Análisis de errores en modelos medioambientales de variables discretas

Álvaro González Dueñas*

Recibido el 12 de mayo de 2014; aceptado el 9 de julio de 2014

Abstract

Some of the variables that try to represent the environment are difficult to measure, so they are usually estimated using models based on other, spreading the errors in the source data. This article is a bibliographic compilation about the most important aspects to consider analyzing the error propagation in models with spatial inputs and discrete outputs data on environmental variables.

Because these models usually have input as discrete variables, their error sources influence is also analized by error propagation. The most important error sources of this type data input are proper completion of the classes, identification of the category —thematic and conceptual errors—, tracing its edges —cartographic precision and accuracy— and scale, allow for variables in the error propagation analysis.

Once we know the source of the error, then different studies of sensitivity analysis of some models that can serve as reference to the analysis of other discrete environmental variables. Monte Carlo is shown as a suitable method for the analysis of error propagation for discrete variables, but we have not find literature that compare different methods for the same data set or model. Certain peculiarities of each data model —raster and vector— and its influence on the results are discussed.

Key words: *model error propagation, discrete data, environment.*

Resumo

Algumas das variáveis que tratam de representar o meio ambiente são de difícil medida, por isso frequentemente são estimadas mediante outros modelos, propagando-se o erro dos dados de origem. O presente artigo é uma recompilação bibliográfica dos aspectos mais importantes a serem considerados ao se analisar a

* Ingeniero de Montes, ETSI, Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, Campus Sur UPM (28031), España. Correos electrónicos: alvaro@iies.es/ alvaro.gonzalez.duenas@alumnos.upm.es

propagação do erro em modelos com dados de entradas espaciais e saídas em variáveis discretas de meio ambiente.

Pelo fato destes modelos geralmente terem como entrada variáveis discretas, também se abordada a influência de suas fontes de erros na propagação nos modelos cartográficos. As fontes mais importantes de erro nos dados de entrada neste tipo são a adequada determinação das classes, a identificação da categoria (erro temático) e conceitual, a localização de suas bordas (precisão) e exatidão cartográfica e, a escala, considerando-se como variáveis nas análises de propagação de erros.

Uma vez conhecida a fonte de erro, posteriormente se analisam diferentes trabalhos de análises de sensibilidade de alguns modelos que podem servir como referência para o estudo de outras variáveis discretas de meio ambiente. Monte Carlo se mostra como um método idôneo para a análise de propagação de erros para variáveis discretas, ainda que não se tenha encontrado bibliografia que compare diferentes métodos para um mesmo conjunto de dados no modelo. Também se analisam certas particularidades de cada modelo de dados *raster* e vetorial e sua influência no resultado.

Palavras chave: *modelo, propagação de erros, dados discretos, meio ambiente.*

Resumen

Algunas de las variables que tratan de representar el medio ambiente son de difícil medida, por lo que suelen estimarse en función de otras mediante modelos, propagando el error de los datos de origen. El presente artículo es una recopilación bibliográfica de los aspectos más importantes a considerar al analizar la propagación del error en modelos con datos de entradas espaciales y salidas discretas sobre variables medioambientales.

Debido a que estos modelos suelen tener datos de entrada en forma de variables discretas, también se aborda la influencia de sus fuentes de error en su propagación por los modelos cartográficos. Las fuentes de error más importantes en los datos de entrada de este tipo son la adecuada determinación de las clases, la identificación de la categoría —error temático y conceptual—, la localización de sus bordes —precisión y exactitud cartográfica— y la escala, considerándose como variables en los análisis de propagación de errores.

Una vez conocida la fuente del error, posteriormente se analizan diferentes trabajos de análisis de sensibilidad de algunos modelos que pueden servir como referencia para el estudio de otras variables medioambientales discretas. Monte Carlo se muestra como un método idóneo para el análisis de propagación de errores para variables discretas, aunque no se ha encontrado bibliografía que compare diferentes métodos para un mismo juego de datos ni modelo. Se analizan ciertas particularidades de cada modelo de datos —raster y vectorial— y su influencia en el resultado.

Palabras clave: *modelo, propagación de errores, datos discretos, medio ambiente.*

Introducción

Durante las últimas décadas ha aumentado la necesidad de mejorar el conocimiento del medio ambiente para optimizar la gestión del territorio y sus recursos. Para ello resulta de especial importancia su distribución espacial. El medio ambiente es un escenario en el que se desarrollan muchas interrelaciones complejas entre muchos de sus elementos, por lo que resulta habitual cartografiar variables medioambientales estimadas a partir de modelos que toman como entrada variables más sencillas. Con este fin, se han desarrollado modelos, tanto para comprender la situación actual de cada recurso (por ejemplo, crecimiento de biomasa forestal a partir de datos de diámetro de fuste, alturas de árboles y calidad de estación, o viabilidad de huevos de truchas a partir de la temperatura del agua), como para predecir su tendencia futura (por ejemplo, predecir la mortalidad de árboles tras una lluvia ácida a partir de datos de su ph o predecir el grado de defoliación de eucaliptos durante la primavera a partir de porcentaje de parasitación de ootecas del defoliador al comienzo de ésta). Algunos modelos se basan en el conocimiento teórico del funcionamiento del recurso, pero otros se basan en la combinación de expresiones empíricas (Finke *et al*., 1999).

Muchos profesionales del medio ambiente toman sus decisiones apoyados en SIG que resultan imprecisos, y lo que es aún más importante, en muchas ocasiones estos profesionales no se ocupan o no conocen su imprecisión. La imprecisión no tiene necesariamente su origen en el propio *software*, sino en el imperfecto conocimiento del fenómeno geográfico o bien en la imperfección de los datos disponibles (Qi *et al*., 2011).

Los datos son por definición generalizaciones de la realidad (Ehlschlaeger, 2000), y mucho más las variables categóricas de origen natural que cualquier otra de origen antrópico. Huang *et al*. (2009), sostienen que todas las capas SIG utilizadas como variables predictivas contienen algún nivel de error, y que además este error se propaga por el proceso de clasificación del modelo según su grado de sensibilidad a dicho error. Por ello, cualquier dato espacial procedente de un modelo siempre contiene error, el que necesita ser cuantificado para poder usarlo de manera adecuada. Además, Finke *et al*. (1999), determinan que la cuantificación del error o incertidumbre de modelos predictivos tiene la utilidad de determinar la relevancia, significación o fiabilidad de los diferentes escenarios modelizados.

El modelado de fenómenos espaciales es mucho más complejo que el modelado de variables sin componente espacial ya que, además de determinar la clase, cada modelo también debe establecer la cantidad de polígonos, además de sus límites, ejes y nodos (Goodchild *et al*., 1992). Esto se ve reflejado cuando la toma de datos

de flora, suelo, etcétera, por diferentes operadores ofrece el mismo resultado si únicamente se les pide determinar la clase de puntos concretos del territorio, pero no ofrece el mismo resultado cuando se les pide cartografiarlo, es decir añadir componente espacial.

También hay que considerar que algunos fenómenos geográficos son difusos por naturaleza, tanto en la identificación de la clase, como por la ubicación de sus límites (Qi *et al*., 2011), lo que también induce error en la salida del modelo, sin que sea atribuible a éste.

Objetivo

El objetivo del presente trabajo es realizar una recopilación bibliográfica que permita conocer qué aspectos hay que considerar y qué metodologías se han utilizado para analizar los errores y su propagación en modelos medioambientales con componente geográfica de variables discretas. La utilización práctica de este conocimiento es colaborar en las estrategias de gestión y reducción del error (Huang *et al.*, 2009). Las etapas a seguir para abordar el problema son las siguientes (Veregin, 1989; Moreno Ruíz *et al.*, 2001, y Huang *et al.*, 2009):

I. Datos de origen:
   i. Identificación de todas las fuentes de incertidumbre de los datos de entrada y parámetros del modelo.
   ii. Caracterización de la incertidumbre mediante su función de densidad de probabilidad.
II. Modelo de datos: propagación de errores, incertidumbre originada por el modelo (Análisis de Incertidumbre) e incertidumbre que ocasionan los datos de entrada en el resultado final (Análisis de Sensibilidad) (Moreno Ruíz *et al.*, 2001, y Huang *et al.*, 2009):
   i. Selección de modelos de errores adecuados para simular dichas incertidumbres.
   ii. Modelización de dichas incertidumbres.
   iii. Propagación de la incertidumbre a través del modelo.
III. Gestión y reducción de errores mediante técnicas adecuadas.

Sin embargo, Satelli *et al*. (2000) definen el análisis de sensibilidad como el estudio de las diferentes fuentes de variación de la salida de un modelo (continuo o discreto) y cómo depende el modelo de los datos de entrada. Por lo tanto, el análisis de sensibilidad del modelo no se ocuparía de las características de la incertidumbre de los datos de entrada (Huang *et al*., 2009), sino que se basaría en las siguientes premisas:

I. Asumir que existe cierto nivel de incertidumbre.
II. Contar con un modelo de error adecuado.
III. Elegir un método de análisis de sensibilidad adecuado.

Los análisis de sensibilidad de modelos matemáticos con variables de entrada y salida continuas que incluyen operaciones SIG sencillas se realizan con modelos matemáticos formales basados en series de Taylor. La misma técnica no puede utilizarse cuando la salida y/o alguna de las variables de entrada es discreta, ya que no son diferenciables (Huang *et al.*, 2009). Según estos autores, la cartografía relacionada con variables medioambientales se expresa frecuentemente en variables discretas, tanto los datos de entrada como los modelos, siendo la propagación de sus errores menos estudiada que las continuas, probablemente por resultar menos sensibles a los datos de entrada.

Error en los datos de entrada en el modelo

El cartografiado de una variable discreta de un mismo territorio por dos observadores diferentes muchas veces da resultados parecidos pero diferentes (Goodchild, 1992). Por eso, el dato tomado por un observador se supone que es simplemente un ejemplo de una población de datos distorsionados de la misma realidad.

En todo trabajo de representación de la realidad, se asume la premisa de que resulta imposible realizar una representación perfecta (Ehlschlaeger, 2000), y más aún al trabajar con la realidad natural que con la realidad antrópica. Por ello, el logro más importante en la investigación de la incertidumbre es lograr la exactitud adecuada de los datos para representar la parte de la realidad más relevante para la aplicación estudiada (Ehlschlaeger, 2000), aun sabiendo que una variable discreta no puede tener un número ilimitado de clases (Qi *et al.*, 2011).

Las fuentes de error más importantes en la cartografía de variables medioambientales discretas son la adecuada determinación de las clases (Qi *et al.*, 2011), la identificación de la categoría —error temático y conceptual— (Ehlschlaeger, 2000), la localización de sus bordes —precisión y exactitud cartográfica— (Goodchild *et al.*, 1992; Huang *et al.*, 2009, y Ehlschlaeger, 2000) y la escala (Finke *et al.*, 1999). Por ello, resulta habitual considerarlas como variables de entrada en los análisis de propagación de errores. Finke *et al.* (1999) estudia la influencia de la escala de los datos de entrada como otro elemento importante en la propagación de errores en el modelo. Para ello, utiliza las variables escala, tamaño mínimo de polígono para vectorial y tamaño de celda para raster, obteniendo que la escala y los errores en la clasificación categórica tienen una elevada influencia en la incertidumbre del modelo.

Recientemente Qi *et al.* (2011) ha señalado que estas fuentes de incertidumbre no son sólo atribuibles al deficiente conocimiento de la realidad física, sino que pueden ser implícitos a la propia variable, pues la frontera entre clases puede estar difuminada en la propia realidad.

Los datos de entrada pueden tomarse en formato vectorial o raster (Goodchild, 1992). Esta decisión no tiene relación con la realidad, pero debe tomarla el analista y, como se verá, resulta de especial importancia e influencia en el resultado del comportamiento del modelo. A continuación se analizan las fuentes de error propias de cada uno de los formatos poniendo especial énfasis en su utilización para datos discretos de variables medioambientales. No se valora ni pondera la influencia de cada una de ellas pues eso depende de cada aplicación concreta.

Errores de los datos de entrada en formato vectorial

Las fuentes de error de los datos de entrada de un modelo vectorial discreto son atribuibles tanto a las limitaciones del propio modelo de datos vectorial para recoger la realidad geográfica, como al proceso de cartografiado:

I. Aquéllas cuyo origen es debido al modelo de datos.
    i. Los bordes de polígonos son una simplificación de la realidad de la zona de transición entre clases discretas.
    ii. Las clases previstas no se ajustan totalmente a la realidad física de ciertas zonas, lo que lleva a representar simplificaciones de la realidad. La clásica clasificación categórica booleana conlleva a cometer errores atribuibles a la propia clasificación, también denominado "efecto prototipo" (Qi *et al.*, 2011).
    iii. La clase asignada a un polígono realmente no es una propiedad homogénea de todo el polígono. De hecho, es habitual encontrar leyendas con términos como "principalmente", "sustancialmente incluye", "mezcla de", etcétera. Para una correcta clasificación categórica, el concepto que representa cada clase debería ser excluyente del resto.
II. Aquéllas cuyo origen está en el proceso de cartografiado.
    i. Delimitación de bordes de polígonos inexacta.
    ii. Etiquetado de polígonos inexacto.
    iii. Las pequeñas islas de una clase dentro un polígono grande de otra clase suelen ser omitidas en alguna de las fases del proceso de cartografiado, especialmente las de tamaño inferior a la "unidad mínima de cartografiado" debido a la escala del mapa.
    iv. Los bordes de polígonos suelen ser deliberadamente suavizados para crear un efecto cartográfico más estético.

Por esto, muchos mapas, principalmente los que proceden de la digitalización de una versión en papel, suelen crear una falsa apariencia de exactitud eliminando cualquier indicio de incertidumbre (Goodchild *et al.*, 1992).

Errores de los datos de entrada en formato raster

El modelo raster es, en algunos aspectos, mejor que el vectorial al registrar de manera mas adecuada la variación continua de la realidad, tanto para variables discretas, como continuas. Además, conceptualmente resulta más fácil estimar para una celda *i* la probabilidad de que la realidad sea de la clase *k* en un modelo raster que para el conjunto de un polígono en un modelo vectorial, ya que puede ocurrir que los mayores niveles de incertidumbre de un polígono no estén en sus bordes, sino en su interior. El modelo raster permite analizar la eventual existencia de zonas de transición entre clases como origen del error (Goodchild *et al.*, 1992).

Cada realización de Monte Carlo debería representar una de las posibles representaciones de la realidad. Debido a que realizar representaciones de la realidad es imposible, Ehlschlaeger (2000) requiere utilizar los siguientes modelos para asemejarse lo más posible a ella en cada realización:

I. Un modelo de probabilidades para cada una de las categorías discretas. Debido a que no existe un modelo sencillo para esto, se pueden utilizar los siguientes enfoques:
   i. Error cartográfico: posición o localización incorrectas.
   ii. Error temático: descripción de las clases incorrectas.
   iii. Error conceptual: asignación de clase incorrecta.

   La distribución de probabilidades de una clase es la probabilidad de que esa clase exista en una posición concreta del mapa. La existencia de zonas del territorio con una distribución uniformemente aleatoria de valores inferiores a la distribución de probabilidades de cierta clase determina ausencias en la identificación de esa clase.

II. Un modelo de autocorrelación espacial para esas mismas clases que determine el nivel de dependencia con las celdas vecinas (error en la delimitación geográfica).
   i. Simulación estocástica, como el proceso de Monte Carlo a partir de un modelo de distribución de probabilidades.
   ii. Simulación condicional, al contrario, crearía modelos de una variable espacialmente distribuida a partir de la información generada en cada realización. Por ejemplo, kriging, media o desviación estándar.

Propagación de errores en el modelo

Una de las aplicaciones de los análisis de incertidumbre y sensibilidad es detectar entre todas las incertidumbres que intervienen en el modelo aquellas que tienen mayor influencia sobre la incertidumbre de la salida (Moreno Ruíz *et al.*, 2001). De esta manera, es posible comprender el comportamiento del modelo, tanto para su aceptación como para su mejora, así como para optimizar los recursos para la reducción de errores de datos de entrada. Solamente una vez asumido y conocido el origen de cierta incertidumbre en los datos de entrada, será posible analizar su propagación por el modelo (Huang *et al.*, 2009).

Para variables categóricas, el error se puede modelar mediante un proceso estocástico capaz de generar una población de versiones distorsionadas de la misma realidad. Cada instancia es un individuo de la población de versiones de la realidad, en el que a cada pixel se asigna una y sólo una categoría. Cada instancia simula cada una de las interpretaciones de la realidad que podría haber hecho un posible operador o proceso de digitalización. De esta manera, es posible calcular la probabilidad que tiene cada pixel de ser asignado a cada categoría tras la aplicación del modelo (Goodchild *et al.*, 1992). En esto se basa el Método de Monte Carlo, el cual permite procesar muchas instancias de una realidad factible, posibilitando el análisis de la propagación del error a través del modelo.

Moreno Ruíz *et al.* (2001) realizan un análisis de sensibilidad para determinar cómo contribuye la incertidumbre de cada una de las fuentes individuales a la incertidumbre de salida mediante la técnica "Extended FAST", la cual adopta la simulación de Monte Carlo como herramienta. Sea un campo A uno de los datos de entrada de un modelo y S la única realización particular de ese campo disponible. Se parte de un modelo genérico del error $Z(x) = S(x) + N(x)$, donde N es el campo de error de media m_N y varianza σ^2_N y Z es el campo perturbado por el error. En la simulación de Monte Carlo, en cada ciclo se genera aleatoriamente un campo N, añadiéndolo a S y obteniendo así una "versión corrupta" Z de A. Este modelo permite introducir dos tipos de incertidumbre: un error sistemático, considerando valores de m_N distintos de cero; y un error estocástico, haciendo que N sea un campo aleatorio, con matriz de covarianza C_{NN} que permite representar la correlación espacial del campo N (Moreno Ruíz *et al.*, 2001).

Huang *et al.* (2009) también muestran la viabilidad del análisis de sensibilidad de Monte Carlo para un modelo de salida discreta. Monte Carlo se muestra como un método flexible, ya que no incluye ninguna restricción sobre f(x). Esto representa la ventaja de poderse aplicar a cualquier modelo para estimar la repercusión del error en el mismo, aunque no se pueda construir un modelo matemático explícito de la incertidumbre. Sin embargo, tiene la ventaja frente a otros enfoques que permite localizar espacialmente el error estimado del modelo como el resto de estadísticos, así como cada una de las simulaciones.

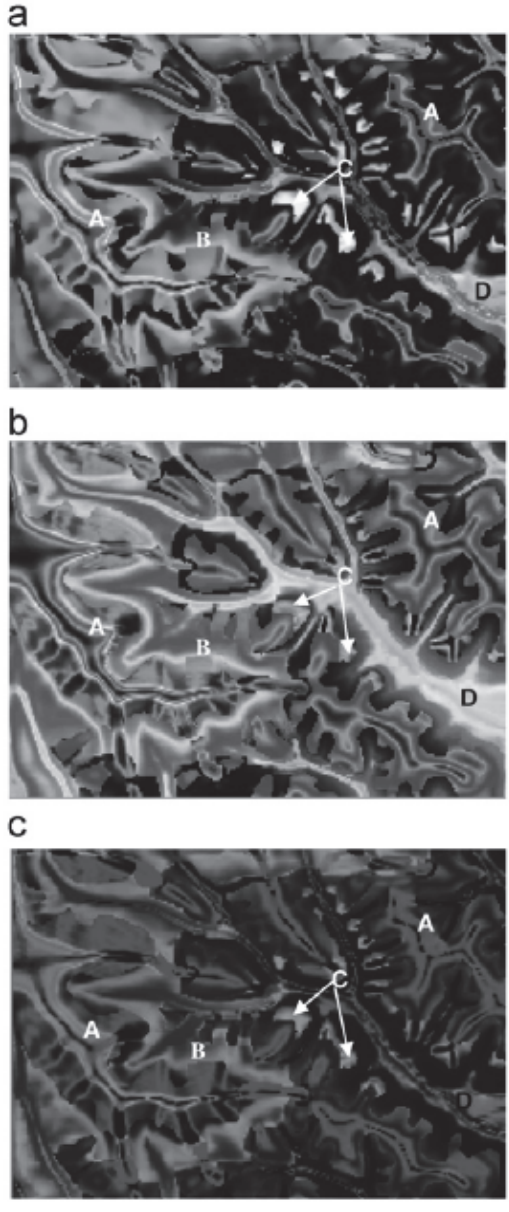

Figura 1. Distribución de la incertidumbre con tres modelos diferentes (a, b y c) para estimar la misma realidad. Los tonos de gris indican el nivel de incertidumbre de la clasificación de cada modelo (Qi *et al.*, 2011).

Resulta especialmente interesante el ejemplo de cartografía de suelos en clases discretas, pues los límites entre clases no siempre son claras en el terreno. Qi *et al.* (2011) analizan la incertidumbre en el cartografiado de clases de suelos comparando diferentes modelos que, en principio, deberían estimar la misma realidad. De esta manera, valora un grado de incertidumbre adicional: la capacidad del modelo de estimar la realidad. El procedimiento resulta particularmente útil para detectar las zonas de inexistencia de incertidumbre y las zonas de elevada incertidumbre. Como se muestra en la Figura 1, A y B son zonas de elevado desconocimiento de la incertidumbre (no es que la incertidumbre sea elevada, sino que en un pixel es baja y en el contiguo es alta, lo que denota desconocimiento de la incertidumbre en esa área) debido a que son zonas de transición entre clases. C y D son áreas con problemas potenciales de estimación mediante modelos: C indica una pequeña zona de elevada incertidumbre en mitad de una divisoria y D indica una incertidumbre excesivamente elevada en el fondo de valle, ya sea por la deficiente exactitud de los datos de entrada o porque la realidad física es muy variada en esta zona.

Particularidades del formato raster

Para aplicar Monte Carlo a variables discretas con formato raster se puede considerar como única causa del error la delimitación de bordes de clases, ya sea por error cartográfico como por error de digitalización (Huang *et al*., 2009 y Goodchild *et al*., 1992). Con esta premisa, para generar cada una de las versiones distorsionadas de los datos de entrada se generan ventanas, asignando al pixel central de la ventana uno de sus valores vecinos dentro de la ventana de manera aleatoria. Para su analizar su bondad, Huang *et al.* (2009) estudian la influencia del tamaño de la ventana en la cantidad de celdas cambiadas en las diferentes versiones de la salida del modelo. Los tamaños de ventana probados son desde 3 por 3 hasta 11 por 11, obteniendo que a mayor tamaño de ventana se obtiene un mayor número de celdas cambiadas en el modelo de manera constante. Este resultado es diferente al obtenido en este trabajo con las variables continuas en las que el aumento del error en los datos de entrada ocasionaron un aumento del error en la salida del modelo de manera exponencial (Figura 2). Los autores concluyen que la cantidad total de celdas con valor cambiado entre las diferentes realizaciones de los datos de entrada siempre es muy inferior en la variable de entrada discreta con respecto a las continuas. También concluyen que la exactitud y precisión de la variable discreta es la mayor en todos los niveles de celdas cambiadas / error respecto de las variables continuas.

También se observa que el aumento de la ventana en la variable discreta tiene poca influencia en la exactitud de salida del modelo.

Huang *et al.* (2009) realizan 1,000 iteraciones de Monte Carlo de los datos de entrada para estimar tipos de bosque a partir de datos de geología, MDT y teledetección, aunque a partir de 500 obtiene una solución estable.

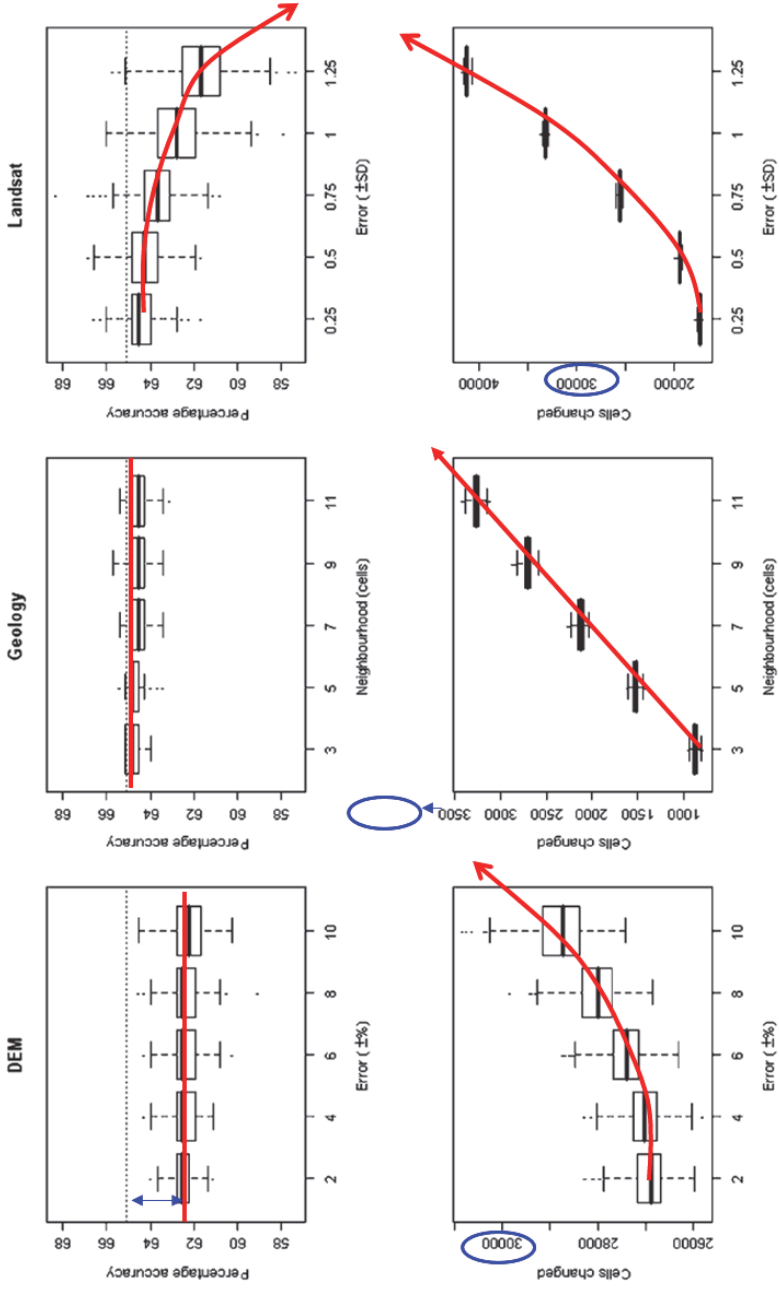

Figura 2. Gráficos de cajas para el análisis de resultados con los criterios de exactitud (superiores) y celdas cambiadas (inferiores). Variable discreta: geología y variables contínuas: MDE y Landsat (Huang *et al.*, 2009).

A partir de datos de teledetección (continuos), Moreno Ruíz *et al.* (2001) realizan 2,190 iteraciones de Monte Carlo para estimar con adecuada precisión los índices de sensibilidad de un modelo de detección (mediante umbrales) de superficies potencialmente quemadas (discreto) aplicado a un formato raster. Los factores considerados en la simulación de Monte Carlo son errores sistemáticos, estocásticos, de definición de máscara por nubes y definición de umbrales que localizan falsos positivos (quemado sí/no).

Particularidades del formato vectorial

Como ya se ha visto, otra manera de presentar los datos de entrada discretos es en formato vectorial. En este caso, el dato puede representar la proporción de cada una de las clases en cada polígono (Goodchild *et al.*, 1992). De esta manera se considera como fuente de error la correcta identificación de la proporción de cada clase dentro de cada polígono. Para aplicar Monte Carlo, se rasteriza el mapa asignando a los pixeles de cada polígono una clase de forma aleatoria, de tal manera que el conjunto del polígono tenga la proporción de clases inicial.

Finke *et al.* (1999) estudian la propagación del error a través del modelo SMART2 (Simulation Model for Acidification's Regional Trends, version 2) considerando como fuente del error la clasificación de datos de entrada discretos y su precisión geométrica. El modelo estima la acidificación del suelo (concentración de nitrato y aluminio discretizada). Para ensayar el modelo se toman dos colecciones de datos de entrada —edafología y uso del suelo— con diferentes niveles de detalle: escala 1:1,000,000, tamaño mínimo de polígono 2,500ha y tamaño de pixel 25ha frente a escala 1:50,000, tamaño mínimo de polígono 6ha y tamaño de pixel 0.0625ha. Con el fin de homogeneizar las clasificaciones de cada fuente de información, se reclasifica cada fuente de información en siete clases de suelo y cuatro tipos de vegetación, obteniendo una única capa de entrada de datos con 28 clases posibles (todas las combinaciones posibles de suelo y vegetación).

La hipótesis del estudio es que los datos de mayor escala son más exactos que los de menor escala, por lo que se pueden usar los primeros para evaluar la exactitud de los segundos.

Para ello realiza los siguientes pasos:

I. Utilización de un modelo del error de los datos de entrada que describa por zonas el grado de confusión de la clasificación (puro o mezclado), empleando una matriz de covarianza (Gómez-Hernández *et al.*, 1992), como el "joint sequential simulation of multigaussian fields", aunque Finke *et al.* (1999) no profundiza en ello.

II. Se encuentra que el modelo transmite homogéneamente el error de los datos de entrada para las clases dominantes, no siendo así para las clases minoritarias. Para evitar este efecto, la simulación se estratifica, subdividiendo el área de estudio en cada una de las 28 clases posibles de suelo-vegetación. De esta manera, se preservan los límites de cada categoría y cada simulación sólo afecta al valor de la clase.

III. Cabe destacar la importancia de representar adecuadamente la incertidumbre de salida del modelo. Para ello hay que representar la incertidumbre de cada clase, así como un modelo de autocorrelación espacial entre las clases (Ehlschlaeger, 2000). La correlación espacial de las clases simuladas se estudia reconstruyendo los variogramas experimentales para cinco clases diferentes dentro de las categorías evaluadas usando simulaciones seleccionadas aleatoriamente. Los modelos de variogramas se adecúan a los parámetros del variograma (meseta, rango y pepita) al compararlos con los parámetros del modelo de error.

IV. Con las anteriores simulaciones de datos de entrada se introducen en el modelo SMART2 para estimar la propagación del error mediante el método de Monte Carlo.

V. Se analiza la parte del error atribuible a las variables de entrada continuas mediante un modelo de bloques agregados (Figura 3).

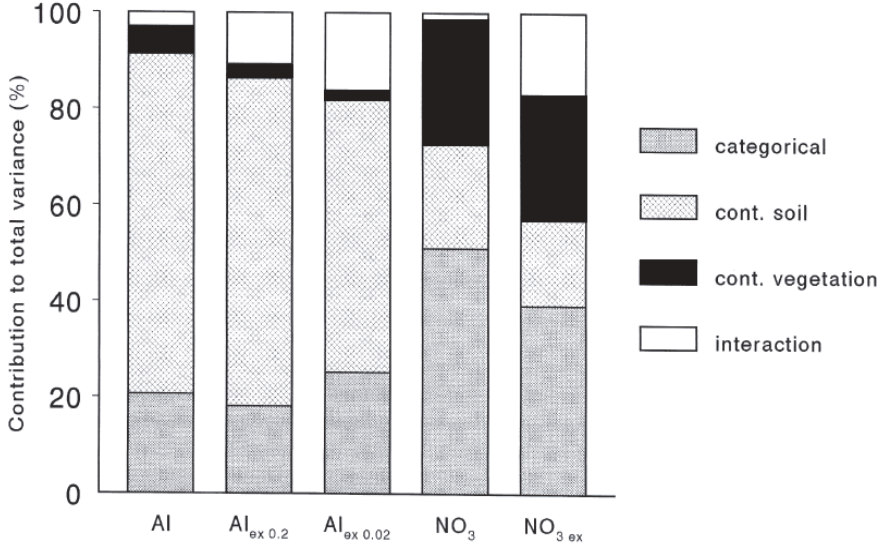

Figura 3. Contribución relativa de la varianza de la salida de cinco modelos con cuatro fuentes de incertidumbre en la entrada del modelo (Finke *et al*., 1999).

En el caso analizado los autores encontraron que la varianza atribuible a las variables categóricas fue mucho menor en casi todos los casos que la atribuible a variables continuas.

VI. Análisis de la varianza para identificar el efecto de las diferentes fuentes del error.

El trabajo de Ehlschlaeger (2000) concluye que los errores de los datos categóricos atribuibles a la confusión de la clasificación y la escala tienen una pronunciada influencia en los resultados del modelo, siendo esta influencia diferente según el parámetro de salida que se considere.

Conclusiones

Para llevar a cabo el análisis de sensibilidad de un modelo medioambiental de salida discreta el formato óptimo es el raster, para lo cual se analizan los siguientes criterios (Huang *et al.*, 2009):

I. Exactitud: diferencia global entre el resultado de cada simulación y el mapa de referencia.
II. Pixeles cambiados entre el resultado de cada simulación y el mapa original.
III. Pixeles cambiados en iteraciones: frecuencia de cada cambio de categoría de pixel a lo largo de todas las iteraciones.

Los trabajos analizados muestran, en general, que al aumentar el error en los datos de entrada se produce un incremento de la sensibilidad del modelo, aunque se han encontrado respuestas diferentes.

El criterio de pixeles cambiados de categoría por ventanas utilizado para cuantificar la incertidumbre en variables discretas muestra un patrón similar al criterio de porcentaje de error de variables continuas.

La sensibilidad del modelo no resulta igual para todas las variables de entrada. En un trabajo concreto Huang *et al.* (2009) concluyen que la variable categórica analizada tiene una respuesta lineal, pero en el caso de las continuas la respuesta es exponencial. En ese caso el modelo se mostró globalmente menos sensible a la variable discreta que a las continuas.

Resulta destacable que dos trabajos (Finke *et al.*, 1999, y Huang *et al.*, 2009) hayan concluido que la propagación de incertidumbre a través del modelo no resulta igual para todas las clases (discretas) de salida del modelo. Al modelar datos discretos, el modelo muestra una sensibilidad similar para cada una de las categorías más abundantes. Sin embargo, cuando el dato de entrada en el modelo es una categoría minoritaria, el modelo se muestra más sensible, ya sea por ser menos frecuente o

bien por presentar una variabilidad mayor. Este comportamiento ha sido detectado por los dos trabajos citados, pero no estudiado expresamente en suficiente profundidad ni por ellos ni por otros autores.

Moreno Ruíz *et al.* (2001) señalan como conclusión que los análisis de sensibilidad y análisis de incertidumbre de los modelos medioambientales también pueden ser utilizados para optimizar la construcción de nuevos modelos medioambientales y la mejora de los ya existentes.

Agradecimientos

El presente artículo ha sido realizado durante los estudios de Máster Universitario en Ingeniería Geodésica y Cartografía de la Universidad Politécnica de Madrid (España). Mi agradecimiento al profesor Dr. Carlos López Vázquez por sus aportaciones.

Bibliografía

Ehlschlaeger, C.R. (2000). "Representing Uncertainty of Area Class Maps with a Correlated Inter-Map Cell Swapping Heuristic", *Computers, Environment and Urban Systems*, vol. 24, núm. 5, September, pp. 451-469.

Finke, P.A.; Wladis, D.; Kros, J.; Pebesma, E.J. y G.J. Reinds (1999). "Quantification and Simulation of Errors in Categorical Data for Uncertainty Analysis of Soil Acidification Modelling", *Geoderma*, núm. 93, pp. 177-194.

Goodchild, M.F.; Guoqing, S. y Y. Shiren (1992). "Development and Test of an Error Model for Categorical Data", *International Journal of Geographical Information Science*, vol. 6, núm. 2, pp. 87-103.

Gómez-Hernández, J.J. y A.G. Journel (1992). "Joint Sequential Simulation of Multigaussian Fields", Soares, A. (ed.), Proceedings of the Fourth Geostatistics Congress Troia (Portugal). *Quantitative Geology and Geostatistics*, núm. 5, Kluwer Academic Publishers, pp. 85-94.

Huang, Z. y S.W. Laffan (2009). "Sensitivity Analysis of a Decision Tree Classification to Input Data Errors Using a General Monte Carlo Error Sensitivity Model", *International Journal of Geographical Information Science*, vol. 23, núm. 11, pp. 1433-1452.

Moreno Ruíz, J.A. y M. Crosetto (2001). "Análisis de la incertidumbre en modelos de teledetección", *Teledetección, Medio Ambiente y Cambio Global*, pp. 538-541.

Qi, F. y A.X. Zhu (2011). "Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps", *Computers & Geosciences*, núm. 37, pp. 1425-1436.

Satelli, A.; Chan, K. y M. Scott (eds.) (2000). *Sensitivity analysis*, Wiley, New York.

Veregin, H. (1989). "Error modelling for the map overlay operation", *Accuracy of spatial databases*, M.F. Goodchild y S. Gopal (eds.), pp. 3-18.