Comparación de dos métodos para evaluar los intervalos de confianza de índices de fiabilidad cartográfica

Jean-François Mas[1]

[1] Centro de Investigaciones en Geografía Ambiental,
Universidad Nacional Autónoma de México,
Antigua Carretera a Pátzcuaro 8701, Col Ex-Hda de San José de La Huerta,
58190 Morelia, Mich, México
jfmas@ciga.unam.mx

Abstract. A model that enables user to simulate classification error on ”true” maps was developed in order to elaborate error-contaminated maps. Both type of maps, true and error-contaminated, were used to evaluate the accuracy assessment procedure using a modeling approach. A large number of random stratified samples were obtained and used to compute the estimates of overall, producer and user accuracy along with their confidence intervals at the 95% confidence level, using two methods: the normal approximation to the binomial distribution and bootstraping. Both methods tend to overestimate accuracy, because the confidence intervals contained the true value of the parameter in less than 95% of the cases and the true value lies below the lower bound of the interval in most of the cases. Producer accuracy, was largely over-estimated in case of a rare class confused with a large one. In such case, increasing the size of the sample was not effective and stratification based on information about likely confused areas is recommended. Such cases of over-estimation of map accuracy could be common when mapping rare classes among dominant classes as, for example, land cover change versus persistence areas.

Keywords: Thematic maps, accuracy assessment, uncertainty mapas temáticos, evaluación de la fiabilidad, incertidumbre

1. Introducción

La evaluación de la fiabilidad de los mapas temáticos se realiza comúnmente a través de la comparación de la categoría del mapa con la categoría asignada a sitios de verificación, con base en el análisis de imágenes de muy alta resolución espacial o de visita de campo. Para seleccionar estos sitios de referencia, los diseños de muestreo más utilizados son los muestreos aleatorios simples, sistemáticos y aleatorios estratificados, entre otros (Stehman, 1995; Stehman y Czaplewski, 1998).

Generalmente, se elabora una matriz de confusión que es una tabla de doble entrada en la cual se reporta el número de sitios de verificación para cada combinación de categoría en el mapa y en los datos de verificación. En caso de muestreos que no conservan la proporción entre el número de sitios por categoría del mapa y la superficie de estas categorías como la mayoría de los muestreos estratificados, es necesario llevar a cabo un ajuste de los valores de la matriz antes de calcular los índices (Card, 1982). Es también importante evaluar la certidumbre del valor estimado de estos índices a través la construcción de un intervalo de confianza. Por ejemplo, sería muy díficil tomar una decisión sobre la conveniencia de utilizar un mapa cuya fiabilidad global es estimada en 75%, con un intervalo de confianza de 20% ya que la fiabilidad varía probablemente entre 55 y 95%.

El objetivo de este estudio, es evaluar la exactitud de los intervalos de confianza de índices de fiabilidad obtenidos con dos enfoques. El primero se basa en la aproximación normal a la distribución binomial y el segundo en las técnicas de remuestreo conocidas como Bootstrap.

Table 1: Matriz de confusión en número de sitios de verificación.

	Referencia						
Mapa	1	2	...	j	...	q	Suma
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	n_{1+}
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	n_{2+}
3	n_{31}	n_{32}	...	n_{3j}	...	n_{3q}	n_{3+}
...	...	...	...	...	...	...	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	n_{i+}
...	...	...	...	...	...	...	...
q	n_{q1}	n_{q2}	...	n_{qj}	...	n_{qq}	n_{q+}
Suma	n_{+1}	n_{+2}	...	n_{+j}	...	n_{+q}	

2. Metodología

Se utilizaron el programa Dinamica EGO, una plataforma gratuita para modelación espacial (www.csr.ufmg.br/dinamica) y el paquete de análisis estadístico R (www.r-project.org) así como herramientas para la evaluación de la fiabilidad cartográfica previamente desarrolladas utilizando estos programas (Mas et al., 2014 y 2015).

La evaluación de los dos métodos de cálculo de los intervalos de confianza se basa en un enfoque de modelación. En una primera etapa, se elaboró un modelo que simula el error en un mapa utilizando las herramientas disponibles en DINAMICA para modelar cambios de cubierta del suelo. El mapa con error se obtiene modificando el mapa original considerado con la información libre de error (”verdad terreno”). Una matriz de Markov permite controlar la cantidad de cada tipo de error, entendiendo por tipo de error la confusión entre dos categorías particulares. La distribución espacial del error es controlada con base en mapas de probabilidad y celulares autómatas.

En una etapa siguiente, se calculó el valor exacto de los índices de fiabilidad con base en la comparación de la totalidad del mapa original y del mapa con el error simulado. Estos índices son i) la fiabilidad global (proporción de área total del mapa correctamente clasificada), ii) la fiabilidad del usuario que es la proporción de sitios mapeados como una categoría dada que fue correctamente clasificada y es por lo tanto relacionada con los errores de comisión y, iii) la fiabilidad del productor que es la proporción de sitios de referencia de una categoría que es correctamente clasificada en el mapa y es relacionada con los errores de omisión.

Luego, se realizó un muestreo aleatorio estratificado por las categorías del mapa con error. Para cada uno de los sitios de verificación se obtiene la información del mapa bajo evaluación (mapa con error simulado) y la información de mapa original (”verdad terreno”). Se elaboró una matriz de confusión (Tabla 2) a la cual se aplicó una corrección del sesgo relacionado con el muestreo siguiendo los pasos descritos a continuación.

En la matriz de confusión bruta (tabla 2), cada celda representa el número n_{ij} de sitios de verificación perteneciendo a la categoría i y j respectivamente en el mapa y en la ”verdad terreno”. Dependiendo del tipo de muestreo utilizado, el número de sitios para cada categoría del mapa no es necesariamente proporcional a la superficie de estas categorías en el mapa. En otras palabras, ciertas categorías son ”sub-representadas” en el muestreo respecto a su representación en el mapa, otras, al contrario, son sobre-representadas en el muestreo. Card (1982) propone un método para ajustar la matriz bruta reemplazando el valor n_{ij} de cada celda por $\hat{p}_{ij}$ que es un estimador no sesgado de la proporción del área (ecuación 1).

$$\hat{p}_{ij} = \frac{\pi_i n_{ij}}{n_{i+}} \quad (1)$$

donde π_i es la proporción de la categoría i en el mapa, n_{ij} el número de sitios de verificación cartografiados como i pero identificados como j en los datos de referencia, y n_{i+} el número de sitios mapeados como i en el mapa. En la matriz ajustada, cada elemento $\hat{p}_{ij}$ representa la probabilidad que una área seleccionada de forma aleatoria sea clasificada como categoría i siendo realmente j en los datos de referencia. La suma de las celdas de cada fila $\hat{p}_{i+}$ es por lo tanto igual a π_i, proporción de la categoría i en el mapa. La matriz ajustada permite compensar los sesgos relacionados con el muestreo.

Se calcularon los siguientes índices de fiabilidad y sus respectivos intervalos de confianza a 95% i) con las ecuaciones propuestas por Card (1982), las cuales se basan en el supuesto de una distribución normal del error (ver ecuaciones en anexo) y, ii) con bootstrap. En el método de bootstrap, los sitios de verificación de cada estrato se sometieron a 500 remuestreos con reemplazamiento y se elaboran igual número de matrices de confusión que se corrigen por el sesgo del muestreo estratificado con la ecuación 1. Se calculan los mismos índices de fiabilidad para cada repetición de bootstrap. Para evaluar los intervalos de confianza de cada índice, se utilizó el método de percentil que consiste en usar los 2.5 y 97.5 percentiles de la distribución de bootstrap como los límites de 95% del intervalo de confianza. Se realizaron 200 muestreos y se calculó la frecuencia con la cual cada índice cae adentro de los intervalos de confianza respectivos.

3. Resultados

Se elaboró un mapa en formato raster con cuatro categorías al cual se aplicaron cambios distribuidos de forma aleatoria. En el mapa domina la categoría 1 (85% del área total). Las categorías presentan diferentes niveles de error. Se realizó una tabulación cruzada entre la totalidad del mapa original y el mapa con error (Tabla 2). La fiabilidad global del mapa simulado es 92.1%. Las categorías presentan valores de fiabilidad entre 64 y 97% (tabla 3).

Se realizaron muestreos aleatorios estratificados con 50 sitios de verificación por categoría, como sugerido por Lillesand et al. (2008). Se calcularon los índices de fiabilidad y sus intervalos de confianza con ambos métodos. Esta operación se repitió 500 veces y se calculó cuantas veces los verdaderos valores de los índices estaban comprendidos en los intervalos de confianza. Como se puede observar en la tabla 4, la proporción de casos en el cual el verdadero valor del índice está dentro del intervalo de confianza está por debajo del valor esperado (95 %) para cinco de nueve índices. En estos casos, el verdadero valor del índice se encuentra por debajo del intervalo de confianza en la mayoría de los casos, es decir que las estimaciones tendieron a ser optimistas. Ambos métodos dieron resultados similares. En el caso de la fiabilidad del productor de la categoría 4, los intervalos de confianza dan una evaluación muy optimista de la fiabilidad. El valor real del índice (79.4%) es inferior al límite inferior del intervalo de confianza en 50% de los casos.

La fiabilidad del productor es la proporción de sitios de referencia de una categoría que es correctamente clasificada en el mapa. A diferencia del caso de la fiabilidad del usuario, es por lo tanto más dificil controlar el número de sitios de verificación en los cuales se basa la estimación. En el caso de la categoría 4 que tiene una área muy reducida, el muestreo es capaz de identificar casos en el cual esta categoría están correctamente clasificas o bien presente errores de comisión. No obstante, es poco probable que se selecionen sitios de las categorías 1 y 2 en el mapa que pertenezcan realmente a la categoría 4. En particular, es muy improbable obtener sitios que representen las confusiones ”categoría 1 en mapa, 4 en realidad” o bien

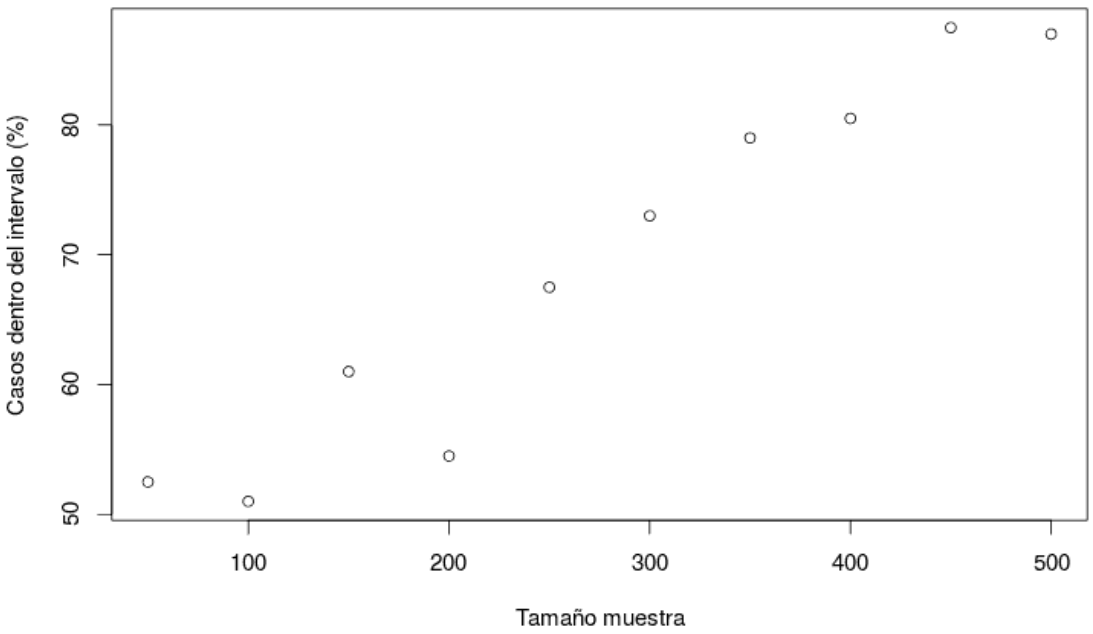

Figure 1: Proporción de casos en los cuales la fiabilidad del productor de la categoría 4 se encuentra dentro del intervalo de confianza en función del número de sitios de verificación para las categorías 1 y 2

”categoría 2 en mapa, 4 en realidad”. Las probabilidades de obtener un sitio de verificación que represente estas confusiones son respectivamente 0.042 (35/8274) y 0.011 (12/1124) como se puede deducir de la tabla 2. Por estas razones, el muestreo falló en representar los errores de omisión de la categoría 4. 50% de los muestreos no tiene ningún sitios que presentan las dos confusiones mencionadas anteriormente y, en estos casos, la fiabilidad del productor de la categoría 4 se estimó en 100%.

Se evaluó cual sería el incremento del tamaño de muestra para las categorías 1 y 2 necesario para mejorar la evaluación de los valores de la fiabilidad del productor de la categoría 4. Como se puede observar en la figura 1, se necesitaría más de 400 sitios de verificación para que el intervalo de confianza sea una representación filedigna del valor del índice en más de 80% de los muestreos.

4. Discusión

La aproximación normal a la distribución binomial, sustenta un rango de pruebas y métodos estadísticos, incluyendo el cálculo de intervalos de confianza precisos. Estos métodos tienen en común la suposición de que la distribución de probabilidad del error sobre una observación es

Table 2: Tabulación cruzada de los mapas completos

	Referencia			
Mapa	1	2	3	4
1	8039	200	0	35
2	296	736	80	12
3	69	0	253	0
4	99	0	0	181

Table 3: Índices de fiabilidad del mapa

Índice de Fiabilidad	Valor en %
Fiabilidad global	92.1
Fiabilidad del usuario categoría 1	97.2
Fiabilidad del usuario categoría 2	65.5
Fiabilidad del usuario categoría 3	78.6
Fiabilidad del usuario categoría 4	64.6
Fiabilidad del productor categoría 1	94.5
Fiabilidad del productor categoría 2	78.6
Fiabilidad del productor categoría 3	76.0
Fiabilidad del productor categoría 4	79.4

normalmente distribuida. Sin embargo, a menudo, no se cumple y existen métodos alternativos para estos casos. Al contrario, el método de bootstrap no se basa en ninguna suposición sobre la distribución del error. Sin embargo, en este estudio, no se encontraron diferencias importantes en los resultados obtenidos por ambos métodos. El valor real del índice fue en general ligeramente sobre-estimado. Para una de las categorías, el error en la evaluación de la fiabilidad del productor fue mucho más importante. No obstante, este error no se debe al método de cálculo del índice y su intervalo de confianza sino al muestreo. Este muestreo, estratificado por las categorías del mapa, no logró representar los errores de omisión de una categoría rara que se confunde con otra categoría muy abundante. Es importante notar que esta configuración podría ser bastante frecuente en cartografía. Por ejemplo, un tipo de cubierta del suelo poco común que tiende a confundirse con otro mucho más representado, podría presentar este tipo de problema. Otro caso son los mapas de cambio de cubierta / uso del suelo, donde las áreas de cambio son generalmente reducidas e inmersas en amplias áreas sin cambio.

En tales casos, el aumento del tamaño del muestreo resulta costoso y poco eficiente y una mejor opción sería estratificar el muestreo utilizando áreas donde se podría esperar que el error de omisión sea importante. Para cartografía de la vegetación, se podrían utilizar áreas potenciales de distribución de los tipos de vegetación mapeados. Para análisis de cambio, mapas de alta probabilidad de cambio o áreas en las cuales el análisis de las imágenes es más problemático (áreas de sombra o de confusión espectral por ejemplo). Finalmente, es importante notar que, en este análisis, no se tomaron en cuenta otros elementos que en ejercicios reales podrían afectar los resultados de la evaluación de la fiabilidad. En particular, la falta de

Table 4: Porcentaje de casos en el cual el verdadero valor del índice se encuentra dentro del intervalo de confianza

Índice de Fiabilidad	Binomial	Bootstrap
Fiabilidad global	84.5	85.0
Fiabilidad del usuario categoría 1	81.0	81.0
Fiabilidad del usuario categoría 2	96.5	96.0
Fiabilidad del usuario categoría 3	93.5	96.0
Fiabilidad del usuario categoría 4	96.5	96.0
Fiabilidad del productor categoría 1	96.0	96.5
Fiabilidad del productor categoría 2	72.0	73.5
Fiabilidad del productor categoría 3	89.0	89.0
Fiabilidad del productor categoría 4	48.0	50.0

fiabilidad de los datos de verificación o la ambigüedad de asignar una categoría a un sitio de verificación en ecotonos o en paisajes complejos.

5. Conclusiones

En este estudio, se mostró que el análisis de la fiabilidad de mapas tiende a dar una evaluación optimista de la calidad de los mapas aunque esté basado en un diseño de muestreo probabilístico y un procesamiento de la matriz de confusión adecuado (Olofson et al., 2014). En particular, resulta muy difícil detectar errores de omisión de una clase rara confundida con otra mucho más extensa. Para estos casos, es conveniente estratificar el muestreo utilizando información adicional que permita concentrar el esfuerzo de muestreo en áreas donde los errores sean más frecuentes. En futuras investigaciones esperamos desarrollar una herramienta que permita calcular los índices de fiabilidad y sus intervalos de confianza para este tipo de estratificación.

Agradecimientos

Este trabajo se realizó con el apoyo del proyecto SEP-CONACYT 178816 *¿Puede la modelación espacial ayudarnos a entender los procesos de cambio de cobertura/uso del suelo y de degradación ambiental?*

Referencias

Card, D.H. Using known map category marginal frequencies to improve estimates of thematic map accuracy. **Photogrammetric Engineering and Remote Sensing**, v. 48, p. 431-439, 1982.

Lillesand, T.M., Kiefer, R. W., Chipman, J. W. **Remote sensing and image interpretation**. Wiley, 2008. 768 p.

Mas, J.F.; Pérez-Vega, A.; Ghilardi, A.; Martínez, S.; Octavio Loya-Carrillo, J.; Vega, E. A Suite of Tools for Assessing Thematic Map Accuracy. **Geography Journal**, v. 2014. Disponível em: <http://downloads.hindawi.com/journals/geography/2014/372349.pdf> Acesso em: 24 oct. 2016.

Mas, J-F; Pérez-Vega, A.; Ghilardi, A.; Martínez, S.; Octavio Loya-Carrillo, J.; Vega, E. Unas herramientas de uso libre para evaluar la fiabilidad temática de datos espaciales. In: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 17., 2015, João Pessoa-PB. **Anais**. São José dos Campos: INPE, 2015. Artigos, p. 1020-1026. On-line. ISBN 978-85-17-0076-8. Disponível em: <http://www.dsr.inpe.br/sbsr2015/files/p0191.pdf>. Acesso em: 10 nov. 2016.

Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. **Remote Sensing of Environment**, v. 148, p. 42-57, 2014.

Stehman, S.V. Thematic map accuracy assessment from the perspective of finite population-sampling. **International Journal of Remote Sensing**, v. 16, p. 589-593, 1995.

Stehman, S.V.; Czaplewski, R.L. Design and analysis for thematic map accuracy assessment: Fundamental principles. **Remote Sensing of Environment**, v. 64, p. 331-344, 1998.

ANEXOS

En este anexo, se muestran las ecuaciones propuestas por Card (1982) para calcular la fiabilidad global, del usuario y del productor así como sus intervalos de confianza utilizando la aproximación normal a la distribución binomial. Los cálculos se basan en la matriz ajustada con base en la ecuación 1 con el fín de compensar los sesgos relacionados con el muestreo. El cálculo de la fiabilidad global, del usuario y del productor se llevan a cabo con base en las ecuaciones 2, 3 y 4.

La fiabilidad global $\hat{O}$ (proporción de área correctamente clasificada) se calcula sumando los elementos de la diagonal de la matriz ajustada (ecuación 2).

$$\hat{O} = \sum_{K=1}^{q} \hat{p}_{kk} \tag{2}$$

donde q es el número de categorías.

Las fiabilidades del usuario $\hat{U}_i$ y del productor $\hat{P}_j$ son respectivamente calculadas utilizando las ecuaciones 3 y 4. La fiabilidad del productor, relacionada con los errores de omisión, es la proporción de sitios de referencia de una categoría que es correctamente clasificada en el mapa. La fiabilidad del usuario, relacionada con los errores de comisión, es la proporción de sitios mapeados como una categoría dada que fue correctamente clasificada.

$$\hat{U}_i = \frac{\hat{p}_{ii}}{\hat{p}_{i+}} \tag{3}$$

$$\hat{P}_j = \frac{\hat{p}_{jj}}{\hat{p}_{+j}} \tag{4}$$

Para muestreos estratificados, los intervalos de confianza de las estimaciones de la fiabilidad global, del productor y del usuario se calculan con base en las ecuaciones 5 a 7 (Card, 1982):

$$HCI_{\hat{O}} = z\sqrt{\sum_{i=1}^{q} \frac{\hat{p}_{ii}(\pi_i - \hat{p}_{ii})}{n_{i+}}} \tag{5}$$

donde $HCI_{\hat{O}}$ es el medio intervalo de confianza de la fiabilidad global y z el número de desviación estándar de una distribución normal (para un nivel de confianza de 95%, z = 1.96).

$$HCI_{\hat{U}_i} = z\sqrt{\frac{\hat{p}_{ii}(\pi_i - \hat{p}_{ii})}{\pi_i^2 n_{i+}}} \tag{6}$$

donde $HCI_{\hat{U}_i}$ es el medio intervalo de confianza de la fiabilidad del usuario de la categoría i.

$$HCI_{\hat{P}_j} = z\sqrt{\hat{p}_{jj}\hat{p}_{+j}^{-4}\left[\hat{p}_{jj}\left(\sum_{i\neq j}^{q} \hat{p}_{ij}(\pi_i - \hat{p}_{ij})/n_{i+}\right) + (\pi_j - \hat{p}_{jj})(\hat{p}_{+j} - \hat{p}_{jj})^2/n_{j+}\right]} \tag{7}$$

donde $HCI_{\hat{P}_j}$ es el medio intervalo de confianza de la fiabilidad del productor de la categoría j.