ - Flexibility in the number of error probabilities fixed, that may be different between columns
	An overall chi-square test	- Is bilateral - We obtain a single p-value - If $\mathbb{H}_0$ is rejected, we can analyze the rejection causes - Flexibility in the number of categories in each column
	k exact multinomial tests	- Is an exact test. - Is unilateral - Bonferroni's correction is required - If H_0 is rejected we can search for categories that not agree with $\mathbb{H}_0$ - Flexibility in the number of error probabilities fixed, that may be different between columns

Frame #1: User's guideline

- The classical binomial test is useful when we are interested in the global rate of well-classified elements, but even in this situation presents several problems.
- In this case, we recommend start with the overall chi square and, if the null hypothesis is rejected, to continue with the k-binomial test. This is useful even with different probabilities between columns.
- If we stablish an entire set of conditions, we recommend start wit the overall chi square and, if the null hypothesis is rejected, to continue with the multiple goodness of fit tests.
- A unilateral alternative in this case is to apply the exact multinomial test.
- The number of categories can vary between columns.

QC of confusion matrices (Frame 2)

MEMCALIG

Frame #2 → Comparison between observed cases

- Two observed confusion matrices are given.
- Here, the set of specifications previously mentioned in Frame #1 are not needed.
- Classical way to address this problem -> comparison between OA (or kappa).
- Drawbacks of classical approach:

  i) use partial information,

  ii) valid for large sample sizes (based on approximation to standard normal distribution).
- In MEMCALIG project we have developed new approaches taking advantages of underlying multinomial distribution.

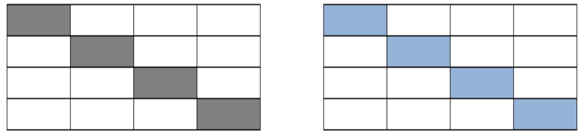
QC of confusion matrices (Frame 2)

Frame #2 → Case 1st

Example: Suppose π_1, π_2 are the OA values of two observed confusion matrices with total sample units *n* and *m*, respectively.

We can state the null hypothesis $\mathbb{H}_0 : \pi_1 - \pi_2 = 0$

$\mathbb{H}_1 : \pi_1 - \pi_2 \neq 0$



$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{m}}} \underset{\mathbb{H}_0}{\longrightarrow} N(0,1)$$

So, $\mathbb{H}_0$ is rejected if $|Z_{obs}| \geq Z_{\alpha/2}$.

In Example 8, $\hat{\pi}_1 = \frac{321}{434} = 0.7396, \hat{\pi}_2 = \frac{246}{336} = 0.7321 \rightarrow Z = 0.2339 < Z_{0.025} = 1.96$

So, we cannot reject $\mathbb{H}_0$.

Kappa analysis is similar¡¡¡

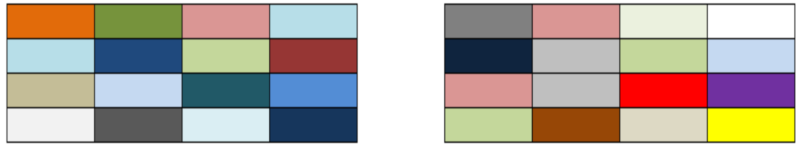
QC of confusion matrices (Frame 2)

MEMCALIG

Frame #2: Case 2nd → Comparison by multinomial law

OUR PROPOSAL: Compare two confusion matrices by means of a discrepancy measure between multinomials

Why? Use the complete information of each matrix



		j=column (Reference Data)			
		1	2	…	k
i=row Classified Data	1	$p_{11} = \frac{n_{11}}{n}$	$p_{12} = \frac{n_{12}}{n}$	…	$p_{1k} = \frac{n_{1k}}{n}$
	2	$p_{21} = \frac{n_{21}}{n}$	$p_{22} = \frac{n_{22}}{n}$	…	$p_{1k} = \frac{n_{1k}}{n}$
	…	…	…	…	
	k	$p_{k1} = \frac{n_{k1}}{n}$	$p_{k2} = \frac{n_{k2}}{n}$	…	$p_{kk} = \frac{n_{kk}}{n}$

First confusion matrix observed
$M(n, p_{11}, \dots, p_{1k}, \dots, p_{k1}, \dots, p_{kk})$

⇓

$M(n, p_1, p_2, \dots, p_M)$

Similar?

$M(m, q_1, q_2, \dots, q_M)$

⇑

Second confusion matrix observed
$M(m, q_{11}, \dots, q_{1k}, \dots, q_{k1}, \dots, q_{kk})$

QC of confusion matrices (Frame 2)

MEMCALIG

Frame #2 → Case 2nd

Let us consider two confusion matrices as two random vectors, X and Y, whose values are grouped into *M=k* x *k* classes with population probabilities, $\mathrm{P} = (P_1, P_2, \ldots, P_M)$ and $\mathrm{Q} = (Q_1, Q_2, \ldots, Q_M)$, respectively.

Confusion matrix produced using Landsat Thematic Mapper Imaginary by Analyst #1

		Reference Data			
		Deciduous	Conifer	Agriculture	Shrub
Classfied Data	Deciduous	65	4	22	24
	Conifer	6	81	5	8
	Agriculture	0	11	85	19
	Shrub	4	7	3	90

Congalton and Green (2009)

Confusion matrix produced using Landsat Thematic Mapper Imaginary by Analyst #2

		Reference Data			
		Deciduous	Conifer	Agriculture	Shrub
Classfied Data	Deciduous	45	4	12	24
	Conifer	6	91	5	8
	Agriculture	0	8	55	9
	Shrub	4	7	3	55

Congalton and Green (2009)

Both analysts reach the same conclusion? ⟹ Test for the equality of both multinomials

QC of confusion matrices (Frame 2)

Frame #2: Case 2nd

Let X and Y be two multinomial distribution with probabilities P and Q, respectively.
Two classification works or two analysts reach the same conclusion if P and Q are “quite” similar.
So our aim is to test

$$\mathbb{H}_0 : P=Q.$$

To measure the nearness between P and Q, let us consider the following discrepancy measure between multinomials

$$D(P,Q) = \sum_{i=1}^{M}\left(\sqrt{P_i} - \sqrt{Q_i}\right)^2, \; D(P,Q) \geq 0, \; D(P,Q) = 0 \; iff \; P = Q.$$

When $\mathbb{H}_0$ should be rejected?

Let $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_m)$ be two independent random samples from X and Y, with sizes *n* and *m*, and p_i and q_i , relative frequencies observed. We use the test statistic

$$T_{n,m} = 4(n+m)\sum_{i=1}^{M}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2 .$$

If $\mathbb{H}_0$ is true -> $T_{n.m} \approx 0$, so, we reject the null hypothesis for “large” values of $T_{n.m.}$

QC of confusion matrices (Frame 2)

Frame #2: Case 2nd

But, what are large values:

- Asymptotic null distribution of $T_{n,m}$ is a χ^2_{M-1}.
- Behavior rather poor of such approximation.
- To overcome this problem, we approximate the null distribution by bootstrapping.
- How bootstrapping works?
  - ✓ Generating a large number of samples under the null hypothesis and calculating the value of the test statistic $T_{n,m}$ from them, we are able to approximate the probability distribution of $T_{n,m}$.
- We reject $\mathbb{H}_0$ if the bootstrap p-value < α.

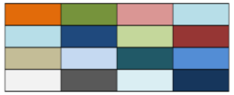
QC of confusion matrices (Frame 2)

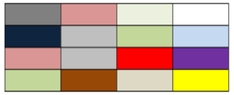
MEMCALIG

Frame #2: Case 2nd

Matrix 1 (P)

Matrix 2 (Q)





Step 1: Calculate $\hat{p}$, $\hat{q}$

Step 2: Calculate T_{obs}

Step 3: $\hat{p}$ $\hat{q}$

Pooled data ← Under $\mathbb{H}_0$

Replication 1

Replication 2

Replication 1000

$\hat{p}_1^*, \hat{q}_1^*$
T_1^*

$\hat{p}_2^*, \hat{q}_2^*$
T_2^*

....

$\hat{p}_{1000}^*, \hat{q}_{1000}^*$
T_{1000}^*

$$\hat{p} = \frac{card(T^* \geq T_{obs})}{1000}$$

QC of confusion matrices (Frame 2)

MEMCALIG

Frame #2: Case 2nd

Example 9

Confusion matrix produced using Landsat Thematic Mapper Imaginery by Analyst #1

		Reference Data			
		Deciduous	Conifer	Agriculture	Shrub
Classfied Data	Deciduous	65	4	22	24
	Conifer	6	81	5	8
	Agriculture	0	11	85	19
	Shrub	4	7	3	90

Congalton and Green (2009)

Confusion matrix produced using Landsat Thematic Mapper Imaginery by Analyst #2

		Reference Data			
		Deciduous	Conifer	Agriculture	Shrub
Classfied Data	Deciduous	45	4	12	24
	Conifer	6	91	5	8
	Agriculture	0	8	55	9
	Shrub	4	7	3	55

Congalton and Green (2009)

In this case,

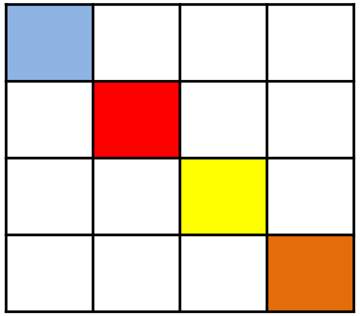
$T_{n,m,obs} = 4(434 + 336)\sum_{i=1}^{16}\left(\sqrt{\hat{p}_i} - \sqrt{\hat{q}_i}\right)^2 = 56.38618$ → p-value=0.5294

p>0.05, the two error matrices reach the same conclusion .

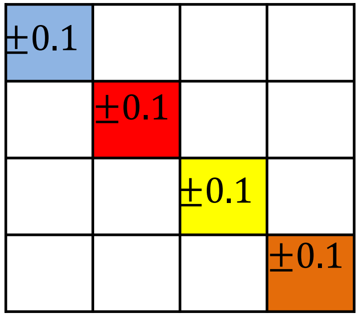
Other advances on confusion matrices

MEMCALIG

Frame #2: Case 3rd → Comparison of small discrepancies



versus



Conditions:

- Sum of variations in the diagonal is 0.
- The off-diagonal cells are collapsed into one

How effective this technique is?

⇓

Power study

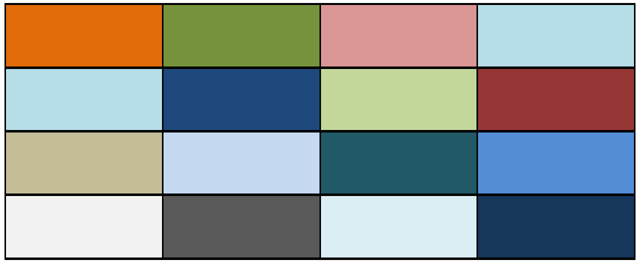
Virtudes Alba-Fernández

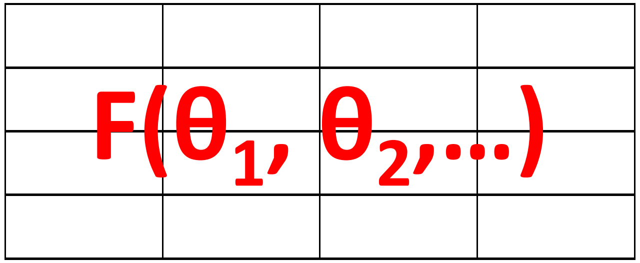
Other advances on confusion matrices

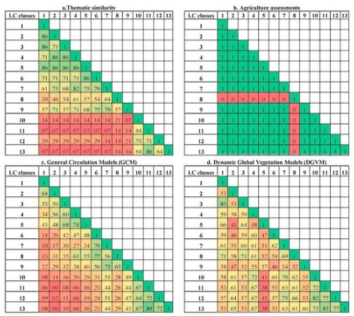
MEMCALIG

Parametric family

- We are investigating the possibility of parametrically modeling the contents of confusion matrices of the same subject and similar conditions.
- The advantage of this modeling would be to have a parametric model from which all the properties of interest on a matrix could be derived.





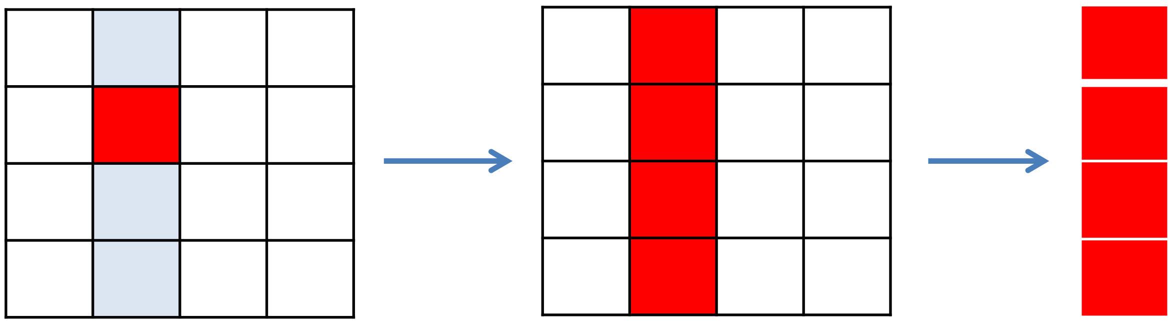


Other advances on confusion matrices

MEMCALIG

Parametric family

Coming back to the perspectives of accuracy parameters derived from a confusion matrix: per class



$$M\left(n_{+j}, p_{1j}, p_{2j}, \dots, p_{kj}\right)$$
$$p_{ij} = \frac{n_{ij}}{n_{+j}}, i = 1, \dots, k$$

$$M\left(n_{+j}, p_{1j}, p_{2j}, \dots, p_{kj}\right)$$

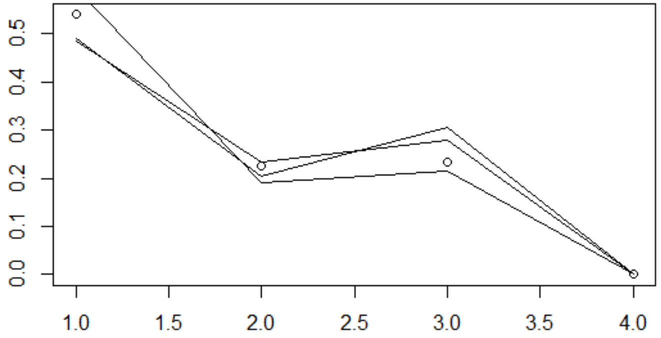
with

$$p_{1j} = p_1(\theta),$$
$$p_{2j} = p_2(\theta),$$
$$\dots$$
$$p_{kj} = p_k(\theta),$$
$$\theta \in \Theta, \Theta \subset \mathbb{R}^s,$$
$$k - s - 1 > 0$$

Non parametric framework ⟶ Parametric framework: a Goodness-of-fit test is needed

Other advances on confusion matrices

Parametric family



Pros:

- Unified and flexible understanding of whole classification work (per class).
- For each cell: more accurate confidence interval that one provided by the relative frequencies.
- One parametric family is validated -> become truth (reference model) for new classification works.

Cons:

- Find the appropriate parametric family “ad doc” because dimension.

Other advances on confusion matrices

Bias detection

- Hypothesis. A confusion matrix shall be symmetric because the chance of confusion for the case A→B is the same as for the case B→A.

- Objective. Detect bias or significative systematic differences between off-diagonal elements of a confusion matrix.

- When. Changes in land use/cover.

- How. A marginal homogeneity tests is need.

- The null hypothesis is the marginal homogeneity, that is, the equality (or lack of significant difference) between the row marginal proportions, π_{i+}, and the corresponding column proportions, π_{+i}.

Other advances on confusion matrices

MEMCALIG

Bias detection

Table 1 Satellite data specifications.				
Data	Year of acquisition	Bands/color	Resolution (m)	Source
Landsat 5 TM imagery	1992	Multi-spectral	250	USGS glovis
SPOT imagery	2012	Multi-spectral	2.5	SUPARCO

Table 2 Classes delineated on the basis of supervised classification.

Sr. No.	Class name	Description
1	Agriculture	Crop fields and fallow lands
2	Settlements	Residential, commercial, industrial, transportation, roads, mixed urban
3	Bare soil/rock	Land areas of exposed soil and barren area influenced by human influence
4	Vegetation	Mixed forest lands
5	Water	River, open water, lakes, ponds and reservoirs

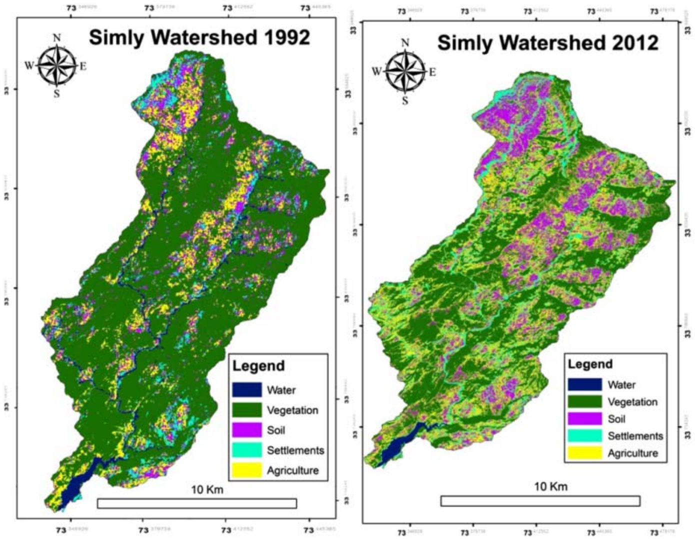


Figure 2 Classified maps of Simly watershed (1992 and 2012).

Table 4 Cross-tabulation of land cover classes between 1992 and 2012 (area in ha).

		1992				
		Agriculture	Bare soil/rocks	Settlements	Vegetation	Water
2012	Agriculture	560	$\boxed{562}$	378	3047	134
	Bare soil/rocks	$\boxed{748}$	622	278	1021	22
	Settlements	212	283	245	982	148
	Vegetation	253	180	133	6286	155
	Water	2	0.3	4	5	144

Our PROPOSAL: Homogeneity test

Other advances on confusion matrices

MEMCALIG

Forthcoming results

- Systematization of the existent and the new methodologies on thematic accuracy assessment (via confusion matrices) in a R package -> Tentative name CoMaAs (in progress).
- Develop a web service.
- Collect a set of confusion matrices in a database in order to facilitate new researchs.

Conclusions

- A general view of the confusion matrix was given.
- The main components of a thematic accuracy assessment has been highlighted.
- The more common and traditional quality control techniques based on confusion matrices has been presented.
- New approaches to thematic quality control have been proposed (at different levels and under different frames).
- Confusion matrices are a useful and complex tool.
- We have detected that the tools to work with them are poorly systematized and that there is a lot of research opportunity.

Malta, February 2018

2nd INTERNATIONAL WORKSHOP ON SPATIAL DATA QUALITY

Invited session on Thematic accuracy assessment by means of confusion matrices

Thank you

Francisco Javier Ariza-López[1] , José Rodríguez-Avi[1] Virtudes Alba-Fernández[1], José Luis García-Balboa[1] and Inmaculada Barranco-Chamorro[2]

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

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