Investigación

Análisis de Calidad Cartográfica mediante el estudio de la Matriz de Confusión

Quality Cartographic analysis by studying Confusion Matrix

José Manuel Sánchez Muñoz

Revista de Investigación

Volumen VI, Número 2, pp. 009–026, ISSN 2174-0410
Recepción: 13 May'16; Aceptación: 1 Jun'16

1 de octubre de 2016

Resumen

En este artículo se expone una metodología para el control de calidad en la producción de cartografía temática, haciendo un análisis de los índices de calidad temática a partir de la obtención de la matriz de confusión o error.

Palabras Clave: cartografía temática, control de calidad, matriz de confusión o error.

Abstract

This article outlines a methodology for quality control in the production of thematic cartography, doing an analysis of the indices of thematic quality from obtaining the confusion or error matrix.

Keywords: thematic cartography, quality control, confusion or error matrix.

1. Introducción

En los últimos 40 años y como consecuencia de la intervención del hombre en la superficie de la Tierra, se iniciaron procesos de degradación del suelo, los cuales repercuten de manera directa sobre las condiciones de vida del ser humano.

Como consecuencia directa, se hace necesario y de vital importancia el conocimiento de las características del uso de la superficie del planeta, así como las dinámicas de evolución del mismo, en tanto en cuanto dicho conocimiento sirve de análisis de los factores medioambientales y humanos que interactúan en el paisaje. Es por ello que en los últimos años se ha producido un espectacular desarrollo en la ingeniería en torno a los satélites de observación terrestre, con el

fin de poder abordar trabajos cartográficos de ocupación de la superficie con mayor precisión y calidad.

Desde este punto de vista surgieron programas a nivel europeo como CORINE (Coordination of Information of Environment) en 1985 cuya finalidad consistía fundamentalmente en la recopilación de datos, la coordinación y la homogeneización de la información sobre el estado del Medio Ambiente y los recursos naturales, teniendo como objetivo principal la creación y actuación permanente de información sobre la cobertura y usos del suelo del territorio europeo, así como la creación de una base de datos numérica y geográfica a escala 1:100.000.

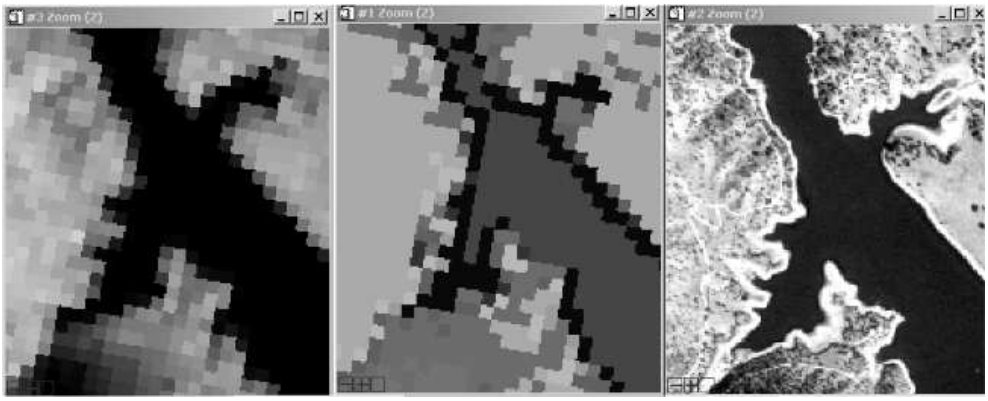


Figura 1. Ejemplo de clasificación temática [1].

Para la correcta generación de dicha producción cartográfica se necesitan procesos de control de calidad bien desde el punto de vista de la exactitud posicional, o bien desde la exactitud temática. En la norma ISO 19113 se definen elementos generales de calidad, para describir el propósito y uso del producto cartográfico, así como el linaje de los datos.

Con respecto a la calidad cartográfica podemos hacer las siguientes afirmaciones:

1. Es propia de la componente temática de la cartografía.
2. **NO** es exclusiva de los denominados mapas temáticos.
3. Cualquier elemento representado en un mapa topográfico pertenece a un tema, relacionándose con el mismo gracias a la leyenda.
4. Ligada a la posición ya que el tema depende de ésta.
5. Su tratamiento independiente de la posición tradicionalmente menos considerado.

Se define exactitud temática al *"grado de conformidad de una entidad de la leyenda respecto a la verdad-terreno"*.

La norma ISO 19114 establece los pasos a seguir para la evaluación de la calidad y trata de asegurar una base estadística para asegurar los resultados representativos de la misma (muestreos):

1. Identificar un elemento, subelemento y ámbitos aplicables.
2. Identificar una medida de calidad.
3. Seleccionar y aplicar un método de evaluación de la calidad.
4. Determinar el resultado de la calidad de los datos.

5. Determinar la conformidad.

Para el seguimiento de la calidad temática de una cartografía, será necesario conocer la naturaleza de los errores cometidos (¿qué entidades se confunde?), la frecuencia con que se comenten (¿probabilidad de que ocurra?), su importancia y magnitud, y la fuente de generación de los mismos (¿pueden minimizarse?). En cualquier caso la herramienta fundamental para llevar a cabo dicho análisis es la *matriz de confusión* que pasamos a ver a continuación.

La tabla 1 muestra las razones por la que se comenten errores en cartografía dependiendo del sujeto de origen que los puede llevar a cabo.

Tabla 1. Errores cartográficos.

1. Toma de datos	a) Datos incompletos
	b) Uso de conceptos equivocados
2. Editor-Autor del mapa	a) Mala elección de los datos
	b) Definir incorrectamente los propósitos del mapa
	c) Incluir excesiva o muy poca información
3. Diseñador cartográfico	a)Variables visuales mal seleccionadas
	b) Diseño erróneo de la simbología
4. Dibujante cartográfico	a) Calidad pobre del dibujo
	b) Colocación de textos incorrecta
5. Usuarios del mapa	a) Incapaz de detectar la información relevante
	b) Nivel cultural y de conocimientos inadecuado
	c) Errónea interpretación de la información

2. Matriz de Confusión

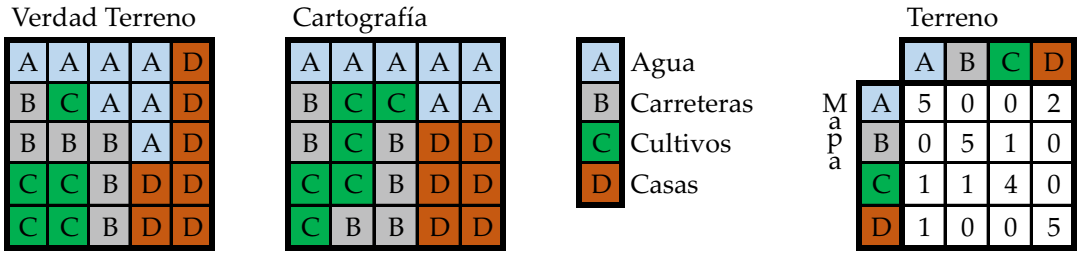
2.1. Descripción

Se la denomina también matriz de error o tabla de contingencia. La matriz de confusión se construye a partir de una imagen de satélite con *N celdillas* clasificadas en *M clases*. Sobre las columnas se ordenan las clases reales (verdad-terreno), y sobre las filas las unidades cartográficas (unidades -o clases- del mapa). Los elementos que aparecen en la diagonal nos indican el número de clasificaciones realizadas correctamente, y aquellos que aparecen fuera suponen migraciones o fugas. Desde el punto de vista de la interpretación de la matriz de confusión, existen dos tipos de errores:

- ✓ Errores de omisión (riesgos del usuario): son los elementos que perteneciendo a esa clase no aparecen en ella por estar erróneamente incluidos en otra (datos por debajo de la diagonal principal de la matriz de confusión).
- ✓ Errores de comisión (riesgos del productor): son los elementos que no perteneciendo a una clase aparecen en ella (datos por encima de la diagonal principal de la matriz de confusión).

La matriz de confusión facilita la detección de errores y además:

- ✓ Permite el análisis descriptivo.
- ✓ Visión general de las asignaciones correctas y de las equivocaciones.
- ✓ Permite el análisis analítico.
- ✓ Distintos niveles de análisis: global, por tipo de entidad, por casos concretos.



Mapa \ Terreno	A	B	C	D
A	5	0	0	2
B	0	5	1	0
C	1	1	4	0
D	1	0	0	5

Figura 2. Ejemplo de obtención de matriz de confusión.

2.2. Fuentes de error

Existen diferentes fuentes posibles de error a la hora de confeccionar una cartografía temática apoyada en teledetección a partir de imágenes por satélite:

- ✓ Diferencias de registro entre los datos de referencia y el mapa de unidades cartográfica.
- ✓ Errores de delineación cuando se marcan las parcelas de seguimiento de exactitud.
- ✓ Errores en la entrada de datos cuando se introducen los datos del muestreo.
- ✓ Cambios en la cubierta entre las fechas de la imagen y de la toma de datos de referencia (error temporal).
- ✓ Variación en la clasificación y delimitación de los datos de referencia debido a inconsistencias de la interpretación humana.
- ✓ Errores en la clasificación de los datos del satélite.
- ✓ Errores en la delineación de los datos del satélite.

2.3. Generación

Con el fin de crear la matriz de confusión muestral, ésta debe tener unas condiciones específicas:

- ✓ Las clases que se establezcan deben ser independientes, mutuamente excluyentes y exhaustivas y en número suficiente.
- ✓ Deben usarse métodos de muestreo que excluyan autocorrelación.
- ✓ Conviene el uso de métodos estratificados para asegurar la presencia de clases extrañas o minoritarias.
- ✓ Para comprobar la bondad de un proceso de clasificación supervisada, no se deben usar las parcelas de entrenamiento del clasificador.

La figura 3 y la tabla 2 muestran un ejemplo sencillo de generación de una matriz de confusión a partir de la comparativa entre la cartografía y los datos reales del terreno.

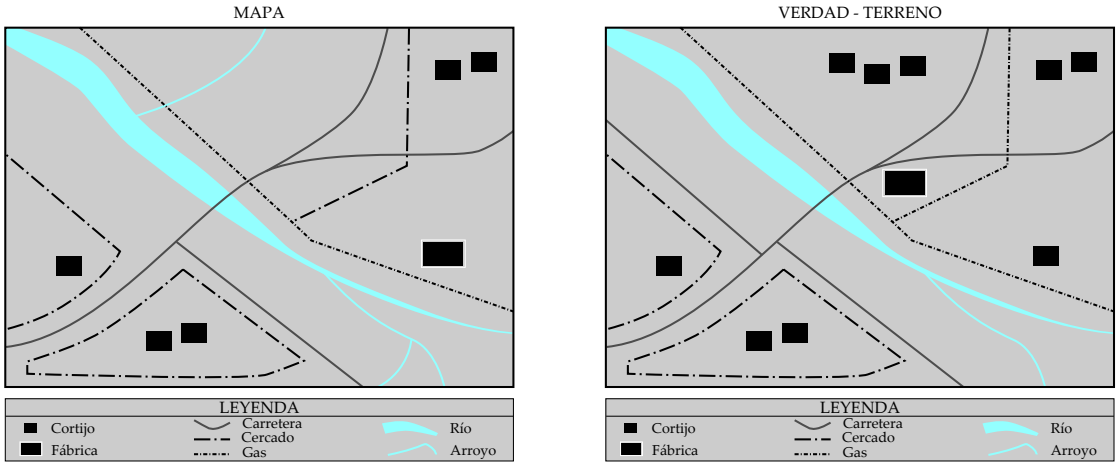


Figura 3. Comparativa entre catografía y verdad-terreno.

Tabla 2. Generación de la matriz de confusión.

		TERRENO							n_{i+}	
		Cortijo	Fábrica	Carretera	Cercado	Gas	Río	Arroyo	Σ	Comisión
MAPA	Cortijo	4							4	1
	Fábrica	1							1	
	Carretera			6					6	
	Cercado				6	2			8	
	Gas					4			4	
	Río						1		1	
	Arroyo							2	2	2
n_{+j}	Σ	5	0	6	6	6	1	2	26	
	Omisión	3	1	1						

2.4. Normalización

Con el fin de facilitar el trabajo de análisis de control de la calidad temática se puede realizar una *normalización* de la matriz de confusión, que permite presentar todos los valores en tanto por uno. Consiste en un proceso iterativo de compensación, de forma que se va consiguiendo el valor unidad en los marginales de las filas y las columnas hasta alcanzar el umbral establecido. Dicho proceso se realiza mediante un cálculo iterativo, y consiste en:

1. En general una matriz X está formada por celdas que denotamos x_{ij}. Definimos x_{i+} a la suma de todos los elementos de la fila i de la matriz X, y x_{+j} la suma de todos los elementos de la columna j de dicha matriz.
2. Se procede a realizar la división de cada elemento a_{ij} de la matriz A de partida por el correspondiente sumatorio por fila a_{i+}, obteniéndose una nueva matriz A^1.
3. A continuación se procede a dividir cada elemento a^1_{ij} de la matriz A^1 por el correspondiente sumatorio por columna a^1_{+j} obteniéndose una nueva matriz A^2, completando la primera iteración.
4. Se repite nuevamente el proceso iterativo hasta que se obtenga en la $n-$ésima iteración una matriz A^{2n} cuyos sumatorios por fila y por columna sean unitarios.

La utilización de este proceso presenta una serie de ventajas e inconvenientes:

- Ventajas:
  - ✓ Los valores dentro y fuera de la diagonal principal representan los errores de comisión y omisión de forma mucho más clara.
  - ✓ Es más fácil de comparar con otras matrices.
  - ✓ Es más fácil de comparar unas clases con respecto a otras en %.
- Inconvenientes:
  - ✓ No se tiene en cuenta el % de superficie que ocupa cada clase.
  - ✓ ¿Todas las clases son igual de importantes?

Si numerosas celdas poseen un valor "0" puede deberse a que si son fijos es porque se produce una limitación natural de la clasificación, y por el contrario, si son aleatorios puede que se haya llevado a cabo un muestreo deficiente o una clasificación extraordinariamente buena. Pueden afectar en gran medida al proceso de normalización. Para eliminarlos, se emplea el *método de sustitución por pseudoceros*, cuya metodología fue establecida por Feinberg y Holland (1970) y consiste en realizar una serie de operaciones sobre la matriz de confusión como describimos a continuación:

1. Partimos de una matriz de confusión M cuyas celdas denominaremos m_{ij}.
2. Creamos una nueva matriz E cuyas celdas denominamos e_{ij}. El valor de cada celda e_{ij} se determina mediante la expresión:
$$e_{ij} = \frac{m_{+j} \cdot m_{i+}}{n}$$
siendo, n el número total de casos de la matriz M, m_{+j} el valor total marginal por columna (suma de todos los elementos de una columna de la matriz M), y m_{i+} el valor total marginal por fila (suma de todos los elementos de una fila de la matriz M). Los elementos de la matriz E son las probabilidades que cabe esperar en cada celda bajo la hipótesis de independencia.
3. Se determina el número ν mediante la expresión:
$$\nu = \frac{n^2 - \sum_{i=1}^{r}\sum_{j=1}^{r} {m_{ij}}^2}{\sum_{i=1}^{r}\sum_{j=1}^{r}(e_{ij} - m_{ij})^2}$$
donde r es el rango de la matriz M.
4. Generamos una nueva matriz que denominaremos P con celdas p_{ij} obtenidas a partir de la expresión:
$$p_{ij} = \frac{e_{ij} \cdot \nu}{n}$$
5. Sumamos las matrices P y M y multiplicamos cada una de las celdas por el factor $\frac{n}{n+\nu}$ obteniendo así el valor del pseudocero en cada celda (i, j).

3. El muestreo

Con el fin de poder llevar a cabo un análisis de control de calidad temática de la cartografía generada, se ha de llevar a cabo un muestreo sobre el terreno para realizar la comparativa pertinente. En cuanto al tamaño de la muestra, ésta debe de tener cierta significación estadística, para lo cual como regla general, se recomienda tomar al menos 50 muestras por cada clase. Existen varios tipos de muestreos como se especifica en la figura 4.

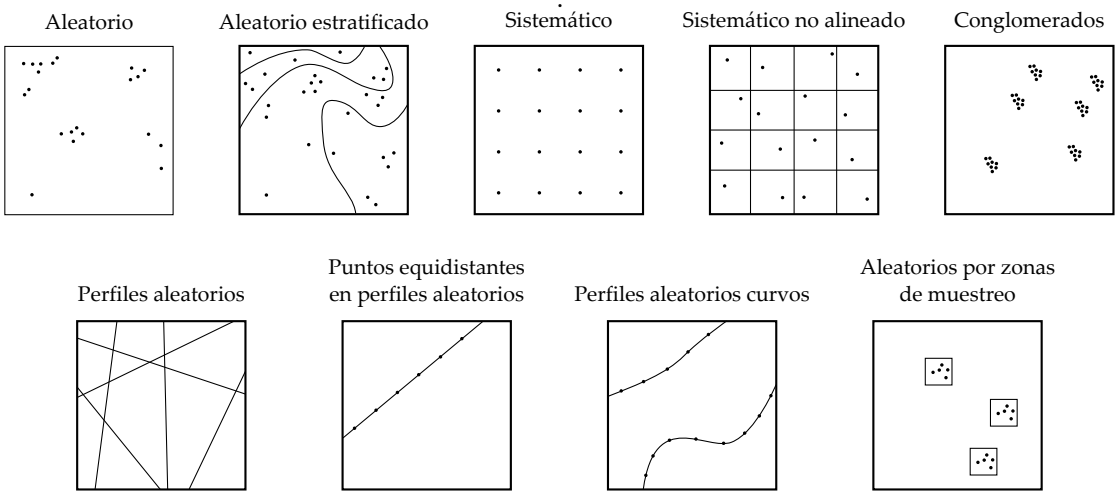


Figura 4. Tipos de muestreo.

4. Índices de calidad temática

Mediante el estudio analítico de la matriz de confusión, podemos llevar a cabo conclusiones acerca de los trabajos cartográficos llevados a cabo. Existen varios tipos de índices, *globales* (ofrecen una valoración de la calidad de toda la clasificación), *por clase* (se emplean en caso de necesitar un mayor nivel de conocimiento sobre una clase concreta), y *por caso* (analizan el comportamiento estadístico de una sola celda de la matriz).

4.1. Índices globales

4.1.1. Porcentaje de acuerdo. P_a

Se trata de un coeficiente sencillo de calcular y muy intuitivo. Sobrestima la bondad de la clasificación dado que no considera los errores entre las clases. Puede considerarse como la probabilidad de estar o no bien clasificado, por ello puede suponerse que su distribución siga el comportamiento de una función binomial. Analíticamente se expresa:

$$P_a = \frac{1}{N}\sum_{i=1}^{M} n_{i,i} = \sum_{i=1}^{M} p_{i,i}$$

donde:

- ✓ M representa el número de clases.

- ✓ N expresa el número total de muestras (número de datos).
- ✓ $n_{i,i}$ representa el número de casos en la diagonal.

El coeficiente tiene una varianza

$$\sigma^2_{P_a} = \sigma^2(P_a) = \frac{P_a(1-P_a)}{N}$$

4.1.2. Coeficiente de acuerdo aleatorio (a priori). $C_{a_{pr}}$

Es un coeficiente que no necesita la matriz de confusión. Es sencillo de calcular. Las probabilidades consideradas son a priori a la clasificación. Cuando todas las probabilidades a priori son iguales, se cumple que

$$C_{a_{pr}} = \frac{1}{M}$$

es decir que es la inversa del número de clases que tengamos, lo que significa que si se consideran muchas clases, su valor disminuye, y su clasificación es más complicada. Su varianza es nula.

4.1.3. Coeficiente de acuerdo aleatorio (a posteriori). $C_{a_{ps}}$

Se basa exclusivamente en las distribuciones marginales de la matriz de confusión; es decir, el de las probabilidades a posteriori de cada una de las clases. Es sencillo de calcular, y representa el porcentaje de acuerdo que cabe esperar al azar teniendo en cuenta que unas clases contienen un mayor número de celdillas que otras y que por lo tanto son más probables de estar bien clasificadas. Su expresión analítica es:

$$C_{a_{ps}} = \sum_{i=1}^{M} P_{i+} \cdot P_{+i} = \frac{1}{N^2} \sum_{i=1}^{M} n_{i+} \cdot n_{+i}$$

4.1.4. Coeficiente Kappa de ajuste. κ

Su uso está muy extendido. Considera las distribuciones marginales de la matriz de confusión, es decir, las probabilidades a posteriori de pertenencia a una clase. Muestra cuánto ha mejorado la clasificación respecto a una asignación aleatoria de N elementos en M grupos. Da idea del % de acuerdo, una vez se ha eliminado la parte debida al azar. Sobrestima la aportación del acuerdo al azar y de esta forma subestima la bondad de la clasificación total. Cuando N es grande puede considerarse que se distribuye según una normal. Su expresión analítica es:

$$\kappa = \frac{P_a - C_{a_{ps}}}{1 - C_{a_{ps}}}; \qquad \sigma^2_\kappa = \sigma^2(\kappa) = \frac{P_a(1-C_{a_{ps}})}{N(1-C_{a_{ps}})^2}$$

4.1.5. Coeficiente Tau de ajuste. τ

Es un coeficiente similar a κ, pero mucho menos utilizado como parámetro de calidad. Su valor, da idea de cuánto ha mejorado el sistema de clasificación respecto a una clasificación aleatoria de los N elementos en M grupos. Se basa en la probabilidad a priori de pertenencia a un grupo. Cuando N es grande puede considerarse que se distribuye según una normal. Su expresión analítica es:

$$\tau = \frac{P_a - C_{a_{pr}}}{1 - C_{a_{pr}}}; \qquad \sigma^2_\tau = \sigma^2(\tau) = \frac{P_a(1-P_a)}{N(1-C_{a_{pr}})^2}$$

4.2. Índices por clase

4.2.1. Exactitud del usuario. *EU*

También denominada *pureza de la unidad cartográfica*. Representa la probabilidad de que un pixel escogido aleatoriamente y clasificado en una unidad cartográfica del mapa, esté correctamente asignado. Es un índice adecuado para acompañar al P_a cuando existen notables diferencias en la pureza de las unidades del mapa. Este índice determina para una clase el porcentaje de los elementos de comprobación realmente bien clasificados. Su expresión analítica es:

$$EU = \frac{x_{i,i}}{n_{i+}}$$

4.2.2. Riesgo del productor. *RP*

Es el complementario a la unidad del índice anterior (EU). Son los elementos que, perteneciendo a distintas clases de la verdad-terreno, se han incluido erróneamente en una misma unidad cartográfica dada. Suponen un riesgo para el *productor*, ya que si el usuario los utiliza como comprobación, puede demostrar que el trabajo del productor no está bien hecho. También se denomina *error de comisión (ERC)*, ya que esas inclusiones son errores por comisión dentro de la unidad cartográfica considerada. Su expresión analítica es:

$$RP(i) = 1 - EU(i)$$

4.2.3. Exactitud del productor. *EP*

Es la probabilidad de que un pixel escogido aleatoriamente y perteneciente a una clase esté correctamente asignado a una unidad cartográfica. Indica por lo tanto, lo que realmente está bien consignado en la unidad cartográfica del producto. Su expresión analítica es:

$$EP(j) = \frac{x_{j,j}}{n_{+j}}$$

4.2.4. Riesgo del usuario. *RU*

Es el complementario a la unidad del índice anterior (EP). Es el porcentaje de elementos mal clasificados y que, por tanto, suponen un riesgo de uso. Se le denomina también *error de omisión (ERO)*, pues los elementos de la verdad-terreno que no se han incluido en la unidad cartográfica son errores del tipo omisión. Su expresión analítica es:

$$RU(j) = 1 - EP(j)$$

4.3. Índices por caso

4.3.1. Coeficiente Kappa por clase. κ_c

Es un coeficiente bastante menos utilizado como parámetro de calidad que el kappa global (κ), aunque su sentido es muy similar. Sus valores varían entre 0 y 1, representando el valor 1 el caso de total acuerdo. Su expresión analítica es:

$$\kappa_c(i) = \frac{P_{ii} - P_{+i} \cdot P_{i+}}{P_{i+} - P_{+i} \cdot P_{i+}}$$

$$\sigma^2(\kappa_c) \approx \frac{n(n_{i+} - n_{ii})}{[n_{i+}(n - n_{+i})]^3}[(n_{i+} - n_{ii})(n_{i+}n_{+i} - n\, n_{ii}) + n\, n_{ii}(n - n_{i+} - n_{+i} + n_{ii})]$$

4.3.2. Probabilidad del caso *i, j*. P_{ij}

Se trata de un índice que estima la probabilidad de la celda i, j de la matriz de confusión. Su expresión analítica es:

$$P_{ij} = \frac{n_{i,j}}{N}$$

5. Test de control

Una vez obtenidos los índices resultado del análisis de la matriz de confusión, se pueden derivar consecuencias a partir de las propiedades estadísticas de éstos. Existen varios tipos que dependen, entre otros factores, de la naturaleza del muestreo llevado a cabo, y entre los que destacamos:

- ✓ Test P_a para muestreos aleatorios simples.
- ✓ Test P_a para muestreos aleatorios estratificados.
- ✓ Test Kappa para muestreos aleatorios simples.
- ✓ Test para la comparación de dos matrices de confusión.

5.1. Test P_a para muestreos aleatorios simples

Podemos encontrarnos con dos casos en función del número de elementos:

- ✓ $H_0 : P_a \geq P_{a_0}$
- ✓ $H_1 : P_a < P_{a_0}$

En el caso de tratarse un muestreo con un número reducido de elementos, haríamos el contraste mediante aproximación binomial, de manera que la regla de decisión sería:

- ✓ Si $P > \alpha(RP)$ se acepta H_0.
- ✓ Si $P < \alpha$ se rechaza.

$$P[r \leq x] = \sum_{r=0}^{x} \frac{n!}{(n-r)! \cdot r!} \cdot {P_{a_0}}^r \cdot (1 - P_{a_0})^{n-r}$$

siendo:

- ✓ x el total de muestras correctamente clasificadas.
- ✓ P_{a_0} el umbral definido para un nivel de confianza marcado $(1 - \alpha)$.
- ✓ n el número de elementos de la muestra.

Si el número de elementos del muestreo fuera lo suficientemente grande, realizamos una aproximación por la normal, siendo la regla de decisión:

- ✓ Si $Z > Z_{1-\alpha}$ se acepta la hipótesis nula H_0.
- ✓ Si $z < Z_{1-\alpha}$ se rechaza.

$$Z = \frac{P_a - P_{a_0}}{\sqrt{\frac{P_{a_0} \cdot (1 - P_{a_0})}{n}}}$$

siendo:

- ✓ Z el estadístico a contrastar.
- ✓ $Z_{\alpha/2}$ el cuantil de la distribución normal tipificada correspondiente a un nivel de confianza bilateral de $(1 - \alpha)$.

5.2. Test P_a para muestreos aleatorios estratificados

Se utiliza fundamentalmente para control de calidad en procesos de clasificación temática por teledetección, fotointerpretación, etc, donde se utilizan muestreos de tipo aleatorio estratificado.

Se calcula el índice P_a asignando a cada clase un peso proporcional a su extensión sobre el terreno.

$$P_a = \sum_{i=1}^{k} \frac{n_{i,i} \cdot a_{i,i}}{n_{i+}}$$

siendo:

- ✓ n_{i+} el total de casos en la fila $i-$ésima de la matriz de confusión.
- ✓ $n_{i,i}$ los casos de la celda $i-$ésima sobre la diagonal principal de la matriz de confusión.
- ✓ $a_{i,i}$ la extensión relativa (%) de la clase $i-$ésima respecto al área total (peso).

La regla de decisión es:

- ✓ Si $Z < Z_{1-\alpha}$ se acepta la hipótesis nula ($H_0 : P_a > P_{a_0}$).
- ✓ Si $Z < Z_{1-\alpha}$ se rechaza.

5.3. Test Kappa para muestreos aleatorios simples

Se utiliza fundamentalmente para control de calidad en procesos de clasificación temática por teledetección, fotointerpretación, etc, donde se utilizan muestreos de tipo aleatorio simple. Para este test no se ha definido una correspondencia estándar, por lo que se podrían definir categorías de exactitud en función de unos umbrales κ_0 admisibles.

Se calcula el índice κ (coeficiente Kappa de ajuste):

- ✓ $H_0 : \kappa \geq \kappa_0$
- ✓ $H_1 : \kappa < \kappa_0$

$$Z = \frac{\hat{\kappa}}{\sqrt{\sigma^2(\hat{\kappa})}}$$

La regla de decisión es:

✓ Si $Z > Z_{1-\alpha}$ se acepta la hipótesis nula.

✓ Si $Z < Z_{1-\alpha}$ se rechaza.

5.4. Test para la comparación de dos matrices de confusión

Se utiliza para llevar a cabo un control de calidad basado en la comparativa de dos trabajos expresados mediante dos matrices de confusión.

Para este test no se ha definido una correspondencia estándar. El resultado deber ser cumple/no cumple para un nivel de significación establecido. Se basa en un contraste de hipótesis sobre dos valores de P_a, κ, o τ.

Se calcula el índice (por ejemplo P_a) en cada una de las clasificaciones:

✓ $H_0 : P_{a_1} - P_{a_2} = 0 \rightarrow P_{a_1} = P_{a_2}$

✓ $H_1 : P_{a_1} - P_{a_2} \neq 0$

$$Z = \frac{|P_{a_1} - P_{a_2}|}{\sqrt{\sigma^2(P_{a_1}) + \sigma^2(P_{a_2})}}$$

La regla de decisión es:

✓ Si $Z > Z_{\alpha/2}$ se rechaza la hipótesis nula.

6. Caso práctico

6.1. Datos de partida

Partimos de una matriz de confusión que contiene los datos especificados en la tabla 3.

Tabla 3. Datos de la matriz de confusión.

		Terreno								
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**
Mapa	**M1**	238051	1	939	0	0	5	0	29	115
	M2	7	4086	5082	0	48	151	105	36	2
	M3	132	188	51817	5	4	119	601	280	0
	M4	0	0	0	11148	834	135	110	0	4
	M5	0	4	34	1618	2853	726	174	0	124
	M6	24	16	500	78	340	6774	155	6	595
	M7	9	45	1867	0	32	75	8257	5	0
	M8	2	1	325	0	0	1	8	2993	0
	M9	189	0	17	0	197	553	0	0	4374

Las filas de la matriz M1, M2, . . . , M9, representan 9 clases de suelos distintos, mientras que las 9 columnas T1, T2, . . . , T9, representan 9 parcelas de terreno distintas.

Los nueve distintos usos del suelo para las 9 parcelas de terreno son:

M1: agua	M4: arroz	M7: olivos y algarrobos
M2: cultivos	M5: frutales	M8: salinas
M3: suelo improductivo	M6: matorral	M9: juncal

Dependiendo de la dispersión de los datos con respecto a la diagonal principal de la matriz de confusión, podemos deducir si la clasificación está bien definida o no. Todos los elementos que están bien clasificados aparecen en la diagonal principal, mientras que los datos por encima de dicha diagonal principal representan los errores de comisión, es decir aquellos que son inventados, que representan el riesgo del productor, y los datos por debajo de la diagonal principal representan los errores por omisión, es decir que faltan, que representan el riesgo del usuario.

A continuación realizamos el sumatorio de todos los datos de las celdas, tanto por filas como por columnas. La tabla 4 representa los correspondientes sumatorios de filas y columnas. En amarillo se pueden ver los elementos de la diagonal principal (que estarían perfectamente definidos)[1].

Tabla 4. Sumatorios por filas y columnas.

		Terreno									
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	
Mapa	**M1**	238051	1	939	0	0	5	0	29	115	239140
	M2	7	4086	5082	0	48	151	105	36	2	9517
	M3	132	188	51817	5	4	119	601	280	0	53146
	M4	0	0	0	11148	834	135	110	0	4	12231
	M5	0	4	34	1618	2853	726	174	0	124	5533
	M6	24	16	500	78	340	6774	155	6	595	8488
	M7	9	45	1867	0	32	75	8257	5	0	10290
	M8	2	1	325	0	0	1	8	2993	0	3330
	M9	189	0	17	0	197	553	0	0	4374	5330
		238414	4341	60581	12849	4308	8539	9410	3349	5214	347005

6.2. Índices

Con los datos anteriormente especificados obtenemos unos índices que nos ayudarán a efectuar el posterior análisis de calidad de la toma de datos correspondiente.

6.2.1. Índices globales

1. Porcentaje de acuerdo.

$$P_a = \frac{1}{N}\sum_{i=1}^{M} n_{i,i} = \sum_{i=1}^{M} p_{i,i} = \frac{1}{347005}\cdot(238051 + 4086 + \ldots + 4374) = 0,952$$

[1] Nótese que la celda correspondiente al valor 347005, corresponde a N, es decir el número de casos distintos, que es la suma de todas las celdas de la matriz de confusión, o bien el sumatorio de las columnas sumatorio, o bien el sumatorio de las filas sumatorio, esto es:

$$3470005 = 239140 + 9517 + \ldots + 5330 = 238414 + 4341 + \ldots + 5214$$

2. Coeficiente de acuerdo aleatorio (a priori).

$$C_{a_{pr}} = \frac{1}{M} = \frac{1}{9} = 0,111$$

3. Coeficiente de acuerdo aleatorio (a posteriori)

$$C_{a_{ps}} = \frac{1}{N^2}\sum_{i=1}^{M} n_{i+} \cdot n_{+i} = \frac{1}{347005^2} \cdot (238414 \cdot 239140 + 4341 \cdot 9517 + \cdots + 5214 \cdot 5330) = 0,504$$

4. Coeficiente Kappa por clase.

$$\kappa_c(i) = \frac{P_{ii} - P_{+i} \cdot P_{i+}}{P_{i+} - P_{+i} \cdot P_{i+}} = \frac{0,952 - 0,504}{1 - 0,504} = 0,903$$

5. Coeficiente Tau de ajuste.

$$\tau = \frac{P_a - C_{a_{pr}}}{1 - C_{a_{pr}}} = \frac{0,952 - 0,111}{1 - 0,111} = 0,946$$

6.2.2. Índices por clase

	EU	*RP*	*EP*	*RU*
agua	$\frac{238051}{239140} = 0,995$	$1 - 0,995 = 0,005$	$\frac{238051}{238414} = 0,998$	$1 - 0,998 = 0,002$
cultivos	$\frac{4086}{9517} = 0,429$	$1 - 0,429 = 0,571$	$\frac{4086}{4341} = 0,941$	$1 - 0,941 = 0,059$
suelo improductivo	$\frac{51817}{53146} = 0,975$	$1 - 0,975 = 0,025$	$\frac{51817}{60581} = 0,855$	$1 - 0,855 = 0,145$
arroz	$\frac{11148}{12231} = 0,911$	$1 - 0,911 = 0,089$	$\frac{11148}{12849} = 0,868$	$1 - 0,868 = 0,132$
frutales	$\frac{2853}{5533} = 0,516$	$1 - 0,516 = 0,484$	$\frac{2853}{4308} = 0,662$	$1 - 0,662 = 0,338$
matorral	$\frac{6774}{8488} = 0,798$	$1 - 0,798 = 0,202$	$\frac{6774}{8539} = 0,793$	$1 - 0,793 = 0,207$
olivos y algarrobos	$\frac{8257}{10290} = 0,802$	$1 - 0,802 = 0,198$	$\frac{8257}{9410} = 0,877$	$1 - 0,877 = 0,123$
salinas	$\frac{2993}{3330} = 0,899$	$1 - 0,899 = 0,101$	$\frac{2993}{3349} = 0,894$	$1 - 0,894 = 0,106$
juncal	$\frac{4374}{5330} = 0,821$	$1 - 0,821 = 0,179$	$\frac{4374}{5214} = 0,839$	$1 - 0,839 = 0,161$

6.3. Conclusiones

Para proceder al análisis, una vez obtenidos los índices correspondientes podemos realizar las siguientes afirmaciones.

6.3.1. Índices globales

1. Según el porcentaje de acuerdo (P_a), el 95, 2 % de los datos están bien clasificados.

2. El coeficiente κ nos da una idea de cuanto mejora el coeficiente de acuerdo aleatorio a posteriori con respecto al porcentaje de acuerdo. En el caso estudiado un $90,3\,\%$.

3. El coeficiente τ nos da una idea de cuanto mejora el coeficiente de acuerdo aleatorio a priori con respecto al porcentaje de acuerdo. En el caso estudiado un $94,6\,\%$.

6.3.2. Índices por clase

Para este estudio nos vamos a fijar en la clase correspondiente a suelo de *cultivos*. El análisis arroja las siguientes conclusiones:

1. Si se llevara a cabo un control de calidad de la parcela, la probabilidad de que se hubiera cometido un error en su clasificación, es de un $57,1\,\%$, es decir el riesgo de que una parcela marcada como cultivo no fuera realmente de cultivo es bastante alto, por lo que sería muy probable que el mapa cartografiado fuera rechazado. Este es el riesgo del productor.

2. Por el contrario, si un usuario va al terreno, se sitúa en una parcela destinada a cultivos, y observa si está bien o mal cartografiada, tendrá únicamente un $5,9\,\%$ de probabilidades de que la parcela donde se ha situado no sea de cultivos. Este es el riesgo del usuario.

3. Estas conclusiones pueden hacerse extensibles al resto de las clases. Este análisis permite hacerse una idea sobre los errores de omisión, que se olvidan, correspondientes al riesgo del usuario, y los errores por comisión, que se inventan, que son los correspondientes al riesgo del productor.

6.4. Normalización

Como última parte de este caso práctico hemos llevado a cabo la normalización de la matriz de confusión, lo cual facilita la comparación de los datos tanto entre sí como con otras posibles matrices. Al normalizar la matriz, obtenemos que tanto los sumatorios de las filas como de las columnas tengan un valor igual a 1. Dicho proceso se realiza por medio de cálculo iterativo ya expuesto en la Sección 2.4.

Este proceso iterativo se puede mecanizar mediante la implementación de una macro de Microsoft Excel que nos permite automatizar las iteraciones. Esta macro tiene el código de programación en Visual Basic que aparece a continuación.

```
Private Sub Botón1_Haga_clic_en()
    Dim Valor(9, 9), Iteraciones As Single
    Dim SumF(9), SumC(9) As Single

'Lee los valores de la matriz de confusión y el número de iteraciones
 For f = 1 To 9
    Valor(f, 1) = Worksheets("C2_Normalización").Range("c" & f + 3).Value
    Valor(f, 2) = Worksheets("C2_Normalización").Range("d" & f + 3).Value
    Valor(f, 3) = Worksheets("C2_Normalización").Range("e" & f + 3).Value
    Valor(f, 4) = Worksheets("C2_Normalización").Range("f" & f + 3).Value
    Valor(f, 5) = Worksheets("C2_Normalización").Range("g" & f + 3).Value
    Valor(f, 6) = Worksheets("C2_Normalización").Range("h" & f + 3).Value
    Valor(f, 7) = Worksheets("C2_Normalización").Range("i" & f + 3).Value
    Valor(f, 8) = Worksheets("C2_Normalización").Range("j" & f + 3).Value
    Valor(f, 9) = Worksheets("C2_Normalización").Range("k" & f + 3).Value
 Next f
```

```
 Iteraciones = Worksheets("C2_Normalización").Range("d15").Value

'Proceso de Iteración
 For i = 1 To Iteraciones 'Realiza la primera/siguiente iteración

   'Primero calcula el sumatorio de Columnas
    For f = 1 To 9
        SumF(f) = 0
            For c = 1 To 9
                SumF(f) = Valor(f, c) + SumF(f)
            Next c
    Next f
   'Y después itera el sumatorio de Columnas y modifica el valor de cada celda
    For f = 1 To 9
        For c = 1 To 9
            Valor(f, c) = Valor(f, c) / SumF(f)
        Next c
    Next f

   'Después calcula el sumatorio de Filas
    For c = 1 To 9
        SumC(c) = 0
            For f = 1 To 9
                SumC(c) = Valor(f, c) + SumC(c)
            Next f
    Next c
   'Y después itera el sumatorio de Filas y modifica el valor de cada celda
    For c = 1 To 9
        For f = 1 To 9
            Valor(f, c) = Valor(f, c) / SumC(c)
        Next f
    Next c

 Next i

'Escribir resultados de la matriz normalizada
 For f = 1 To 9
    Worksheets("C2_Normalización").Range("c" & f + 18).Value = Valor(f, 1)
    Worksheets("C2_Normalización").Range("d" & f + 18).Value = Valor(f, 2)
    Worksheets("C2_Normalización").Range("e" & f + 18).Value = Valor(f, 3)
    Worksheets("C2_Normalización").Range("f" & f + 18).Value = Valor(f, 4)
    Worksheets("C2_Normalización").Range("g" & f + 18).Value = Valor(f, 5)
    Worksheets("C2_Normalización").Range("h" & f + 18).Value = Valor(f, 6)
    Worksheets("C2_Normalización").Range("i" & f + 18).Value = Valor(f, 7)
    Worksheets("C2_Normalización").Range("j" & f + 18).Value = Valor(f, 8)
    Worksheets("C2_Normalización").Range("k" & f + 18).Value = Valor(f, 9)
 Next f

End Sub
```

Ha de comentarse que el caso práctico estudiado se trata de una matriz cuadrada de rango 9. En cualquier otro caso el código de programación puede ser perfectamente modificable. Si la matriz fuera de rango r, habría que cambiar 9 por dicho factor r. Por lo tanto a la hora de leer

en el primer proceso iterativo y de escribir en el último, en lugar de 9 líneas tendríamos *r* líneas. El nombre de la hoja donde está la macro es "C2_Normalización". El número de iteraciones llevado a cabo en proceso se puede leer en la celda 'D15'. La macro comienza a leerlos elementos de la matriz de confusión desde la celda 'C4' hasta la 'K12' (f + 3), y de igual forma, escribe el resultado final del proceso iterativo en la celda 'C19' hasta la 'K27' (f + 18), ya que el contador f va de 1 a 9.

La tabla 5 muestra el resultado final de la aplicación de esta macro para la normalización de la matriz original con las nueve parcelas distintas y los nueve tipos de suelo diferenciados. Dicho resultado final muestra la aplicación de un proceso de 1000 iteraciones a la matriz de confusión.

Tabla 5. Normalización de matriz de confusión mediante programación de macro de Visual Basic [5], [3].

	T1	T2	T3	T4	T5	T6	T7	T8	T9	Σ
M1	238051	1	939	0	0	5	0	29	115	239140
M2	7	4086	5082	0	48	151	105	36	2	9517
M3	132	188	51817	5	4	119	601	280	0	53146
M4	0	0	0	11148	834	135	110	0	4	12231
M5	0	4	34	1618	2853	726	174	0	124	5533
M6	24	16	500	78	340	6774	155	6	595	8488
M7	9	45	1867	0	32	75	8257	5	0	10290
M8	2	1	325	0	0	1	8	2993	0	3330
M9	189	0	17	0	197	553	0	0	4374	5330
Σ	238414	4341	60581	12849	4308	8539	9410	3349	5214	347005

Número de iteraciones:	1000

Resultado tras el proceso de iteración:

	T1	T2	T3	T4	T5	T6	T7	T8	T9	Σ
M1	0,992	0,000	0,003	0,000	0,000	0,000	0,000	0,001	0,004	1,000
M2	0,000	0,913	0,061	0,000	0,007	0,011	0,006	0,002	0,000	1,000
M3	0,003	0,058	0,858	0,000	0,001	0,013	0,045	0,022	0,000	1,000
M4	0,000	0,000	0,000	0,829	0,151	0,013	0,007	0,000	0,001	1,000
M5	0,000	0,001	0,001	0,164	0,703	0,092	0,016	0,000	0,023	1,000
M6	0,001	0,005	0,009	0,007	0,077	0,785	0,013	0,001	0,102	1,000
M7	0,000	0,021	0,046	0,000	0,010	0,012	0,911	0,001	0,000	1,000
M8	0,000	0,001	0,022	0,000	0,000	0,000	0,002	0,974	0,000	1,000
M9	0,005	0,000	0,000	0,000	0,051	0,074	0,000	0,000	0,870	1,000
Σ	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	9,000

Del mismo modo podemos utilizar el método de sustitución por pseudoceros para eliminar algunos ceros que vemos que aparecen en la matriz de confusión inicial obteniendo así la matriz de la tabla 6. Esta técnica se utiliza normalmente de forma previa al proceso de normalización de la matriz de confusión en el caso en que ésta presente una cantidad considerable de ceros en sus celdas que dificulten dicho proceso.

Tabla 6. Método de sustitución por pseudoceros [6].

	T1	T2	T3	T4	T5	T6	T7	T8	T9	Σ
M1	238049,822	1,04776352	939,651792	0,14142334	0,04741628	5,09390519	0,10357177	29,0363978	115,055552	239140
M2	7,10431966	4085,93664	5081,94537	0,00562819	48,0011204	151,001329	105,002445	36,000892	2,00225192	9517
M3	132,581071	188,007616	51816,3206	5,03134979	4,01047382	119,018986	601,013419	280,00372	0,01275385	53146
M4	0,13421259	0,00244372	0,03410342	11147,8292	833,989105	135,002651	110,00354	0,00188528	4,00287128	12231
M5	0,06071444	4,00104159	34,0148845	1617,97743	2852,95553	725,99058	173,999617	0,00085286	123,999347	5533
M6	24,0927568	16,0014403	500,015681	78,0037739	339,996253	6773,89515	155,001201	6,00121251	594,992534	8488
M7	9,11276997	45,0013372	1866,99887	0,00608533	32,0015292	75,0028463	8256,87258	5,00150624	0,00246937	10290
M8	2,03650864	1,00064935	325,004094	0,00196931	0,00066027	1,00129276	8,00131446	2992,95271	0,00079913	3330
M9	189,055468	0,00106492	17,01459	0,00315207	196,997911	552,993263	0,00230843	0,00082156	4373,93142	5330
Σ	238414	4341	60581	12849	4308	8539	9410	3349	5214	347005

Referencias

[1] ARIZA, Francisco J. *Calidad en producción cartográfica*, Universidad de Jaén, 2000.

[2] ATKINSON, Alan D.J. *Apuntes de Investigación aplicada en producción cartográfica, Tema 3 (parte II). Producción cartográfica. Exactitud posicional, exactitud temática y procesos de generalización y georeferenciación de la información*, Máster Universitario de Especialización en Geotecnologías Topográficas en la Ingeniería, Universidad de Extremadura, 2011.

[3] CONGALTON, Russel G.; GREEN, Kass. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers, 1998.

[4] FALLAS, Jorge. *Normas y Datos para Estándares Geoespaciales*, Laboratorio de Teledetección y Sistemas de Información Geográfica Programa Regional en Manejo de Vida Silvestre y Escuela de Ciencias ambientales Universidad Nacional, Heredia, Costa Rica, 2002.

[5] FEINBERG, Stephen E. *An iterative procedure for estimation in contingency tables*, The Annals of Mathematical Statistics, pp. 907–917, Vol. 41, No. 3, 1970.

[6] FEINBERG, Stephen E.; HOLLAND, Paul W. *Methods for Eliminating Zero Counts in Contingency Tables*, Random Counts in Scientific Work (G.P. Patil, editor), Pennsylvania State University, University Park, Pennsylvania, pp. 233–260, No. 1, 1970.

[7] RUESCAS ORIENT, Ana Belén. *Cartografía de Usos del Suelo por Teledetección: La Cuenca del Carraixet*, Cuadernos de Geografía, pp. 65–66, 103–121, Valencia, 1999.

Sobre el autor:
Nombre: José Manuel Sánchez Muñoz
Correo electrónico: jmanuel.sanchez@gmx.es
Institución: Ingeniero de Caminos, Canales y Puertos. Profesor de Enseñanza Secundaria. Grupo de Innovación Educativa "Pensamiento Matemático", Universidad Politécnica de Madrid, España.