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

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

INCERTIDUMBRE DE MODELOS ESTADÍSTICOS ASOCIADA A LOS NIVELES DE AGREGACIÓN DE LA INFORMACIÓN ESPACIAL

JEAN F. MAS[1], AZUCENA PÉREZ VEGA[2], ARACELI ANDABLO REYES[3], MIGUEL ANGEL CASTILLO SANTIAGO[4]
[1]Centro de Investigaciones en Geografía Ambiental Universidad Nacional Autónoma de México.
Antigua Carretera a Pátzcuaro No. 8701,
Col. Ex-Hacienda de San José de la Huerta.
C.P. 58190. Morelia Michoacán, México
jfmas@ciga.unam.mx
[2]Universidad de Guanajuato
Bellavista, 36730 Salamanca, Gto., México
azu_pvega@hotmail.com
[3]Centro de Investigación en Alimentación y Desarrollo A.C.
Carretera a La Victoria km 0.6, Hermosillo, Sonora, México, C.P. 83304, México
aandablo@ciad.mx
[4]El Colegio de la Frontera Sur, Unidad San Cristóbal de las Casas
Periférico Sur s/n, María Auxiliadora, 29290 San Cristóbal de las Casas, Chis., México
aandablo@ciad.mx

RESUMEN

La modelación de fenómenos como los cambios de cubierta / uso del suelo se basa en la evaluación de la relación entre el cambio y variables explicativas utilizando métodos estadísticos como los modelos de regresión. Las variables explicativas utilizadas describen las condiciones físicas y socioeconómicas del territorio. La información disponible se presenta a menudo de forma agregada espacialmente en unidades político-administrativas como municipios. Sin embargo, los resultados de análisis estadísticos no son independientes de la configuración espacial de las unidades utilizadas para agregar la información. En este estudio, analizamos los efectos de este fenómeno, conocido como el problema de la unidad de área modificable (MAUP por sus siglas en inglés), sobre la evaluación de los factores de la distribución de la cubierta forestal en México a diferentes niveles de agregación. Utilizamos variables del censo de población junto con variables topográficas y de accesibilidad, dicha agregación utiliza áreas geoestadísticas básicas, municipios y estados. Los resultados muestran que el nivel de agregación de la información afectó los valores del

Recibido: 15/08/2017
Aceptada versión definitiva: 28/05/2018
Editor al cargo: Dra. Anna Badia

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

coeficiente de correlación y el ajuste de los modelos de regresión. El MAUP tuvo un efecto sustancial en estos modelos, en particular, cuando no hay una fuerte relación entre la variable dependiente y las variables explicativas. Estos resultados sugieren que las relaciones e inferencias obtenidas usando datos agregados con unidades administrativas como condados, provincias o municipios, deben ser interpretados con precaución.

Palabras clave: unidad de área modificable; MAUP; agregación espacial, censo; análisis estadístico.

UNCERTAINTY OF STATISTICAL MODELS ASSOCIATED WITH THE LEVELS OF AGGREGATION OF SPATIAL INFORMATION.

ABSTRACT

The modeling of phenomena such as land use / cover changes is based on the evaluation of the relationship between change and explanatory variables using statistical methods as regression models. The explanatory variables describe the physical and socioeconomic conditions of the territory. The available information is often presented in a spatially aggregated form based on political-administrative units such as municipalities. However, results of statistical analyses are not independent from the spatial configuration of the units used to aggregate the information. In this study, we analyze the effects of this effect, known as the modifiable areal unit problem (MAUP), on the evaluation of the factors of the distribution of forest cover in Mexico at different aggregation levels. We used population census variables along with topographic and accessibility variables aggregated using basic geostatistical areas, municipalities and states. The results show that the level of aggregation of the information affected the values of the correlation coefficient and the fitting of the regression models. The MAUP had a substantial effect on these models, in particular, when there is no strong relationship between the dependent variable and the explanatory variables. These results suggest that the relationships and inferences obtained using aggregated data with administrative units such as counties, provinces or municipalities, should be interpreted with caution.

Keywords: Modifiable areal unit; MAUP; Census; Spatial aggregation, statistical analysis.

1. Introducción

La construcción de modelos "espacialmente explícitos" como los modelos de análisis de los patrones de distribución de los tipos de cubierta / uso del suelo o de los cambios de cubierta / uso del suelo (CCUS) implica dos principales etapas, una de orden conceptual, que se construye con base en la delimitación del problema de investigación y el marco o modelo conceptual de referencia elegido por el investigador. En esta primera etapa, se definen los principales procesos y relaciones de interés para la investigación, así como las posibles variables que los representarán en el modelo (Vliet *et al.*, 2016). La segunda etapa implica la traducción del modelo conceptual en un modelo operacional codificado y preferentemente computarizado, para luego pasar a la calibración y validación, y por último, a la experimentación e interpretación de los resultados (Vliet *et al.*, 2016; Camacho Olmedo *et al.*, 2018).

Estos modelos se utilizan para representar, describir, explicar y predecir los CCUS (NRC, 2013). Los resultados y escenarios derivados de estos ejercicios ofrecen información e instrumentos

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

para la toma de decisiones y la generación de políticas en diversos ámbitos (Mahmood *et al.*, 2016). Por esta razón, la fase de validación del modelo es esencial para garantizar su fiabilidad y el alcance de sus resultados (Brown *et al.*, 2013; Vliet *et al.*, 2016; Camacho Olmedo *et al.*, 2018). La incertidumbre en la modelación depende de múltiples factores (Soares-Filho, 2013), uno de los principales es la disponibilidad de información de las variables elegidas en la escala adecuada para resolver las preguntas centrales de la investigación.

Entre los insumos básicos para elaborar los modelos de CCUS se incluyen las imágenes satelitales, cartografía de rasgos ambientales, información sobre actividades socioeconómicas derivada de fuentes primarias, como la generada en campo, y de fuentes secundarias como los datos censales (Kindu *et al.*, 2015; Vitali *et al.*, 2018; Camacho Olmedo *et al.*, 2018). Con base en esta información se construyen variables físicas y socioeconómicas que se organizan en un Sistema de Información Geográfica para establecer relaciones con los índices de cambio obtenidos del procesamiento de imágenes de satélite multi-temporales (Tapia Silva & López Flores, 2017). El análisis de estas relaciones se realiza con base en métodos estadísticos, como modelos de regresión, con el fin de evaluar el efecto de las variables consideradas conductoras o explicativas de los CCUS (Bravo Peña *et al.*, 2017; Camacho Olmedo *et al.*, 2018).

Particularmente, la información disponible en fuentes oficiales, como los censos, se presenta agregada espacialmente en unidades político-administrativas tales como provincias, municipios o entidades federativas. Un problema frecuente que genera incertidumbre sobre la modelación es que los resultados del análisis estadístico dependen de la escala y la configuración espacial de las unidades utilizadas para agrupar la información. Según Openshaw (1984), este problema, conocido como el Problema de la Unidad de Área Modificable (MAUP), tiene dos componentes: el efecto de la escala y el de zonificación. El problema de la escala es la modificación de los resultados observados cuando los datos se agrupan en conjuntos de unidades de agregación cada vez mayores. El problema de la zonificación está relacionado con las variaciones en los resultados observados cuando el análisis se lleva a cabo utilizando unidades alternativas del mismo tamaño.

Este trabajo tiene como objetivo evaluar los efectos del MAUP sobre el comportamiento de los parámetros estadísticos de modelos de la distribución de las cubiertas forestales en la República Mexicana. Para evaluar el efecto que tiene la agregación diferenciada de los datos, se construyen modelos de regresión global y local utilizando algunas variables derivadas de los Censos de Población y Vivienda (INEGI, 2010a y 2010b) que están disponibles en tres niveles de agregación: entidades federativas (estados), municipios y áreas geoestadísticas básicas (AGEBs); además de algunas variables físicas para las que se calcularon índices en los mismos niveles de agregación de las variables censales. La hipótesis que se plantea es que si los resultados de la modelación no varían en función del nivel de agregación, el MAUP no afecta dicho análisis.

2. Antecedentes

Los efectos del MAUP han sido reportados en diversos temas como la ecología del paisaje (Jelinski y Wu, 1996), la cartografía de la pobreza (Hayward y Parent, 2009), el efecto de la contaminación del aire sobre problemas de salud (Parenteau y Sawada, 2011), la relación entre viajes activos y características del entorno urbano (Clark y Scott, 2013), la definición de regiones

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

económicas (Pietrzak, 2014), la representación de tasas de desempleo (Weir-Smith, 2016) y la evaluación de programas de conservación (Avelino *et al.*, 2016).

Una de las principales líneas de investigación desarrolladas se enfoca en la búsqueda de una escala "ideal" de representación de los datos, que minimice o elimine los efectos del MAUP. Weir-Smith (2016) recomienda que la agregación de variables socio-económicas se haga a partir de las unidades más pequeñas disponibles. Gerell (2017) también recomienda pequeñas unidades de análisis para explicar la ocurrencia de incendios provocados en una ciudad de Suecia. En cambio, Avelino *et al.* (2016) encuentran que unidades demasiado pequeñas presentan ruido, y que existe un compromiso entre unidades demasiado pequeñas y demasiado grandes. Varios autores proponen llevar a cabo los análisis con base en unidades que tienen sentido respecto al proceso estudiado (Jelinski y Wu, 1996; Parenteau y Sawada, 2011; Avelino *et al.* 2016). Sin embargo, podemos imaginar que para el mismo proceso, diferentes variables actúen a diferentes escalas. Diez Roux (2008) señala que este problema se deriva de la presencia de múltiples niveles de organización de los elementos que intervienen en el desarrollo de un fenómeno social o ecológico, por lo que sería adecuado adoptar un enfoque multinivel con preguntas pertinentes para cada nivel de organización del fenómeno.

Otra línea de investigación aborda la homogeneidad de forma geoestadística basada en evaluaciones de la varianza de las unidades o la autocorrelación espacial. Algunos autores proponen usar unidades homogéneas (Pietzrak, 2014b), permitiendo minimizar la variación dentro de las unidades y maximilizarla entre unidades (Butkiewicz *et al.*, 2010). Sin embargo, Fotheringham y Wong (1991) observan que no es posible obtener una zonificación óptima para todas las variables. Finalmente, otros autores proponen algunos métodos para detectar los casos en los cuales el efecto del MAUP es severo utilizando diferentes enfoques como análisis de sensibilidad (Xu *et al.*, 2014) o análisis multi-fractal (Sémécurbe *et al.*, 2016). Butkiewicz *et al.* (2010) proponen herramientas que permiten visualizar los datos y detectar casos en los cuales el MAUP puede afectar los resultados.

Finalmente, existe una discusión sobre el efecto del MAUP en análisis locales como los modelos de regresión ponderada geográficamente (RPG). Cuando existe una alta correlación entre variables explicativas, el ajuste del modelo de regresión puede volverse inestable en el sentido que pequeños cambios pueden producir cambios importantes en la estimación de los parámetros del modelo (Belsley *et al.*, 1980). Estos cambios pueden ser causados por el orden de presentación de las variables en el modelo o bien la variación de los valores de las observaciones producida por una agregación espacial diferente. Wheeler y Tiefelsdorf (2005) opinan que los modelos de RPG son altamente susceptibles a los efectos de la multicolinealidad, aun cuando estos problemas no se presenten en los datos globales. Al contrario, Fotheringham y Oshan (2016) opinan que las RPG no son más sensibles a la autocorrelación que cualquier modelo de regresión.

2. Materiales

En México, la información censal se encuentra disponible principalmente en tres niveles de agregación: entidad federativa (estado), municipio y AGEBs (Figura 1). Existen 32 estados y 2456 municipios, cuya área varía de 2.2 a más de 53,000 km^2 con una superficie promedio de 796 km^2 y más de 54,000 AGEBs. Para el estudio se eligió trabajar con las AGEBs rurales, subdivisiones

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

municipales que en promedio abarcan 11,000 ha, se especializan en el uso de suelo agropecuario o forestal, pueden incluir varias localidades rurales, pero existen algunas sin localidades, y están delimitadas por rasgos naturales como ríos y/o culturales como carreteras y límites prediales (INEGI, 2010a).

Figura 1. Diferentes niveles de agregación: Estado, municipio y AGEB

Para la construcción de la base de datos, se utilizaron algunas variables socio-económicas del Censo de Población y Vivienda de 2010 del INEGI a nivel de localidad (INEGI, 2010b), y el índice de marginación calculado por la Comisión Nacional de Población con base en la información de vivienda, nivel de educación e ingresos del INEGI (CONAPO, 2010). Las variables físicas de elevación y pendiente se obtuvieron del modelo digital de elevación del *Shuttle Radar Topography Mission* (SRTM, http://www2.jpl.nasa.gov/srtm/) y los datos de cobertura arbórea se tomaron de la base de datos sobre Cambio Forestal Global (Hansen *et al.*, 2013; *Global Forest Change*: https://earthenginepartners.appspot.com/). Los mapas a escala municipal y estatal, se generaron a partir del mapa de las AGEBs rurales. La Tabla 1 muestra la fuente y la resolución original de las variables utilizadas en el estudio.

Table 1. Características de los insumos

Variable original	Fuente	Resolución	Variable agregada	Nombre abreviado
Cubierta Forestal 2000	Global Forest Change	30 m, remuestreado a 300 m	Proporción de bosque (%)	PB
Número de habitantes	Censo INEGI	Localidad	Densidad de población (hbs/km^2)	DensPob
Hacinamiento	Censo INEGI	Localidad	Hacinamiento (%)	Hacina
Viviendas con piso de tierra	Censo INEGI	Localidad	Viviendas con piso de tierra (%)	Piso
Índice de Marginación	CONAPO	Localidad	Índice de Marginación	Marg
Elevación	MDE SRTM	90 m, remuestreado a 300 m	Elevación promedio (m)	Elev
Pendiente	SRTM DEM	90 m, remuestreado a 300 m	Pendiente promedio (grado)	Pend
Distancia a carretera	INEGI	90 m, remuestreado a 300 m	Distancia a carretera promedio (m)	DistCarr

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

Los datos en formato *raster* se remuestrearon a 300 m para reducir el tamaño de los archivos y los tiempos de procesamiento conservando un detalle suficiente para la agregación de los datos. Todos los análisis espaciales y estadísticos se llevaron a cabo utilizando el programa de código abierto R (R Core team, 2018).

3. Métodos

El área de estudio abarca el territorio continental de México con cerca de dos millones de km^2. Para evaluar la variabilidad en los resultados de algunos análisis estadísticos asociada al MAUP, se agregó la información de los insumos (Tabla 1) a nivel de AGEB, municipio y estado. Los mapas de municipios y estados, se obtuvieron agregando la información de las AGEBs rurales, por lo tanto se trata de la misma fuente de información en los tres niveles de agregación.

Para cada polígono del mapa de agregación, se calculó la proporción de bosque (*PB*), la elevación promedio (*Elev*), la pendiente promedio (*Pend*), la densidad de población (*DensPob*), el índice de marginación promedio (*Marg*), la proporción de viviendas con piso de tierra (*Piso*), la proporción de viviendas con hacinamiento (*Hacina*) y la distancia promedio a carreteras (*DistCarr*). Estas variables fueron seleccionadas, porque en diversos estudios sobre CCUS se reportan como variables importantes para explicar los cambios (Lo y Yang, 2002; Mertens y Lambin, 2004; Pineda *et al.* 2008; Bravo Peña *et al.*, 2017; Marshall *et al.*, 2017). Con base en cada nivel de agregación, se realizaron algunos análisis estadísticos que se detallan a continuación. Si los resultados de un cierto análisis estadístico no varían en función del nivel de agregación, se considera que el MAUP no afecta dicho análisis.

Los análisis estadísticos incluyen el cálculo del coeficiente de correlación de Pearson y el ajuste de modelos de regresión. El coeficiente de correlación de Pearson entre la proporción de bosque y cada una de las variables explicativas es una medida de la relación lineal entre dos variables cuantitativas cuyo valor varía entre -1 y 1. El índice con un valor cercano a uno indica una fuerte dependencia entre las dos variables: cuando una de ellas aumenta, la otra aumenta también. Un valor cercano a -1, indica una fuerte relación inversa entre las dos variables: cuando una aumenta, la otra disminuye. Un valor cercano a 0 indica una ausencia de relación lineal entre ambas variables. Existen pruebas estadísticas para evaluar si el coeficiente es significativamente diferente de cero. En el caso de este estudio, el coeficiente de Pearson permite determinar las variables explicativas que son más relacionadas con la proporción de bosque así como la forma (positiva o negativa) de esta relación. Los modelos lineales usaron como variable dependiente la proporción de bosque y como variables explicativas las variables físicas y socioeconómicas de la tabla 1. Se ajustaron utilizando el método por pasos hacia adelante y atrás, que consiste en agregar y eliminar variables explicativas de forma iterativa y se basa en el índice de Akaike (AIC) para seleccionar un subconjunto de las variables explicativas. El AIC es una medida de la calidad de un modelo que permite un compromiso entre la bondad de ajuste y la complejidad del modelo. En el caso de este estudio, los parámetros (en particular su signo y su significancia) asociados a cada variable explicativa permite evaluar el efecto de cada variable física o socio-económica para explicar la proporción de bosque en el área de estudio. Ambos análisis (Pearson y modelos de regresión) se llevaron a cabo de forma global (para todo el territorio nacional) y local (análisis basado en ventanas que abarcan porciones del territorio) con base en los datos a diferentes niveles de agregación.

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

4. Resultados

La Tabla 2 muestra la correlación global entre la proporción de bosque y las variables explicativas usando la información agregada a nivel de AGEB, municipio y estado. Se puede observar que todas las variables presentan importantes diferencias del valor de correlación dependiendo del nivel de agregación. A excepción de *DensPob* y *DistCarr*, todas dejan de ser significativas en el nivel de agregación mayor (estados). Se puede observar que la elevación presenta valores de correlación positivo a nivel de AGEBs y negativo a nivel de municipio, ambos valores indicando una correlación muy débil pero estadísticamente significativa.

Tabla 2. Coeficiente de correlación entre PB y cada una de las variables explicativas en los tres niveles de agregación. * p <= 0.05, ** p <= 0.01, *** p <= 0.001.

Nivel de agregación	DensPob	Marg	Hacina	Piso	DistCarr	Elev	Pend
AGEB	-0.24***	0.36***	0.15***	0.33***	0.36***	0.05***	0.45***
Municipio	-0.33***	0.25***	0.14***	0.25***	0.43***	-0.08***	0.52***
Estado	-0.71***	0.07	0.09	0.10	0.68***	-0.33	0.06

El ajuste de los modelos de regresión lineal se realizó excluyendo la variable *Marg* que tenía una correlación muy alta (> 0.8) con otras variables explicativas (*Piso* y *Hacina*). El método paso a paso condujo a la selección de todas las variables en cada nivel de agregación (Tabla 3) a excepción de la variable *Elev* que fue eliminada para los estados debido a su falta de significancia estadística. Los parámetros asociados a la misma variable no cambian de signo dependiendo del nivel de agregación.

Tabla 3. Parámetros de los modelos lineales globales

Variable	AGEB	Municipio	Estado
Ordenada al origen	$2.59.10^{-1}$	$2.17.10^{-1}$	$6.17.10^{-2}$
DensPob	$-6.67.10^{-4}$	$-9.72.10^{-4}$	$-4.84.10^{-3}$
Hacina	$4.68.10^{-4}$	$1.66.10^{-3}$	$9.49.10^{-3}$
Piso	$-6.74.10^{-4}$	$-2.19.10^{-3}$	$-5.10.10^{-3}$
DistCarr	$2.71.10^{-5}$	$4.03.10^{-5}$	$3.07.10^{-5}$
Elev	$-1.97.10^{-5}$	$-4.69.10^{-5}$	
Pend	$2.63.10^{-2}$	$2.69.10^{-2}$	$2.42.10^{-2}$

Para los análisis locales, se definieron ventanas de 500 x 200 km^2 y se analizaron solo las 17 ventanas que contienen por lo menos 30 municipios para garantizar la robustez de los análisis estadísticos. Los análisis se realizaron a nivel de municipio y AGEB tomando en cuenta las AGEBs de estos municipios. Se detectaron cinco ventanas para las cuales se observaron ciertas contradicciones entre los resultados obtenidos a nivel de AGEB y de municipio por lo menos para una variable: 1) una diferencia de, por lo menos, 0.3 entre los valores del coeficiente de correlación

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

y 2) un coeficiente de correlación con significancia ($p <= 0.05$). No se observó ningún caso en el cual dos coeficientes de correlación significativos tenían un signo opuesto.

Para tratar de entender mejor las causas de estas discrepancias, se analizó a detalle una de estas ventanas. En este caso, se puede observar que la densidad poblacional, significativa a nivel de AGEB ($r=-0.25$, $p < 0.001$), ya no muestra correlación a nivel de municipio mientras ocurre lo contrario para la elevación (Tabla 4). Para el desarrollo de los modelos de regresión, se siguió el mismo método que para los modelos a escala nacional. Debido a que la densidad de población tenía algunos valores nulos y otros muy elevados, se aplicó una transformación logarítmica. Los modelos de regresión lineal se basan en diferentes conjuntos de variables seleccionadas (Tabla 5). Por lo tanto, dependiendo del nivel de agregación de los datos, se llega a conclusiones diferentes al identificar las variables más relacionadas con la proporción de bosque. Mientras que a nivel de AGEB la densidad poblacional y la marginación son variables importantes para explicar la proporción de bosque, a nivel de municipio estas variables tienen poca importancia. En cambio, las variables topográficas figuran en ambos niveles de agregación.

Tabla 4. Correlaciones entre la proporción de bosque y las variables explicativas en los dos niveles de agregación.

Nivel	Log(DensPob + 1)	Marg	Hacina	Piso	DistCarr	Elev	Pend
AGEB	-0.25***	0.35***	0.21***	0.24***	0.27***	-0.06	0.42***
Municipio	0.00	0.18	0.48**	0.01	0.21	-0.42*	0.31

Tabla 5. Parámetros del modelo de regresión lineal entre la proporción de bosque y las variables explicativas en los dos niveles de agregación.

Nivel	DensPob	Marg	Hacina	Piso	DistCarr	Elev	Pend
AGEB	$-4.35.10^{-3}$	$7.84.10^{-2}$	-	-	-	$-3.13.10^{-4}$	$5.09.10^{-2}$
Municipio	-	-	$6.60.10^{-3}$	-	$1.96.10^{-5}$	$-2.69.10^{-4}$	$3.34.10^{-2}$

Las figuras 2 a 5 representan las variables *PB*, *Elev* y *DensPob* en los datos originales y agregados a nivel de AGEB y de municipio. Se puede observar que los municipios agrupan AGEBs con características heterogéneas y que, por consecuencia, presentan valores "suavizados" por el efecto del promedio que elimina los valores extremos. Este efecto es particularmente severo para la densidad poblacional ya que las localidades más pobladas están agregadas en el espacio. Hayward y Parent (2009) observan efectos similares en mapas de índices de pobreza realizados con diferentes niveles de agregación. Estas variaciones en el valor de las variables modifican la relación entre ellas: por ejemplo, en las figuras 6 y 7 se puede observar que la relación entre *PB* y *Elev*, por un lado y *PB* y *DensPob* por el otro no son muy estrechas (el coeficiente de Pearson es inferior a 0.5). Sin embargo, una relación observable a un cierto nivel de agregación se desvanece totalmente en el otro.

Por lo tanto, en este caso de estudio, observamos que, en los tres niveles de agregación utilizados, la relación entre la proporción de bosque y un conjunto de variables socioeconómicas, de

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

accesibilidad y topográficas es débil. El MAUP provoca cambios en los valores de las variables de las unidades de agregación que modifican de forma importante la correlación entre variables y el ajuste de los modelos de regresión.

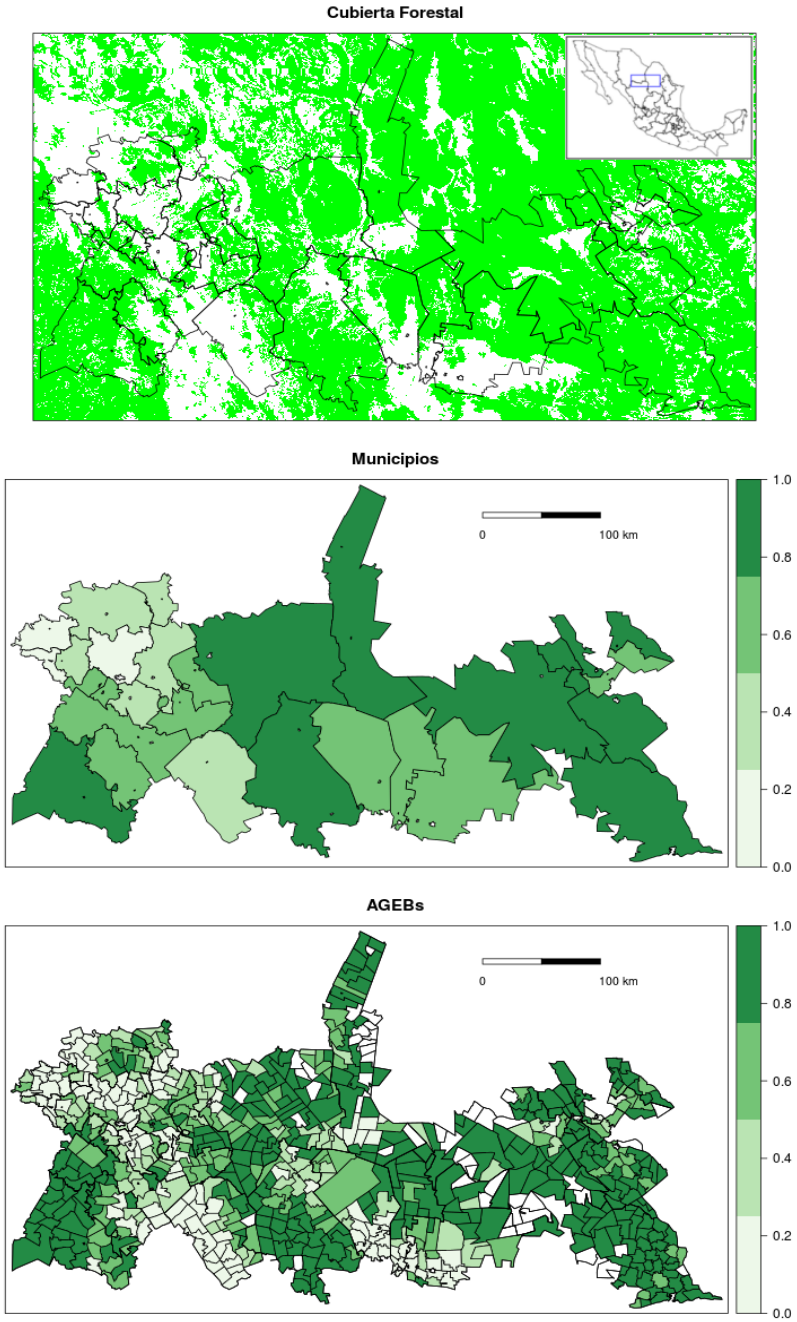

Figure 2. Cubierta forestal: Datos originales y proporción de bosque por municipio y por AGEB

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

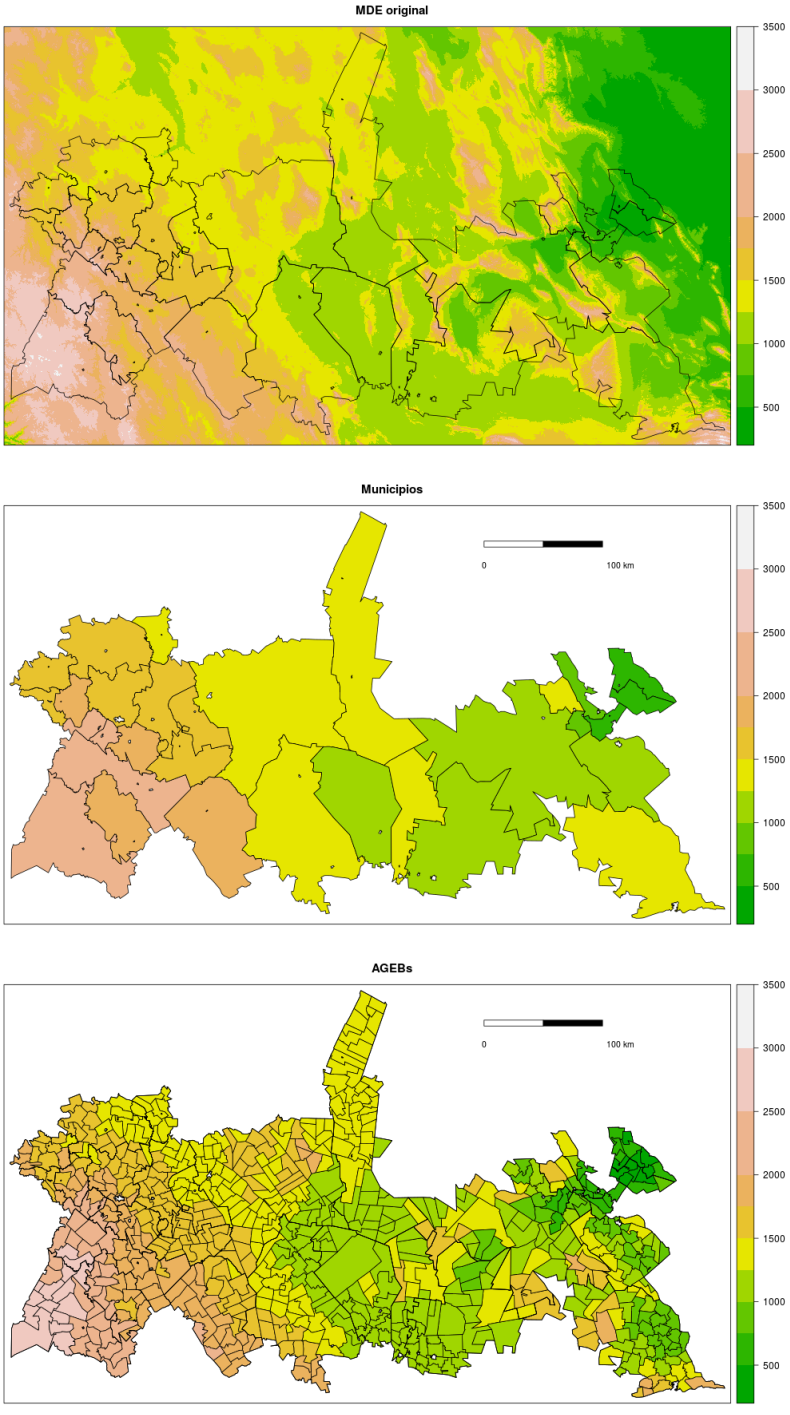

Figura 3. Elevación: DEM original y elevación promedio por municipio y por AGEB.

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

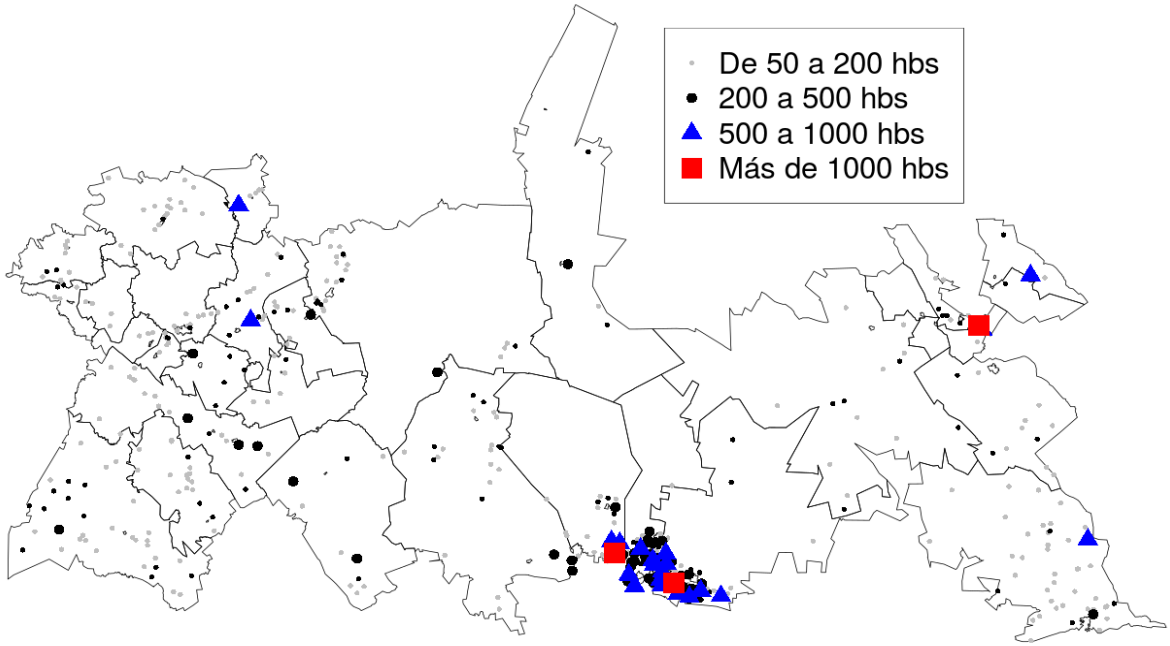

Figure 4. Distribución de las poblaciones de más de 50 habitantes

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

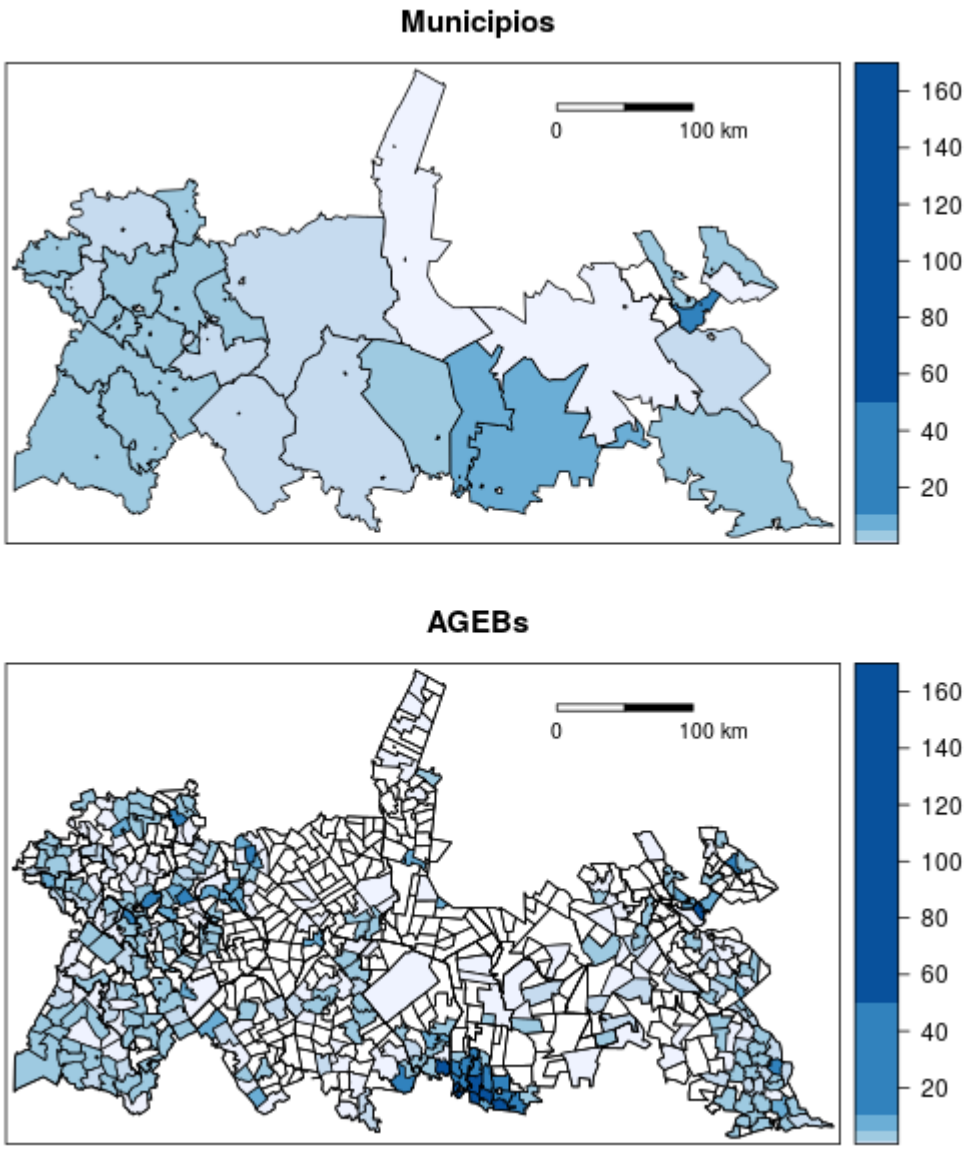

Figure 5. Densidad poblacional calculada con base en los municipios y las AGEBs

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

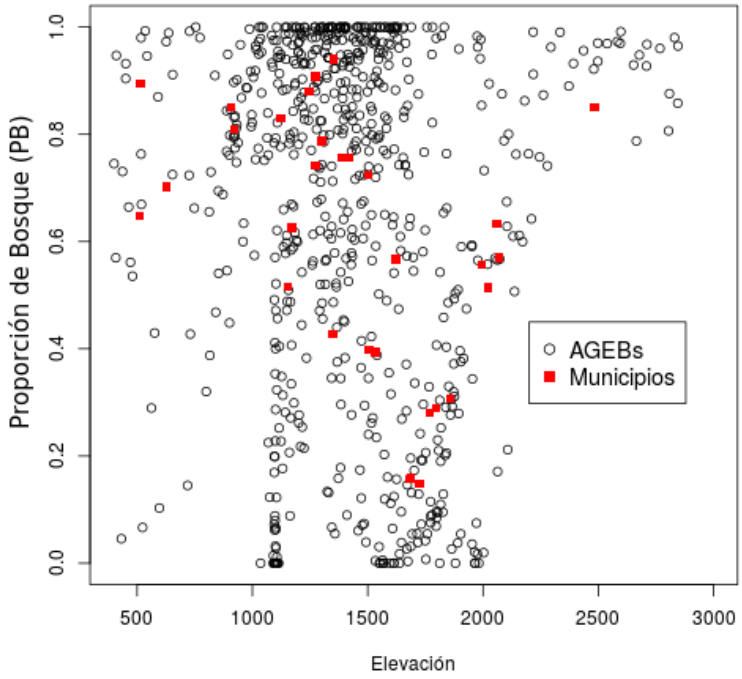

Figura 6. Relación entre *PB* y *Elev* en los dos niveles de agregación. A nivel municipal, se observa una relación negativa (r = -0.42) mientras que a nivel AGEB r es nulo.

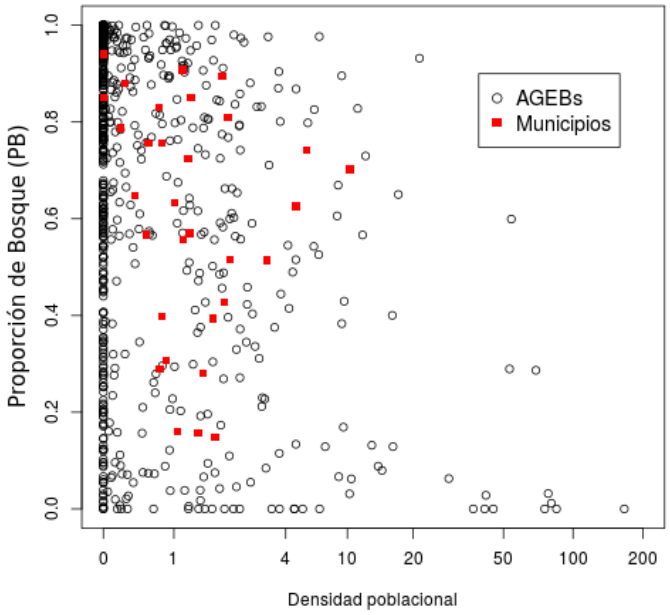

Figura 7. Relación entre *PB* y *DensPob* en los dos niveles de agregación. A nivel municipal r es nulo mientras que a nivel de AGEB r = -0.25.

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

5. Discusión y conclusión

Flowerdew (2011) calculó el coeficiente de correlación entre variables del censo del Reino Unido utilizando tres niveles de agregación y concluyó que, en general, las diferencias son mínimas pero que, en algunos casos, pueden ser sustanciales. Al contrario, algunos estudios reportan un muy fuerte efecto del MAUP como, por ejemplo, variaciones de los valores de las correlaciones entre -1 y 1 (Openshaw y Rao, 1995). En nuestro caso, las diferencias son más importantes que las reportadas por Flowerdew y mucho menos extremas que las de Openshaw y Rao. Obtuvimos diferencias mayores que Flowerdew (2011) probablemente porque, en vez de usar únicamente variables del censo, usamos variables de otras fuentes y naturaleza. Openshaw y Rao (1995) observaron grandes variaciones usando unidades de agregación muy convolucionadas, lo cual no es el caso de las unidades de agregación administrativas que usamos.

Es probable que las diferentes variables explicativas estudiadas actúen a diferentes escalas. Por ejemplo, en el caso de los procesos de deforestación, la pendiente podría actuar muy localmente: se conserva el bosque en una barranca mientras que se deforesta en áreas vecinas con pendientes más suaves. Al contrario, la densidad de población podría tener un efecto más difuso en el espacio porque las actividades humanas no se circunscriben a las inmediaciones de los lugares de residencia. En el caso de las divisiones utilizadas en México, podemos esperar que las AGEBs sean unidades pertinentes para el estudio de procesos socioambientales ya que en su delimitación se busca tener una homogeneidad del tipo de uso del suelo y de la tenencia de la tierra. En cambio, los municipios presentan criterios más vagos para su delimitación, aunque factores tan importantes en la toma de decisiones sobre el uso de suelo como las políticas públicas, se organizan a nivel municipal. Sin embargo, ambos tipos de unidades presentan grandes variaciones en el tamaño de los polígonos. En particular, como se puede observar en la tabla 6, los municipios más grandes (cuantil a 90 %) son 65 veces mayores que los más pequeños (cuantil 10 %) y el coeficiente de variación de la superficie es superior a 200 % (Tabla 6).

Tabla 6. Heterogeneidad en el tamaño de los polígonos por cada nivel de agregación (ha)

Unidades de agregación	Cuantil 10 %	Cuantil 90 %	Promedio	Desviación estándar	Coeficiente de variación
AGEBs	4283	16592	10762	10162	94 %
Municipios	2858	185432	76327	173855	228 %

En este estudio, se trató de limitar la multicolinealidad excluyendo del análisis las variables con altas correlaciones entre sí. Ambos modelos, global (nacional) y local (ventana) presentaron una variación de los resultados dependiendo de la agregación espacial. Sin embargo, no observamos que el problema se exacerbe en los modelos locales.

A fin de diseñar políticas, basadas en evidencias, orientadas a disminuir las tendencias en la pérdida de cobertura forestal, es necesario contar con datos y modelos confiables sobre la deforestación y las causas asociadas. Las variables socioeconómicos pueden contribuir a explicar los patrones de cambio en el uso del suelo, por ejemplo la densidad poblacional, los niveles de pobreza o los montos de los subsidios agrícolas, han sido variables que se han relacionado con la

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

presencia de bosques (Chowdhury, 2006; Dimobe *et al.*, 2015; Samndong *et al.*, 2018). Sin embargo, como puede derivarse de los resultados del presente trabajo, también es necesario que dichos datos estén desagregados a una escala similar a la de la cobertura forestal, de lo contrario se corre el riesgo de obtener conclusiones erróneas sobre este tipo de relaciones. Algunas instituciones encargadas de generar y distribuir datos socioeconómicos frecuentemente los presentan agregados a nivel municipal, resultados como los acá mostrados podrían justificar la importancia de generar información con mayor resolución espacial.

En conclusión, se observó que el nivel de agregación de la información afectó los valores del coeficiente de correlación y el ajuste de los modelos de regresión. El MAUP tuvo un efecto sustancial en particular, cuando no hay una fuerte relación entre la variable dependiente y las variables explicativas, como es a menudo el caso en la modelación de procesos complejos que involucran variables físicas y sociales. En el caso de México, la disponibilidad de datos censales para las unidades más detalladas (AGEBs rurales, manzanas en las áreas urbanas) podría aminorar los efectos del MAUP. No obstante, sin duda este efecto genera incertidumbre sobre la validez del modelo y la interpretación de sus resultados y deja abiertas algunas cuestiones a evaluar sobre la escala adecuada en la que deben plantearse las preguntas de investigación, la unidad de análisis espacial congruente con estas preguntas y por tanto el alcance explicativo de las variables elegidas para construir el modelo estadístico. Cuestiones que deberían ser expuestas en la presentación de cada modelo que se proponga explicar los CCUS tanto en la etapa de validación como en la de interpretación de resultados.

Agradecimientos

Este estudio se llevó a cabo en el ámbito del proyecto *Análisis espacio-temporal de la vulnerabilidad del paisaje utilizando percepción remota y métodos espaciales: un estudio interdisciplinario y multiescalar en cuatro regiones del país* financiado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) y el Instituto Nacional de Estadística y Geografía (INEGI). El artículo se elaboró durante una estancia sabática del primer autor en la *Universidade Federal da Bahia* y la *Universidade Estadual de Feira de Santana*, Brasil con el apoyo de PASPA-DGAPA – UNAM. Gabriela Cuevas participó en la elaboración de las bases de datos espaciales.

Referencias bibliográficas

Avelino, A.F.T., Baylis, K., & Honey-Rosés, J. (2016): "Goldilocks and the Raster Grid: Selecting Scale when Evaluating Conservation Programs", *PLoS ONE,* 11(12), e0167945. Disponible en http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0167945

Belsley, D.A., Kuh, E., & Welsch, R.E. (1980): Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New Jersey: John Wiley & Sons.

Bravo Peña L.C., Torres Olave M. E., Alatorre Cejudo L.C., Castellanos Villegas A.E., Moreno Murrieta R.L., Granados Olivas A., Uc Campos M., González León M., & Wiebe Quintana L.C. (2017): "Áreas probables de degradación-deforestación de la cubierta vegetal en Chihuahua, México. Una exploración mediante regresión logística para el período 1985-2013", GeoFocus, 20, 109-137. http://dx.doi.org/10.21138/GF.545

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

Brown, D.G., Verburg, P.H. Pontius, R.G., & Lange, M.D. (2013): "Opportunities to improve impact, integration, and evaluation of land change models", Current Opinion in Environmental Sustainability, 5, 452–457.

Butkiewicz, T., Meentemeyer, R.K., Shoemaker, D.A., Chang, R., Wartell, Z. & Ribarsky, W. (2010): "Alleviating the Modifiable Areal Unit Problem within Probe-Based Geospatial Analyses." Computer Graphics Forum 29 (3): 923–932.

Camacho Olmedo, M.T., Paegelow, M., Mas, J.F, & Escobar, F. (Editores) (2018). Geomatic Approaches for Modeling Land Change Scenarios, Series: Lecture Notes in Geoinformation and Cartography. New York: Springer.

Chowdhury, R.R. (2006): "Landscape change in the Calakmul biosphere reserve, Mexico: Modeling the driving forces of smallholder deforestation in land parcels", Applied Geography, 26: 129-152.

Clark, A., & Scott, D. (2013): "Understanding the Impact of the Modifiable Areal Unit Problem on the Relationship between Active Travel and the Built Environment", Urban Studies, 1-16.

CONAPO (2010): Consejo Nacional de Población, índices de marginación, México.

Diez Roux, A.V. (2008): "La necesidad de un enfoque multinivel en epidemiología", Región y sociedad, 20(spe2), 77-91.

Dimobe, K., Ouédraogo, A., Soma, S., Goetze, D., Porembski, S., & Thiombiano, A. (2015): "Identication of driving factors ofland degradation and deforestation in the Wildlife Reserve of Bontioli (Burkina Faso, West Africa)", Global Ecology and Conservation, 4, 559-57

Flowerdew, R. (2011): "How serious is the Modifiable Areal Unit Problem for analysis of English census data?", *Population Trends*, 145, 106-118.

Fotheringham, A.S., & Wong, D.W.S. (1991): "The modifiable areal unit problem in statistical analysis", *Environment and Planning A*, 23, 1025-1044.

Fotheringham, A.S., & Oshan, T.M. (2016): "Geographically weighted regression and multicollinearity: dispelling the myth." Journal of Geographical Systems 18 (4): 303-329.

Gerell, M. (2017): "Smallest is Better? The Spatial Distribution of Arson and the Modifiable Areal Unit Problem", *Journal of Quantitative Criminology*, 33(2), 293-318.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., & Townshend, J.R.G. (2013): "High-Resolution Global Maps of 21st-Century Forest Cover Change", *Science,* 15, 342 (6160), 850-853.

Hayward, P., & Parent, J. (2009): "Modeling the influence of the modifiable areal unit problem (MAUP) on poverty in Pennsylvania", *The Pennsvlvania Geographer*, 47(1), 120-135.

INEGI (2010a): Compendio de criterios y especificaciones técnicas para la generación de datos e información de carácter fundamental: Marco geoestadístico, 16 p. Disponible en http://www.inegi.org.mx/inegi/SPC/doc/INTERNET/16-marco_geoestadistico_nacional.pdf

INEGI, (2010b): Censo de Población y Vivienda 2010, Aguascalientes, México.

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

Jelinski, D. E. & Wu, J. (1996): "The modifiable areal unit problem and implications for landscape ecology", *Landscape Ecology*, 11, 129-140.

Kindu, M., Schneider, T., Teketay, D., & Knoke, T. (2015): "Drivers of land use/land cover changes in Munessa-Shashemene landscape of the south-central highlands of Ethiopia", Environmental Monitoring and Assessment,187(7), 4671.

Lo, C.P., & Yang, X. (2002): "Drivers of Land-Use/Land-Cover Changes and Dynamics Modeling for the Atlanta, Georgia Metropolitan Area", Photogrammetric Engineering & Remote Sensing, 68(10), 1073-1082.

Mahmood, R, Pielke, R.A., & McAlpine, C.A. (2016): "Climate-relevant land use and land cover change policies", Bull. Am. Meteorol. Soc., 97, 195–202.

Marshall, M., Norton-Griffiths, M., Herr, H., Lamprey R., Sheffield, J., Vagen, T., & Okotto-Okotto, J. (2017): "Continuous and consistent land use/cover change estimates using socio-ecological data", Earth System Dynamics, 8, 55-73.

Mertens, B., & Lambin, E. (2000): "Land-Cover Change Trajectories in Southern Cameroon", Annals of the Association of American Geographers, 90(3), 467-494.

NRC (2013) *Advancing Land Change Modeling: Opportunities and Research Requirements*. National Research Council, Washington, DC: The National Academies Press

Openshaw, S. (1984): "The modifiable areal unit problem Concepts and Techniques" in Modern Geography No. 28 Geo Books, Norwich.

Openshaw, S., & Rao, L. (1995): "Algorithms for reengineering 1991 Census geography", *Environment and Planning A*, 27(3), 425-446.

Parenteau, M.P., & Sawada, M.C. (2011): "The modifiable areal unit problem (MAUP) in the relationship between exposure to NO_2 and respiratory health", *International Journal of Health Geographics,* 10:58. Disponible en http://www.ij-healthgeographics.com/content/10/1/58

Pietrzak, M.B. (2014): "The Modifiable Areal Unit Problem – Analysis of Correlation and Regression", *Equilibrium Quarterly Journal of Economics and Economic Policy*, 9(3), 113-131.

Pineda, N., Bosque, J., Gómez, M., & Plata, W. (2009): "Análisis de cambio del uso del suelo en el Estado de México mediante Sistemas de Información Geográfica y técnicas de regresión multivariantes. Una aproximación a los procesos de deforestación", Investigaciones Geográficas, 69, 33-52.

R Core Team (2018): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Samndong, R.A., Bush, G., Vatn, A., & Chapman, M. (2018): "Institutional analysis of causes of deforestation in REDD+ pilot sites in the Equateur Province: Implications for REDD+ in the Democratic Republic of Congo", Land Use Policy. In press

Soares-Filho, B., Rodrigues, H., & Follador, M. (2013): "A hybrid analytical-heuristic method for calibrating land-use change models". Environmental Modelling & Software, 43, 80-87.

Tapia Silva, F. O., & López Flores E. (2017): "Variabilidad espacio-temporal de la cobertura terrestre en la cuenca del río Tecolutla, México", GeoFocus, 20, 163-182.

Mas, J-F., Pérez Vega, A., Andablo Reyes, A., Castillo Santiago, M. Á. (2018): "Incertidumbre de modelos estadísticos asociada a los niveles de agregación de la información espacial", GeoFocus (Artículos), nº 21, p. 169-186. ISSN: 1578-5157 http://dx.doi.org/10.21138/GF.585

Vitali, A., Urbinati, C., Weisberg, P. J., Urza, A.K. & Garbarino, M. (2018): "Effects of natural and anthropogenic drivers on land-cover change and treeline dynamics in the Apennines (Italy)", *Journal of Vegetation Science*, 1-11.

Vliet, J., Bregt, A. K. Brown, D. G. Delden, H. Heckbert, S. & Verburg, P. H. (2016): "A review of current calibration and validation practices in land-change modeling". Environmental Modelling & Software, 82, 174-182.

Weir-Smith, G. (2016): "Changing boundaries: Overcoming modifiable areal unit problems related to unemployment data in South Africa", *South African Journal of Science*, 112(3/4), 8 p. Disponible en http://dx.doi.org/10.17159/sajs.2016/20150115

Wheeler, D., & Tiefelsdorf, M. (2005): "Multicollinearity and correlation among local regression coefficients in geographically weighted regression." Journal of Geographical Systems 7 (2): 161-187.

Xu, P., Huang, H., Dong, N., & Abdel-Aty, M. (2014): "Sensitivity analysis in the context of regional safety modeling: Identifying and assessing the modifiable areal unit problem," Accident Analysis & Prevention, 70, 110-120.