Revista Internacional de Ciencia y Tecnología de la Información Geográfica
International Review of Geographical Information Science and Technology

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

CONTROL ESTRICTO DE MATRICES DE CONFUSIÓN POR MEDIO DE DISTRIBUCIONES MULTINOMIALES

FRANCISCO JAVIER ARIZA-LÓPEZ[1], JOSÉ RODRÍGUEZ-AVI[2], VIRTUDES ALBA-FERNÁNDEZ[2].
[1]Dep. de Ingeniería Cartográfica, Geodésica y Fotogrametría. Universidad de Jaén. Campus Las Lagunillas s/n. Edificio A3. 23071, Jaén. España
[2]Dep. de Estadística e Investigación Operativa. Universidad de Jaén.
fjariza@ujaen.es
Campus Las Lagunillas s/n. 23071, Jaén. España
jravi@ujaen.es, mvalba@ujaen.es

RESUMEN

Las matrices de confusión son la forma más usual y estándar de informar sobre la exactitud temática de productos derivados de la clasificación de datos procedentes de imágenes. En este marco, son ampliamente utilizados dos índices: el porcentaje de acuerdo y el índice Kappa. Ambos son índices globales y no permiten un control categoría por categoría y, aún menos, establecer dentro de una categoría condiciones específicas. En este trabajo se propone un método novedoso basado en la distribución multinomial y en un test estadístico exacto. De esta forma, se pueden establecer las preferencias de exactitud para cada categoría y también establecer cierto grado de mala clasificación entre distintas categorías.

Palabras clave: control de calidad, exactitud temática, distribución multinomial, matriz de confusión

STRICT CONTROL OF CONFUSION MATRICES BY MULTINOMIAL DISTRIBUTIONS

ABSTRACT

Confusion matrices are the most usual and standard way of reporting the thematic accuracy of products derived from the classification of data from images. In this context, two indices are widely used: the overall agreement percentage and the Kappa index. Both are global indices and do not allow for category-by-category control, and even less, to establish specific conditions within a category. In this paper we propose a novel method based on multinomial distribution and an exact statistical test. In this way, you can set the accuracy requirements for each category and also

Recibido: 02/10/2017
Aceptada versión definitiva: 05/07/2018
Editor al cargo: Dr. Lluís Pesquer

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

establish some degree of misclassification between different categories. A practical example is given.

Key words: quality control, thematic accuracy, multinomial distribution, confusion matrix

1. Introducción:

Una matriz de confusión o matriz de error, es una tabla de contingencia que sirve como herramienta estadística para el análisis de observaciones emparejadas. Como indican Comber *et al.* (2012), la matriz de confusión ha sido adoptada, de facto y de jure, como un estándar para informar sobre la exactitud temática de cualquier producto de datos derivados de la teledetección. Por supuesto, esta misma herramienta puede ser utilizada para evaluación de la calidad temática de cualquier tipo de dato espacial (p.ej. parcelas catastrales, cubiertas vegetales, red viaria, base de datos topográfica, etc.). En esta línea, la matriz de confusión aparece reconocida en la Norma Internacional ISO 19157, relativa a la calidad de la información geográfica, como un mecanismo para ofrecer los resultados de la calidad temática de productos vectoriales o derivados de imágenes (p.ej. clasificación de una imagen).

El contenido de una matriz de confusión es un conjunto de valores que contabilizan el grado de semejanza entre observaciones emparejadas: un conjunto de datos bajo control (CDC) y un conjunto de datos de referencia (CDR), para los que se ha establecido una clasificación. Usualmente el CDR es la verdad terreno, es decir, la realidad, y suele conocerse por medio de un muestreo. La matriz de confusión puede construirse a partir de píxeles, agrupaciones de píxeles o cualquier tipo de objeto geográfico (p.ej. polígonos). Con independencia de su tipología, los elementos del CDC se comparan con sus homólogos en el CDR.

De esta forma, se trata de una matriz cuadrada de dimensión $M \times M$ (filas × columnas), donde M denota el número de clases en consideración. Las clases del CDR las denominamos Γ_R (clases referencia) y las clases del CDC las denominamos G_P (clases producto) Cada uno de los M^2 elementos de la matriz los denominamos celdas de la matriz. Las celdas de la diagonal de la matriz de confusión contienen las cantidades correspondientes a los ítems bien clasificados (coincide una G_P con su correspondiente Γ_R). Estas celdas las denominamos C_C (*celdas coincidencia*). Las celdas de fuera de la diagonal contienen las cantidades correspondientes a las confusiones, los errores debidos a las omisiones y comisiones. Estas celdas las denominamos C_E (*celdas error* o *no coincidencia*). La Ec.1 indica el conteo que se almacena en cada una de las celdas de la matriz.

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

Una matriz de confusión ofrece una visión completa de la distribución de los acuerdos y errores entre clases, pero es difícil de manejar de una manera sencilla, y por esta razón existen distintos índices derivados para resumir su información por medio de un valor, o por un conjunto reducido de valores. Existen numerosas medidas o índices de exactitud temática derivados de una matriz de confusión, véase Liu *et al.* (2007) para un análisis comparativo. Dos índices globales ampliamente adoptados son el porcentaje de acuerdo (PA) y el coeficiente Kappa (κ). El primero es

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

la ratio entre el total de elementos correctamente clasificados (celdas de la diagonal principal), y el total de elementos en la matriz (Ec.2). El coeficiente Kappa (Ec.4) es una medida basada en la diferencia entre el porcentaje de acuerdo indicado por los valores de la diagonal principal y el acuerdo aleatorio a posteriori (Ec.3), estimado a partir de los valores marginales (totales de las filas y columnas). Podemos considerar que el coeficiente Kappa (Ec.4) es una corrección de PA en orden a descontar la cantidad de clasificación correcta que ocurre aleatoriamente. Ambos índices están ampliamente adoptados en trabajos y herramientas de software, pero existe cierta crítica sobre su uso por causa de los problemas de subestimación y sobre-estimación que introducen (Veregin 1989, Nishii and Tanaka 1999). Ambos índices presentan una aproximación binomial, es decir, consideran sólo dos estados posibles, bien clasificado y mal clasificado, y se modelizan según una distribución estadística de ley multinomial. Además de los índices globales, también están muy extendidos los índices por clase. Directamente relacionados con el PA pueden calcularse la exactitud del usuario (filas), y la exactitud de productor (columnas). También existe la posibilidad de calcular Kappa por clase.

$$PA = \frac{1}{N}\sum_{i=1}^{M} n_{ii} = \sum_{i=1}^{M} p_{ii} \qquad (2)$$

$$Ca_{ps} = \sum_{i=1}^{M} P_{i+}P_{+i} \; = \frac{1}{N^2}\sum_{i=1}^{M} n_{i+}n_{+i} \qquad (3)$$

$$\kappa = \frac{P_a - Ca_{ps}}{1 - Ca_{ps}} \qquad (4)$$

Cuando se controlan diferencias, por ejemplo, entre un CDC y la realidad, o entre trabajos realizados por distintas empresas o equipos, entre distintas fuentes de datos, entre situaciones de fechas distintas, etc., utilizando matrices de confusión se puede realizar una comparación cuantitativa y tomar una decisión de semejanza o no por medio de un contraste o test estadístico. En la comparación de valores de *PA* y Kappa, se utiliza un test Z. Por ejemplo, la Ec.5 presenta el estadístico Z para comprobar un valor observado de PA frente a una tolerancia establecida PA_t (H0: $PA=PA_t$) y el mismo test Z puede ser aplicado para probar la misma hipótesis usando el coeficiente Kappa. Pruebas semejantes pueden aplicarse a las exactitudes de productor y usuario, pues se trata de variables binomiales, sin embargo, su uso no está nada extendido. Véase Congalton and Green (2009) para más detalles sobre las técnicas básicas de análisis que son más usuales.

$$Z = \frac{PA - PA_T}{\sqrt{PA(1-PA)/N}} \qquad (5)$$

Pero, un control de calidad sobre la exactitud temática de un producto que se haya realizado utilizando PA o Kappa es muy general, la exactitud global alcanzada en la clasificación puede ser suficientemente buena, pero algunas categorías particulares pueden no estar suficientemente bien clasificadas y tener calidades parciales o de clase bajas, o que no cumplan con las especificaciones.

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

El control descrito no es un control basado en categorías, es un control global basado en la diagonal de la matriz: cuanto mayor son los valores de la diagonal tanto mayor es la exactitud global de la clasificación. Obviamente, los valores fuera de la diagonal indican asignaciones incorrectas (errores), pero se entiende fácilmente que no todos los errores tienen la misma importancia ni oportunidad de ocurrencia. El control por categorías podría realizarse sin mayor problema pues las funciones de distribución de los parámetros por clase son conocidos (por ejemplo, las exactitudes de usuario y productor se comportan como una variable binomial). Sin embargo, estos contrastes no son nada usuales, por ejemplo, en Colgalton y Green (2009), no aparecen explicados los posibles contrastes estadísticos de hipótesis para estos casos.

En este trabajo se propone una nueva forma de controlar la exactitud temática a partir de la matriz de confusión por medio de categorías y tal que permite establecer preferencias de calidad, categoría por categoría, y también permite establecer limitaciones a las confusiones entre categorías. Es decir, la propuesta que se realiza en este trabajo no adopta la perspectiva binomial sobre las categorías (bien o mal clasificada), la novedad reside en que se propone un modelo multinomial ordenado tal que es posible establecer el nivel mínimo de calidad de la categoría de interés, y valores máximos de confusión con las demás categorías. Por ejemplo, considérese la matriz de confusión de 4×4 presentada en la Tabla 1, y ahora considérese que existen unas exigencias sobre los niveles de calidad de cada clase o categoría y sobre las confusiones entre las clases. En esta línea, la Tabla 2 puede ser un ejemplo de especificaciones de un trabajo que se deben contrastar tras su ejecución. Como puede verse, la Tabla 2 establece diferentes exigencias de calidad para cada categoría y también considera la posibilidad de cierto grado de confusión entre ellas. Por ejemplo, en un trabajo concreto podemos suponer que la confusión entre las categorías sin-vegetación y pastos es menos importante que una confusión entre sin-vegetación y forestal. En consecuencia, se puede establecer una gradación u orden de la importancia de los errores de confusión, tal y como se ha indicado en la Tabla 2. Por tanto, a la hora de realizar un control de calidad se deberán tener en cuenta estas situaciones por medio de una probabilidad máxima de ocurrencia para cada caso. Un control con estas condiciones es mucho más riguroso que los comentados anteriormente para el caso de los índices globales PA y Kappa, y que para controles binomiales por clase. Ha de indicarse que la Tabla 2 se ha determinado de manera arbitraria para este ejemplo, pero su determinación en casos reales debe estar basada en la experiencia y en la adecuación al uso del producto.

Tabla 1. Ejemplo de matriz de confusión (fuente Gary *et al.*, 1995)

		Referencia			
		F	P	N	A
Conjunto de datos	F	47	3	0	0
	P	4	40	6	0
	N	0	5	45	0
	A	0	0	2	48
F=forestal, P=pastos, N=sin vegetación, A=agua					

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

Tabla 2. Ejemplo de niveles de calidad exigidos para cada categoría y grado de confusión permitido

Forestal • Al menos 95 % de exactitud de clasificación (≥ 95 %) • Puede existir cierta confusión con pastos, pero no más de 4 % (≤ 4 %) • No puede ser confundida con terrenos sin vegetación o con agua, no más del 1 % (≤ 1 %)
Pastos • Al menos 85 % de exactitud de clasificación (≥85 %) • Puede existir cierta confusión con terrenos sin vegetación, pero no más del 10 % (≤10 %) • Puede existir cierta confusión con forestal o con agua, no más del 5 % (≤5 %)
Sin vegetación • Al menos 90 % de exactitud de clasificación (≥90 %) • Puede existir cierta confusión con pastos, pero no más del 8 % (≤8 %) • No puede existir confusión con forestal o con agua, no más del 2 % (≤2 %)
Agua • Debe estar perfectamente clasificada (≥ 99 %) • No puede estar confundida con ninguna otra categoría, no más del 1 % (≤1 %)

El objetivo de este trabajo es desarrollar la base estadística de un método que permita controlar una matriz de confusión (p.ej. Tabla 1) con restricciones como las indicadas por la Tabla 2. El fundamento estadístico lo conforman dos pilares, en primer lugar, una aproximación multinomial a cada una de las columnas de matriz de confusión y en segundo un test exacto. Conviene indicar aquí que una distribución mutinomial es un caso general de una distribución binomial. En el caso binomial como resultado de un experimento solo caben dos estados o categorías (p.ej. bien clasificado y mal clasificado), pero en el caso multinomial se pueden considerar como opciones válidas todas las celdas de la matriz de confusión.

Este artículo se organiza de la siguiente manera, tras esta introducción se presenta el método de manera general. En la siguiente sección se considerará y discutirá un ejemplo real tomado de las referencias. Finalmente, se exponen unas breves conclusiones.

2. Método estadístico.

2.1 La distribución multinomial

La obtención de una matriz de confusión puede ser considerada como un experimento consistente en la realización de n observaciones o pruebas. Después, esas n observaciones se clasifican en una y sólo una de las M^2 categorías que conforman la matriz. De aquí en adelante, sean $\Gamma_1, \dots, \Gamma_M$ las M clases reales (C_R) y $G_1, \dots, G_M$ (C_P) las clases asignadas en el proceso de clasificación. En consecuencia, la asignación de una observación a la celda (i, j) implica que ha sido clasificado por el procedimiento de clasificación aplicado (p.ej. fotointerpretación, clasificación automática, clasificación OBIA, etc.) a la categoría i en el CDC, pero que realmente pertenece a la categoría j según lo evidencia la fuente de mayor exactitud. Este es el caso de las celdas de tipo C_E. Adicionalmente, la clasificación en una celda (i, i) (diagonal) significa una buena clasificación (celdas de tipo C_C), mientras que una clasificación en una celda $(i, j), i \neq j$ (fuera de la diagonal) implica un error en el proceso de clasificación. Finalmente, n_{ij} indica el número de ítems

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

asignados a la celda (i,j), y n_{+j} es la suma de los ítems que pertenece a la categoría verdadera Γ_j (lo que se denomina valor de la marginal), de forma que $n_{1+} + \cdots + n_{M+} = n$.

Tabla 3. Notación para la matriz de confusión

		Referencia			
		Γ_1	Γ_2	...	Γ_M
Conjunto de datos	G_1	n_{11}	n_{12}	...	n_{1M}
	G_2	n_{21}	n_{11}	...	n_{2M}
	⋮	⋮	⋮	⋮	⋮
	G_M	n_{M1}	n_{11}	...	n_{MM}
Total		n_{+1}	n_{+2}		n_{+M}

Si se considera independencia y aleatoriedad en el proceso de muestreo, una primera aproximación es ver la matriz de confusión, toda entera, como una realización de una distribución multinomial con parámetros: *n* (el tamaño total de la muestra) y un conjunto de M^2 probabilidades. De esta forma, si se realizan los *n* experimentos (observaciones) independientes y se clasifica cada resultado en una de las *M* categorías, se obtendrán las probabilidades $\pi_1, \dots, \pi_{M^2}$, tal que $\pi_1 + \cdots + \pi_{M^2} = 1$, y entonces la función de masa de probabilidad de la multinomial $\mathcal{M}(n, \pi_1, \dots, \pi_{M^2})$ está dada por:

$$P[(X_1 = N_1, \dots, X_{M^2} = N_{M^2}) = \frac{N!}{N_1! \dots N_{M^2}!} \pi_1^{N_1} \dots \pi_{M^2}^{N_{M^2}} \qquad (6)$$

No obstante, considerar la matriz de confusión completa como una única multinomial tampoco es lógico: en este caso no son posibles todos los casos de una multinomial con parámetro *n*, pues no se puede dar un valor n_{ij} en la celda *(i,j)* que sea mayor que n_{+j} y tampoco puede ser nulo, así, por ejemplo, no es posible obtener una multinomial de valor $(n, 0, \dots, 0)$ que, sin embargo, es teóricamente posible en una multinomial con *n* observaciones. Por otra parte, las confusiones ocurren en el CDC y no en el CDR. En la realidad no pueden existir confusiones, los elementos de la realidad son lo que son y no pueden ser otra cosa, por ello n_{+j} es fijo para cada Γ_j. Así, desde la perspectiva del control sobre la verdad terreno, la confusión de una categoría Γ_j sólo ocurre en su columna (un elemento cierto de la realidad es asignado a una categoría equivocada del CDC). Por tanto, se ha de proponer un modelo más ajustado a la realidad. Así, para una matriz de confusión como la presentada en la Tabla 3 nuestra propuesta consiste en considerar cada columna de manera independiente. Desde esta perspectiva cada columna puede ser modelada como una multinomial cuyo primer parámetro es su valor marginal correspondiente (n_{+j}). Esta aproximación al problema permite asumir requisitos de calidad específicos para cada categoría bajo consideración.

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

2.2 Test para el control de calidad

De manera general, en un test estadístico aplicado al control de calidad la hipótesis principal es:

$\mathbb{H}_0$: El producto cumple con las especificaciones (probabilidades de error).

frente a la hipótesis alternativa:

$\mathbb{H}_1$: El producto no cumple con las especificaciones

No obstante, si consideramos la matriz de confusión como una concatenación de *M* distribuciones multinomiales independientes, se puede proponer *M* hipótesis nulas distintas (una para cada clase), y en consecuencia se pueden realizar *M* test estadísticos. En cada uno de ellos se podrán establecer las hipótesis nulas particulares de cada clase. Bajo esta perspectiva, un CDC pasará el control de calidad relativo a unas especificaciones como las expuestas en la Tabla 2 si pasa en todos y cada uno de los controles de calidad (test estadísticos) planteados para cada una de las clases (condición de Y lógico).

2.2 Test multinomial para cada categoría

Se describe a continuación cómo se plantea y se realizan los cálculos del test multinomial que se establece para cada una de las *M* categorías de la matriz de confusión. Para cada clase (columna) se pueden fijar las exigencias de calidad en orden a establecer:

a) La proporción mínima de elementos bien clasificados en la clase *i*. Es decir, en la celda *(i,i)* correspondiente a la diagonal.
b) Hasta un total de $M-1$ especificaciones sobre las asignaciones erróneas, ordenadas por su importancia (probabilidad), tal como se indicará más adelante. Aquí la probabilidad que se asigna es un límite superior (no se permite un porcentaje mayor que el indicado para esa combinación de clases confundidas). De esta forma, la especificación, en forma de multinomial de clase, debe mantener la misma estructura que la columna de la matriz de confusión original o puede agrupar cualquier conjunto de confusiones dentro de la clase, puesto que en cada columna se puede proponer un número distinto de especificaciones. Así, si hay M categorías, en una categoría dada se pueden considerar las M posibilidades (asignación correcta y las $M-1$ confusiones) o llegar sólo al caso de considerar dos categorías (bien clasificado y mal clasificado), que sería el caso binomial.

De esta forma, para cada columna el modelo multinomial propuesto supone una gradación de los posibles errores de confusión, como se realiza en una escala de Likert, desde la opción preferida (correcto) hasta una opción inaceptable. Estas condiciones llevan a asumir que una columna sigue una distribución multinomial de parámetros $(n_{+j};\ \pi_{1j};\ldots;\pi_{k_j,j})$, en donde k_j es la dimensión final tras las especificaciones de la multinomial correspondiente a la columna j (con lo que $k_j \leq M$). Desde un punto de vista práctico, en orden a establecer una especificación estadística más clara, cada una de las M columnas de la matriz es reordenada como un vector multinomial tal que el primer elemento se refiere al caso bien clasificado (el de la diagonal en la matriz original), y

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

el resto de los elementos se ordenan de manera descendente en probabilidades especificadas de error, según se ha indicado. La Tabla 4 presenta cómo quedaría expresada cada una de las multinomiales con las que se ha de trabajar, de forma que, siempre, la primera clase corresponde a los casos correctamente clasificados. Por su parte, la Tabla 5 presenta las especificaciones de calidad, en % y ordenadas según se han presentado las multinomiales en la Tabla 4.

Tabla 4. Multinomiales por clase para el ejemplo de la Tabla 1

Multinomial	$\mathcal{M}_1$ (F)		$\mathcal{M}_2$ (P)		$\mathcal{M}_3$ (N)		$\mathcal{M}_4$ (A)	
	$n_{1,i}$	casos	$n_{2,i}$	casos	$n_{3,i}$	casos	$n_{4,i}$	casos
Bien clasificado	47	F/F	40	P/P	45	N/N	48	A/A
Confusiones	4	F/P	5	P/N	6	N/P	0	A/F
	0	F/N	3	P/F	0	N/F	0	A/N
	0	F/A	0	P/A	2	N/A	0	A/P

Tabla 5. Exigencias de calidad para cada clase expresadas para su uso en forma de hipótesis nula

Clase	F		P		N		A	
	%	casos	%	casos	%	casos	%	casos
Bien clasificado	95	F/F	88	P/P	90	N/N	99	A/A
Confusiones	4	F/P	10	P/N	8	N/P	1	A/F-N-P
	1	F/N-A	2	P/F-A	2	N/F-A		

Por tanto, estamos interesados en detectar si una clase *j* está clasificada realmente como igual o mejor que la condición que se le ha establecido, (lo que implica y se propone un test unilateral para la siguiente hipótesis nula sobre la clase *j*:

$$\mathbb{H}_0: \pi_{1j} \geq \pi_{1j}^0;\ \pi_{2j} \leq \pi_{2j}^0;\ \ldots;\ \pi_{k_j,j} \leq \pi_{k_j,j}^0 \qquad (7)$$

frente a la hipótesis alternativa para la clase M_j:

$\mathbb{H}_1$: al menos una de las condiciones anteriores no es cierta, es decir:

$$\pi_{1j} < \pi_{1j}^0 \ o\ \pi_{2j} > \pi_{2j}^0 \text{ etc.}$$

Deseamos destacar que, de la forma indicada, bajo la hipótesis nula para la clase *j*, la distribución multinomial está completamente fijada, es decir, bajo esta hipótesis la distribución es:

$$\mathcal{M}(n_{+j}, \pi_{1j}^0;\ \ldots;\ \pi_{k_j,j}^0)$$

Consecuentemente, bajo la hipótesis nula la distribución exacta del estadístico del test es conocida y por ello se propone un test exacto. Esto implica que el test se define sin ninguna asunción paramétrica y que puede ser evaluado sin usar aproximaciones. En Mehta y Patel (1983) o en Storer y Choongrak (1990), entre otros, se pueden ver ejemplos de test exactos.

El estadístico $\boldsymbol{T}^j$ del test para la clase *j* es, en este caso, un vector k_j-dimensional $(T_1^j, \ldots, T_{k_j}^j)$ que contiene el número de elementos observados (valores de las celdas de la columna *j*) y ordenados según se ha indicado anteriormente. Así, el primer elemento se corresponde con el valor observado en la celda de la diagonal (elemento bien clasificado), y el resto de los valores son el número de observaciones en cada caso de confusión considerada en esa categoría, pero de manera

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

ordenada con respecto a lo indicado en la hipótesis nula. En consecuencia, el estadístico del test se obtiene de manera directa a partir de la matriz de confusión y para cada categoría M_j puede tener dimensiones distintas.

Para un modelo estadístico dado, el *p-valor* es la probabilidad de obtener un valor como el resultado observado (el estadístico) o más extremo, en el sentido que indique la hipótesis alternativa. Esta probabilidad es calculada asumiendo que $\mathbb{H}_0$ es cierta y, por esta razón, se requiere que $\mathbb{H}_0$ esté fijada previamente y sea conocida. En este caso, para la clase *j* dado que se trata de un test exacto, la probabilidad se calcula de manera exacta a partir de la distribución multinomial, así el *p-valor* es la suma de las probabilidades de ocurrencia de los posibles casos multinomiales $(L_1, \dots, L_{k_j})$ que son iguales o peores que el estadístico observado. Es decir, que cumplen alguna de las siguientes condiciones:

$$\mathrm{L}_1 < \mathrm{T}_1^{\mathrm{j}}$$
$$L_1 = \mathrm{T}_1^{\mathrm{j}} \text{ y } \mathrm{L}_2 > \mathrm{T}_2^{\mathrm{j}}$$
$$L_1 = \mathrm{T}_1^{\mathrm{j}}; \mathrm{L}_2 = \mathrm{T}_2^{\mathrm{j}}, \text{ y } \mathrm{L}_3 > \mathrm{T}_3^{\mathrm{j}}$$
$$\dots, \dots,$$

Para una clase *j* dada se rechazará la hipótesis si el *p-valor* obtenido es menor que el nivel de significación *α*. No obstante, es el caso de una matriz de confusión con *M* clases, se calcularán M *p-valores*. En orden a asegurar que el error de tipo I global no supere el valor de significación dado *α*, la decisión de aceptación de aceptación/rechazo sobre la calidad del conjunto de datos se toma usando el criterio de Bonferroni. En consecuencia, se rechaza una hipótesis nula si al menos el *p-valor*$_j$ de una clase *j* es menor que α/M. Es importante indicar que este control es bastante robusto, ya que no requiere ninguna hipótesis sobre independencia de *p-valores* y no depende de cuántas de las hipótesis nulas son ciertas (Goeman and Solari, 2014). Adicionalmente, se pueden hacer contrastes parciales sobre cualquier subconjunto de categorías de la matriz

Debe indicarse que el procedimiento descrito para el cálculo de los *p-valores* puede implementarse de manera sencilla en paquetes estadísticos como R. Mediante un mecanismo que barra todas las posibilidades de reparto del valor total de la marginal n_{+j} de la clase *j* entre todas las componentes $L_1, \dots, L_{k_j}$ iguales o peores que el valor del estadístico $\boldsymbol{T}^j$, el p-valor se calcula como la suma de todas las probabilidades de esos casos.

3. Aplicación

Se va a aplicar el test propuesto a los datos de la Tabla1. En esta tabla existen 4 categorías y, en consecuencia, se requiere fijar cuatro hipótesis nulas, una para cada categoría o columna, y calcular cuatro *p-valores*. Tomando las especificaciones indicadas en la Tabla 2 y su organización tal y como se ha presentado en la Tabla 5, el proceso es el siguiente:

3.1 Contrastes por categorías

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

Forestal. De acuerdo a las especificaciones, esta categoría puede ser considerada como la siguiente distribución multinomial con tres categorías:

$$(51,\ \pi_F,\ \pi_{F,P},\ \pi_{F,NA}).$$

Y la hipótesis que se deriva de la Tabla 5 es:

$$\mathbb{H}_0: \pi_F = 0.95;\ \pi_{F,P} = 0.04;\ \pi_{F,NA} = 0.01$$
$$\mathbb{H}_1: \pi_F < 0.95\ OR\ \{\pi_F = 0.95\ \text{ AND }\ \pi_{F,P} > 0.04\}$$

En este caso el estadístico del test viene dado por la agrupación correspondiente de la columna F de la Tabla 4:

$$T_F = (47, 4, 0)$$

El *p-valor* se calcula sumando en una multinomial $\mathcal{M}(51;\ 0.95,\ 0.04;\ 0.01)$ las probabilidades de T_F y de todos los casos de combinaciones posibles donde el primer valor es menor de 47, o si el primer valor es 47, el segundo valor es mayor que 4. En este caso: *p-valor*$_1$= *0.1932*.

Pastos. De acuerdo a las especificaciones, la categoría pastos puede ser considerada como una distribución multinomial con parámetros $(48,\ \pi_P, \pi_{P,N}, \pi_{P,FA})$, y la hipótesis planteada a partir de la Tabla 4 es:

$$\mathbb{H}_0: \pi_P = 0.85;\ \pi_{P,N} = 0.10;\ \pi_{P,FA} = 0.05$$
$$\mathbb{H}_1: \pi_P < 0.85\ OR\ \{\pi_P = 0.95\ \text{ AND }\ \pi_{P,N} > 0.10\}$$

El estadístico se obtiene por la agrupación de celdas de la columna *P* de la Tabla 4:

$$T_P = (40, 5, 3)$$

El *p-valor* se obtiene sumando en una multinomial $\mathcal{M}(48;\ 0.85,\ 0.10;\ 0.05)$ las probabilidades de los casos en los que el primer valor es menor que 40, o si el primer valor es 40, que el segundo valor sea mayor que 5. En este cas

o: *p-valor*$_2$*= 0.3255*.

Sin vegetación. De acuerdo a las especificaciones, la categoría sin vegetación puede ser considerada como una distribución multinomial de parámetros $(53, \pi_N, \pi_{N,P}, \pi_{N,FA})$, y la hipótesis planteada a partir de la Tabla 4 es:

$$\mathbb{H}_0: \pi_N = 0.90;\ \pi_{N,P} = 0.08;\ \pi_{N,FA} = 0.02$$
$$\mathbb{H}_1: \pi_N < 0.90\ OR\ \{\pi_N = 0.90\ \text{AND}\ \pi_{N,P} > 0.08\}$$

El estadístico se obtiene por la agrupación de celdas de la tercera columna de la Tabla 4:

$$T_N = (45, 6, 2)$$

El *p-valor* se obtiene sumando en la multinomial $\mathcal{M}(53;\ 0.90,\ 0.08;\ 0.02)$ las probabilidades de los casos en los que el primer valor es menor que 45, o si el primer valor es 45, que el segundo valor sea mayor que 6. En este caso: *p-valor*$_3$*=0.0942*.

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

Agua. La especificación dada por la Tabla 2 para esta categoría indica que no se acepta más de un 2 % de confusión. En consecuencia, sólo se requieren dos categorías, y por ello el modelo estadístico a aplicar es el binomial con parámetros $(48, \pi_A)$ y la hipótesis nula es:

$$\mathbb{H}_0: \pi_A = 0.98$$
$$\mathbb{H}_1: \pi_A < 0.98$$

El estadístico se deriva de la columna A de la matriz de confusión de la Tabla 4. Así $T_A = 48$, y el *p-valor* se obtiene de igual manera que para un test estándar de la distribución binomial como la probabilidad de que $B(48, 0.02)$ sea menor o igual a 48. En este caso: $p\text{-}valor_4=1$.

3.2 Decisión global.

Para la matriz de confusión de la Tabla 1 se han ejecutado cuatro test de hipótesis, y en orden a garantizar el error de tipo I en el nivel de significación global $\alpha = 5\%$ aplicamos la regla de Bonferroni para compensar el número de pruebas que se realizan. En consecuencia, se rechazará la hipótesis de que la matriz de confusión del CDC cumple con la especificación si un $p\text{-}valor_j$ $(j \in [1,4])$ obtenido es menor o igual a $\alpha/4 = 0.0125$. Esta situación no ocurre en este ejemplo, y por ello se concluye que no se han encontrado evidencias de que, globalmente, la matriz de confusión de la Tabla 1 no cumpla con las especificaciones establecidas en la Tabla 2.

4. Conclusiones

Se ha propuesto un nuevo método de control de calidad aplicable a matrices de confusión cuyo objeto es el control de calidad temático. Este método es mucho más potente y completo que los basados en índices globales y por clase al uso. El método permite un control global a partir del control clase por clase y también permite considerar ciertos niveles de confusión dentro de cada clase. Estas capacidades lo dotan de gran versatilidad y de gran potencialidad de aplicación en controles de aceptación de productos temáticos.

El método se basa en la aplicación de la distribución multinomial clase por clase, lo que supone una aproximación ajustada al comportamiento real de las confusiones. Como caso particular el método también permite la aproximación binomial a categorías concretas. Se trata de un método de control de la calidad y no de estimación. Como método de control se fundamenta en un contraste de hipótesis donde en todo momento se controla el error de tipo I, o riesgo del productor.

El cálculo del *p-valor* y, por tanto, la decisión estadística se realiza mediante un test exacto y no sobre aproximaciones, lo que ofrece mayor seguridad. La implementación en herramientas estadísticas como R es inmediata, por lo que su cálculo no conlleva mayor complejidad.

Se ha desarrollado un ejemplo práctico sobre una matriz de confusión de 4×4 clases a la que se han establecido 11 especificaciones, 4 relativas a la calidad de las clases puras y el resto de especificaciones a limitar las posibles confusiones de esas mismas clases. El ejemplo ha mostrado

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018): "Control estricto de matrices de confusión por medio de distribuciones multinomiales", GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.591

que el método permite considerar las confusiones de una manera flexible, como confusión de dos o más categorías, según la conveniencia del caso.

Como línea de trabajo futuro queda analizar el comportamiento del error de tipo II, lo cual se puede realizar por medio de simulaciones.

5. Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación del Reino de España (CTM2015-68276-R).

Referencias bibliográficas

Comber, A., Fisher, P., Brunsdon, C., and Khmag, A. (2012). "Spatial analysis of remote sensing image classification accuracy". *Remote Sensing of Environment*, 127 pp. 237-246. [Consulta: 01-09-2017]. https://doi.org/10.1016/j.rse.2012.09.005.

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

Gary, M., Senseman, C., Bagley, F., Scott, A. (1995). *Accuracy Assessment of the Discrete Classification of Remotely-Sensed Digital Data for Landcover Mapping*. USACERL Technical Report EN-95-04. [Consulta: 01-09-2017].http://www.dtic.mil/get-tr-doc/pdf?AD=ADA296212.

Goeman, JJ.; Solari, A. (2014). "Multiple Hypotheses Testing in Genomics". *Statistics in Medicine*. 33(11) pp. 1946–1978. [Consulta: 01-09-2017]. doi:10.1002/sim.6082.

Liu, C., Frazier, P., Kumar, L. (2007). "Comparative assessment of the measures of thematic classification accuracy". *Remote Sensing of the Environment*, 107(4) pp. 606-616. [Consulta: 01-09-2017]. doi:10.1016/j.rse.2006.10.010.

Mehta, C.R., Patel, N.R. (1983). "A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables". *Journal of the American Statistical Association*, 78(382) pp.427-434.

Nishii, R., Tanaka, S. (1999). "Accuracy and inaccuracy assessments in land-cover classification". *Geoscience and Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing*, 37(1) pp. 491-498. [Consulta: 01-09-2017].DOI: 10.1109/36.739098.

Storer, B.E., Choongrak, K. (1990). "Exact properties of some exact test statistics for comparing two binomial proportions". *Journal of the American Statistical Association*, 85(409) pp.146-155

Veregin, H. (1989). *A Taxonomy of error in spatial databases. Technical papers of the National Center for Geographic Information and Analysis*. Santa Barbara, USA. University of California. [Consulta: 01-09-2017]. http://ncgia.ucsb.edu/technical-reports/PDF/89-12.pdf.