Miscelánea Matemática **58** extraordinario (2014) 87-107 SMM

# Modelos estocásticos espaciales aplicados a la construcción de mapas de distribución de especies biológicas: Comparación de metodologías

Carlos Díaz Avalos
Instituto de Investigaciones en Matemáticas Aplicadas
y en Sistemas
Universidad Nacional Autónoma de México, México D.F.
carlos@sigma.iimas.unam.mx

## 1. Introducción

La elaboración de mapas ha sido una actividad practicada desde hace muchos siglos, ya que estos son una forma gráfica sumamente útil para ubicar objetos o entes de interés dentro de un área geográfica. Quizá el primer uso que se dió a los mapas fue la ubicación de zonas dende la caza y la recolección eran adecuadas. Actualmente, los mapas son utilizados en la mayoría de las actividades humanas, por lo que la elaboración de mapas precisos resulta crucial para la planificación y ejecución de las labores cotidianas. Una de estas actividades es el manejo y conservación de la biodiversidad.

La ciencia de la biodiversidad es el estudio de la tendencia mostrada por la riqueza de ambientes biológicos Su principal objetivo es la generación y análisis de la información necesaria para el manejo y conservación de los recursos naturales, actividad compleja *per se* debido a las interacciones existentes entre las especies animales y vegetales con su medio ambiente. Para estudiar un sistema tan complejo como este, es necesario hacer una serie de simplificaciones y aproximaciones, con el objetivo de hacer posible el estudio de la biodiversidad dentro de un marco de referencia manejable desde el punto de vista tanto computacional como conceptual. Un componente esencial para el manejo y conservación de la biodiversidad son las bases de datos con la localización de los sitios donde las diferentes especies han sido observadas y las características medio ambientales de dichas localidades. Estas bases de

datos son potencialmente grandes, representan información fragmentada y muestran una gran variabilidad en cuanto a precisión y extensión.

La creciente disponibilidad de equipo de cómputo de alta velocidad permite el manejo de bases de datos de gran tamaño en la búsqueda de mejores respuestas al problema de conservación y manejo de especies. Asimismo ha permitido el desarrollo de métodos y algoritmos que buscan la construcción de mapas basados en algún criterio de optimalidad. Una problemática en la construcción de estos mapas es que en la mayoría de los casos, los datos solo incluyen sitios en donde se observó a la especie, pero no se registran localidades en donde la especie se buscó y no fue observada (Peterson *et al.*, 2002) es decir, la información disponible no son datos de presencia-ausencia. Existen varios enfoques para la elaboración de los mapas de distribución de especies. Debido a la naturaleza dicotómica de la información base, la mayoría de ellos utilizan variantes de la regresión logística (Osborne y Tigar, 1992, Buckland y Elston, 1993) con covariables basadas en información sobre las características medio ambientales de los sitios en que la especie ha sido observada. Es importante recalcar que para propósitos de toma de decisiones, es de mayor relevancia contar con mapas de probabilidad de presencia de la especie que con mapas de distribución, por lo que el problema de construcción de mapas de biodiversidad se puede considerar como un problema de reconstrucción de imágenes tal como propone Besag (1986). Entre estos métodos estan los basados en el ajuste de envolturas climáticas alrededor de los sitios donde la especie de interés fue observada (FLORAMAP, Busby, 1991), árboles de decisión (Stockwell et al. 1990, Moore *et al.* 1991), redes neuronales y clasificación Bayesiana (GARP, Stockwell 1993) y campos aleatorios markovianos (Augustin *et al.* 1996). Todos estos métodos han sido utilizados con mayor o menor éxito para la predicción de sitios de distribución de especies. Sin embargo, en todos los casos el planteamiento del problema es el mismo: Dada una región $\mathcal{D}$ dividida en $\mathcal{N}$ subregiones, los métodos mencionados anteriormente buscan clasificar cada región como habitada o no por la especie de interés a partir de información incompleta sobre presencia de la misma en un conjunto finito de puntos $\mathbf{s}_i, i = 1, \ldots, K$ , $K < \mathcal{N}$.

Debido a la forma en que funcionan los diferentes métodos, existe siempre una fracción desconocida de pixeles o regiones que son clasificados erróneamente, ya sea porque se estima que la especie está ausente en sitios donde sí habita o porque se estima que la especie está presente en áreas donde no habita. Estos errores se conocen como error de falso negativo y falso positivo respectivamente. Dado que métodos como FLORAMAP, GARP y el modelo autologístico son utilizados ampliamente en el análisis y cuantificación de la biodiversidad, es de interés comparar su desempeño en el análisis de datos. En este documento se

reportan los resultados de un estudio comparativo de los errores de clasificación cometidos con 4 diferentes metodologías: FLORAMAP, dos versiones del algoritmos GARP y una modificación al modelo autologístico de Besag (1974) desarrollada por el autor . Las comparaciones se hicieron con base en un mapa de distribución de una especie hipotética, obtenido mediante simulación a partir de datos de isotermalidad, isoyetas, elevación y vegetación potencial y para la cual la distribución geográfica exácta es conocida. A diferencia del uso de datos reales, el uso de datos hipotéticos permite cuantificar las tazas de error de clasificación.

## 2. Marco teórico y supuestos

Consideremos una región $\mathcal{D}$ dividida en $\mathcal{N}$ subregiones. Estas subregiones pueden ser por ejemplo pixeles en una imagen o los municipios de algún estado. El proceso de estimación del área de distribución geográfica de una especie en una región $\mathcal{D}$ puede concebirse como un proceso jerárquico, en donde el verdadero estado de la naturaleza es $\mathbf{x}$ para $\mathbf{x} \in \{0,1\}^{\mathcal{N}}$, y del cual contamos solo con información fragmentada. En la notación anterior, $\mathbf{x} = \{x_1, \ldots, x_{\mathcal{N}}\}$, donde $x_i$ es el estado de la $i$-ésima localidad (pixel, municipio, etc), de manera que $x_i$=1 denota la presencia de la especie. Los tres algoritmos comparados en este trabajo buscan reconstruir la imagen $\mathbf{x}$ a partir de observaciones $\mathbf{y}$ en las que $y_i = 1$ indica que la especie fue observada en la $i$-ésima localidad (y por lo tanto $x_i = 1$). Nótese que el hecho de que $y_i = 0$ no necesariamente implica que $x_i = 0$. De este modo, el problema a resolver en los tres casos es el de clasificar en forma binaria los pixeles (o regiones) en donde $y_i = 0$.

GARP es un algoritmo de clasificación basado en una serie de reglas lógicas y es considerado un proceso de modelación inferencial. Las reglas lógicas utilizadas por GARP se clasifican como:

- Atómicas, del tipo «Si $x$, $y$ y $z$, entonces se toma la decisión A».
- BIOCLIM. Estas reglas se basan en envolturas climáticas construídas a partir de los datos $\mathbf{y}$ y de un conjunto de variables ambientales $\mathbf{Z}$ utilizando distribuciones multivariadas para obtener $P(\ x_i = 1|\mathbf{Z})$. Estas reglas se construyen con base en el análisis de discriminantes.
- De rango, similares a las reglas BIOCLIM.
- Logit, basadas en el cálculo de probabilidad de presencia utilizando el predictor lineal de regresión logística evaluado en la localidad

$i$, es decir, si $\mathbf{z}_i$ es el vector de covariables correspondiente a la localidad $i$,

$$P(x_i = 1|\mathbf{z_i}, \beta) = \frac{\exp\{\mathbf{z}_i\beta\}}{1 + \exp\{\mathbf{z}_i\beta\}}.$$

En todos los casos, si la probabilidad calculada es mayor a un umbral (digamos 0.75) se concluye que $x_i = 1$. El valor del umbral es fijado por el analista.

El algoritmo utilizado por FLORAMAP se basa en el cálculo de probabilidades de presencia a partir de envolturas climáticas calculadas con base en el análisis de componentes principales y el análisis de discriminantes a partir de información sobre variables bioclimáticas. Un supuesto implícito es que los componentes principales obtenidos a partir de las covariables tienen distribución normal multivariada.

### 2.1 Modelo autologístico con covariables

Besag (1974) propuso el modelo autologístico

$$p(x_i|\mathbf{x}_{-i}) = \frac{\exp\left\{x_i[\gamma \sum_{j\sim i} x_j]\right\}}{1 + \exp\left\{\gamma \sum_{j\sim i} x_j\right\}} \tag{1}$$

para modelar retículas en las que la variable de interés es binaria. La notación $j \sim i$ significa que las localidades $i$ y $j$ son vecinas entre sí. Los detalles técnicos del modelo autologístico están fuera del contexto de este reporte y se sugiere al lector interesado lea el artículo original. El modelo autologístico nos dice que la probabilidad de que $x_i = 1$ dado el estado de los demás sitios depende del número de vecinos del sitio $i$ que tengan valor igual a 1. Nótese que $\gamma > 0$ promueve la formación de conglomerados (manchones), $\gamma < 0$ promueve la repulsión entre sitios con valores iguales. La estimación de parámetros en el modelo autologístico puede hacerse por pseudo-máxima verosimilitud o por el método Monte Carlo con cadenas de Markov. Sin embargo, estos métodos de estimación implican el conocimiento del estado de $\mathbf{x}$ en toda el área de interés. En las aplicaciones que aquí nos interesan, el verdadero estado de la naturaleza no es observado, aunque puede de cualquier modo idealizarse como un proceso espacial binario. Por facilidad de notación tomaremos

$$x_i = \begin{cases} 1, & \text{si la especie habita en el sitio } i; \\ 0, & \text{en otro caso} \end{cases}$$

y supondremos que todas aquellas localidades donde se sabe que la presencia de la especie es imposible han sido eliminadas del análisis. Así por ejemplo, en el caso de especies terrestres todos aquellos sitios

que representen lagos, presas, mar y lagunas costeras dentro de $\mathcal{D}$ serán eliminadas del análisis.

### 2.2 Especificación del modelo

Los valores de $\{x_i\}$ son desconocidos y en la práctica la información disponible proviene de observadores de campo, quienes reportan evidencia fragmentada del tipo $Y_i = 1$ solo para aquellas localidades donde $x_i = 1$, es decir, si $Y_i = 1$ necesariamente inferimos que $x_i = 1$. La conclusión anterior implica que las especie a la que pertenece el organismo observado se determina sin error. Nótese asimismo que para una localidad $i$, $y_i$ puede tomar el valor 0 ya sea porque la especie no habita en la localidad $i$ o porque aún habitando esa localidad, no fue observada durante la prospección. Si asignamos arbitrariamente el valor $Y = 0$ para estos casos tenemos

$$Y_i = \begin{cases} 1 & \text{Si la especie fue observada en el sitio i} \\ 0 & \text{la especie no fue detectada o la localidad } i \text{ no fue visitada} \\ 0 & \text{la especie no habita en la localidad } i \end{cases}$$

Siempre que se hace una prospección de organismos biológicos existe el riesgo de no detectar a la especie objetivo aún dentro de su área de distribución. Denotaremos como $g$ a la probabilidad de no detectar a la especie de interés aún cuando esta habite en la región $i$. La densidad de las observaciones $\mathbf{y}$ es una función que depende de la distribución verdadera $\mathbf{x}$ y de su detectabilidad $g$. Para cada observación, su función de densidad de probabilidad está dada por

$$f(y_i|x_i, g) = g^{x_i(1-y_i)}(1 - g^{x_i})^{y_i}$$

Nuestro interés sin embargo es la reconstrucción de la imagen $\mathbf{x} = (x_1, \ldots, x_n)$ a partir de las observaciones $\mathbf{y} = \{y_{1,\ldots,}y_n\}$. Si suponemos que las observaciones $y_i$ son condicionalmente independientes, la verosimilitud de la muestra es

$$L(\mathbf{x}, g; \mathbf{y}) = \prod_{i=1}^{n} g^{x_i(1-y_i)}(1 - g^{x_i})^{y_i} \tag{2}$$

Nuestro objetivo es hacer inferencias sobre el estado de la naturaleza $\mathbf{x}$ el cual no es observable. Dado nuestro conocimiento $\mathbf{y}$, es razonable hacer las inferencias sobre $\mathbf{x}$ vía su distribución condicional dada $\mathbf{y}$ es decir, utilizando

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})L(\mathbf{x}, g; \mathbf{y})$$

En el lenguaje bayesiano $p(\mathbf{x})$ es la distribución inicial o *a priori* de $\mathbf{x}$, la cual se obtiene ya sea con base en conocimiento previo sobre el fenómeno

que se estudia o en supuestos razonables sobre el mismo. El estimador máximo verosimil de $\mathbf{x}$ se obtendría simplemente maximizando (2) con respecto a $\mathbf{x}$. Esto sin embargo es complicado debido al elevado número de posibles configuraciones de $\mathbf{x}$, lo cual dificulta la evaluación de la constante de proporcionalidad en $p(\mathbf{x}|\,\mathbf{y})$. Una alternativa es hacer las inferencias sobre $\mathbf{x}$ utilizando métodos bayesianos, como se describe a continuación.

### 2.3 Estimación bayesiana de x

Supondremos que parte de la distribución espacial de la especie está explicada por un conjunto de covariables $\mathbf{z} = (z_1, \ldots, z_p)$. Estas covariables en general estan asociadas a las características físicas y climáticas del hábitat de la especie y se relacionan al vector $\mathbf{x}$ por medio de un predictor lineal del tipo $\alpha = \mathbf{z}^T\beta$. Otra parte de la estructura espacial se supondrá que es explicada por interacciones entre localidades vecinas.

Con base en los supuestos anteriores, la probabilidad de presencia de la especie en el sitio $i$ dados los valores de las covariables asociadas a este y dado el estatus de $\mathbf{x}$ en los sitios vecinos a $i$ se puede modelar como un proceso autologístico (Besag, 1974)

$$p(x_i|\beta, \gamma, \mathbf{x}_{-i}) = \frac{\exp\left\{x_i[z_i^T\beta + \gamma s(x_i)]\right\}}{1 + \exp\left\{z_i^T\beta + \gamma s(x_i)\right\}} \tag{3}$$

donde $s(x_i) = \#\left\{x_j : x_j = 1\right\} - \#\left\{x_j : x_j = 0\right\}$ y $\gamma$ es un parámetro que controla la interacción espacial entre sitios. Nótese que si $\gamma = 0$ la distribución de probabilidad de $\mathbf{x}$ corresponde a la obtenida mediante regresión logística. Bajo este esquema, la densidad conjunta de $\mathbf{x}$ puede escribirse como

$$\begin{aligned} p(\mathbf{x}|\beta, \gamma) &= \frac{\exp\left\{\mathbf{x}^T\mathbf{Z}\beta + \gamma \sum x_i x_j)\right\}}{\sum_{\mathbf{x}} \exp\left\{\mathbf{x}^T\mathbf{Z}\beta + \gamma \sum\sum x_i x_j\right\}} \\ &\propto \exp\left\{\mathbf{x}^T\mathbf{Z}\beta + \gamma \sum\sum x_i x_j\right\} \end{aligned}$$

De esta manera, la densidad conjunta de $\mathbf{x}$ y $\mathbf{y}$ dada $g, \beta$ y $\gamma$ es proporcional a

$$\left[\prod_{i=1}^{n} g^{x_i(1-y_i)}(1-g^{x_i})^{y_i}\right] \times \exp\left\{\mathbf{x}^T\mathbf{Z}\beta + \gamma \sum\sum x_i x_j\right\}$$

Bajo el enfoque bayesiano consideraremos también a $\beta$, $g$ y a $\gamma$ como variables aleatorias con distribuciones iniciales $N_p(0, \sigma^{-2}I)$, $\beta(a, b)$ y $\Gamma(c_1, d_1)$ respectivamente. Supondremos además que $\sigma^{2\sim}\Gamma(c_2, d_2)$ y que $\sigma^2, \beta$, $g$ y $\gamma$ son independientes. El modelo queda completamente especificado si además suponemos que las observaciones $y_i$ son condicionalmente independientes. La distribución conjunta de los componentes

del modelo es proporcional a

$$\left\{\prod_{i=1}^{N} g^{x_i(1-y_i)}(1-g_i^x)^{y_i}\right\} \exp\left\{\mathbf{x}^T\mathbf{Z}\beta + \gamma \sum\sum x_i x_j\right\} \sigma^{n/2} \times$$
$$\times \exp\{-\tfrac{\sigma^2}{2}\beta^T\beta\}\gamma^c \exp\{-d\gamma\}\sigma^{2c_2} \exp\{-d_2\sigma^2\}\tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} g^a(1-g)^b. \quad (4)$$

A partir de estas expresiones las condicionales completas son

$$p(x_i|\mathbf{x}_{-i},\gamma,g,y_i) = \frac{g^{x_i(1-y_i)}(1-g^{x_i})^{y_i}\exp\left\{x_i(z_i^T\beta+\gamma s(x_i))\right\}}{1+g^{x_i(1-y_i)}(1-g^{x_i})^{y_i}\exp\left\{z_i^T\beta+\gamma s(x_i)\right\}} \quad (5)$$

donde $\mathbf{x}_{-i}$ representa el estado del proceso $\mathbf{x}$ en todas las localidades excepto la $i$-ésima. En particular,

$$p(x_i|\mathbf{x}_{-i},\gamma,g,y_i) = \frac{g^{x_i}\exp\left\{x_i(z_i^T\beta+\gamma s(x_i)\right\}}{1+g^{x_i}\exp\left\{z_i^T\beta+\gamma s(x_i)\right\}} \quad (6)$$

es la distribución de probabilidad en los sitios donde la especie no ha sido reportada, que son en los que nos interesa hacer inferencias.

Para el resto de los parámetros en el modelo, las condicionales completas son

$$p(\beta|\cdot) \propto \left[\prod_{i=1}^{n} \frac{\exp\left\{x_i z_i^T\beta\right\}}{1+\exp\left\{z_i^T\beta+\gamma s(x_i)\right\}}\right] \exp\left\{-\frac{\sigma^2}{2}\beta^T\beta\right\}$$

$$p(\gamma|\cdot) \propto \left[\prod_{i=1}^{n} \frac{\exp\left\{x_i\gamma\sum_{i\sim j}x_j\right\}}{(1+\exp\left\{z_i^T\beta+\gamma s(x_i)\right\})}\right] \gamma^{c_1}\exp\{-d_1\gamma\}$$

$$p(\sigma^2|\cdot) \propto \sigma^{\frac{p}{2}+2c_2}\exp\{-(d_2+\sum\beta_j^2)\gamma\}\tilde{}\Gamma(\frac{p}{2}+2c_2, d_2+\sum\beta_j^2)$$

$$p(g|\cdot) \propto \left[\prod_{i=1}^{N} g^{x_i(1-y_i)}(1-g_i^x)^{y_i}\right] g^a(1-g)^b.$$

Con las distribuciones condicionales completas, la estimación de los parámetros del modelo se hizo utilizando el método Monte Carlo con Cadenas de Markov (MCCM). En este método se dan valores iniciales $\mathbf{x}^{\{0\}}, \beta^{\{0\}}, \gamma^{\{0\}}, g^{\{0\}}, \sigma^{2\{0\}}$ a los parámetros. En una segunda etapa, cada parámetro es actualizado mediante un valor simulado a partir de su distribución condicional completa. Por ejemplo, $\mathbf{x}^{\{1\}}$ se obtiene simulando una realización de $\mathbf{x}$ a partir de

$$p(x_i^{\{1\}}|\mathbf{x}_{-i},\gamma^{\{0\}},g^{\{0\}},y_i) = \frac{g^{\{0\}x_i}\exp\left\{x_i^{\{1\}}(z_i^T\beta^{\{0\}}+\gamma^{\{0\}}s(x_i)\right\}}{1+g^{\{0\}x_i}\exp\left\{z_i^T\beta^{\{0\}}+\gamma^{\{0\}}s(x_i)\right\}}.$$

A continuación se actualiza $\beta$ simulando una realización de

$$p(\beta^{\{1\}}|\cdot) \propto \left[\prod_{i=1}^{n} \frac{\exp\left\{x_i^{\{1\}} z_i^T \beta^{\{1\}}\right\}}{1+\exp\left\{z_i^T \beta^{\{1\}} + \gamma^{\{0\}} s(x_i)\right\}}\right] \exp\left\{-\frac{\sigma^2}{2}\beta^{\{1\}T}\beta^{\{1\}}\right\}$$

y así sucesivamente hasta que todos los parámetros han sido actualizados. Una actualización de todos los parámetros define una iteración del algoritmo MCCM. A medida que el número de iteraciones crece, las realizaciones obtenidas en el proceso de simulación se parecen cada vez más a valores realizados a partir de las distribuciones condicionales completas, y en ese momento podemos suponer que el algoritmo ha convergido. Al número de iteraciones necesarias para alcanzar la convergencia del MCCM se le conoce como «tiempo de calentamiento». A partir de este momento, los resultados de iteraciones subsecuentes se deben almacenar a fin de utilizarlos en la estimación de las cantidades de interés mediante promedios ergódicos. Por ejemplo, si se hicieron 10000 iteraciones y el tiempo de calentamiento fué de 3000 iteraciones, la probabilidad de presencia de la especie en la localidad $i$ se estima como

$$P[x_1 = 1] = \frac{1}{7000} \sum_{j=3001}^{10000} x_i^{\{j\}}. \tag{7}$$

Por ser promedios ergódicos estas cantidades convergen al verdadero valor del parámetro a medida que el número de iteraciones tiende a infinito, aunque 500 ó más iteraciones suelen dar una aproximación más que aceptable.

### 2.4 Aplicación del modelo a datos de presencia de una especie hipotética

El modelo anterior se aplicó a datos de presencia construidos artificialmente para una especie hipotética distribuída en la República Mexicana. La construcción de los datos, a los que de aquí en adelante nos referiremos como $\mathbf{x}_t$, se hizo utilizando coberturas para isotermalidad, precipitación, elevación y vegetación potencial de Rzedowski proporcionadas por la CONABIO. Se escribió un programa en lenguaje FORTRAN para implementar el proceso de estimación Bayesiana a partir de las distribuciones condicionales completas descritas en la sección anterior. En forma breve, si $\theta = (\theta_1, \ldots, \theta_k)$ son los parámetros del modelos,el algoritmo consiste en actualizar un parámetro $\theta_i$ a partir de una muestra simulada de su distribución condicional completa $p(\theta_i|\theta_{-i})$, manteniéndose los valores de los demás parámetros fijos. Con el valor obtenido $\theta_i^{(t)}$, se procede a actualizar otro parámetro $\theta_k$ a partir de

| Parámetro | Valor |
|---|---|
| $\beta_0$ (gran media) | 0.20 |
| $\beta_1$ (ISOTM) | 0.37 |
| $\beta_2$ (ISOYT) | 0.24 |
| $\beta_3$ (ELEV) | 0.001 |
| $\beta_4$ (VPOTR 1) | 0.00 |
| $\beta_5$ (VPOTR 2) | 0.00 |
| $\beta_6$ (VPOTR 3) | 0.00 |
| $\beta_7$ (VPOTR 4) | 0.00 |
| $\beta_8$ (VPOTR 5) | 0.00 |
| $\beta_9$ (VPOTR 6) | 0.00 |
| $\beta_{10}$ (VPOTR 7) | 0.00 |
| $\beta_{11}$ (VPOTR 8) | 0.50 |
| $\beta_{12}$ (VPOTR 9) | 0.00 |
| $\gamma$ | 0.00 |
| $g$ | 0.50 |

**Cuadro 1.** Valores de los parámetros del modelo autologístico utilizados en la simulación de un mapa de distribución para una especie hipotética.

una muestra simulada de $p(\theta_k|\theta_{-k}, \theta_i^{(t)})$. El cíclo en el cual se hace una actualización de todos los parámetros define una iteración del proceso.

## 3. Comparación de métodos de construcción de mapas de distribución

Para poder comparar el desempeño de los diferentes algoritmos de construcción de mapas de distribución es necesario conocer el verdadero estado de la naturaleza $\mathbf{x}_t$, es decir, es necesario conocer para cada pixel si $x_i = 1$ o $x_i = 0$ para toda $i$. Esto es prácticamente imposible para especies biológicas reales, ya que si se tuviera un método exacto para conocer $\mathbf{x}_t$ los algoritmos aquí comparados no serían necesarios. Una práctica común para comparar algoritmos es construir un conjunto de datos artificialmente mediante simulación y utilizar dicho conjunto para generar pseudo-observaciones.

En nuestro estudio el procedimiento de obtención de $\mathbf{x}_t$ fue a través de datos simulados del modelo autologístico (3), manteniendo fijos los valores de $\beta$y $\gamma$ de acuerdo a lo mostrado en el cuadro 1.

Con los valores especificados en el cuadro se hicieron 1000 simulaciones del modelo autologístico. El resultado de la iteración número 1000 se tomó como la imagen verdadera de presencia de la especie ($\mathbf{x}_t$ ) en el territorio nacional. El mapa de distribución geográfica resultante

se presenta en la figura 1. De los 380000 pixeles que conforman dicha imagen, se tomó al azar una muestra de tamaño 10000, los cuales representan en nuestro experimento los sitios visitados por observadores de campo. De estos 10000 pixeles, solo se retuvieron aquellos en donde $x_i = 1$, dando un total de 975 sitios en donde la especie se reportó como presente y que en nuestro experimento representarán a las observaciones $\mathbf{y}$. De estos 975 puntos se tomaron submuestras de tamaños 490, 240, 176, 50, 30, 10 y 5 con el objeto de evaluar el efecto del número de sitios con registro de la especie sobre la capacidad clasificatoria de los tres algoritmos comparados en este estudio. La localización geográfica de los puntos de cada submuestra se presenta en la figura 2.

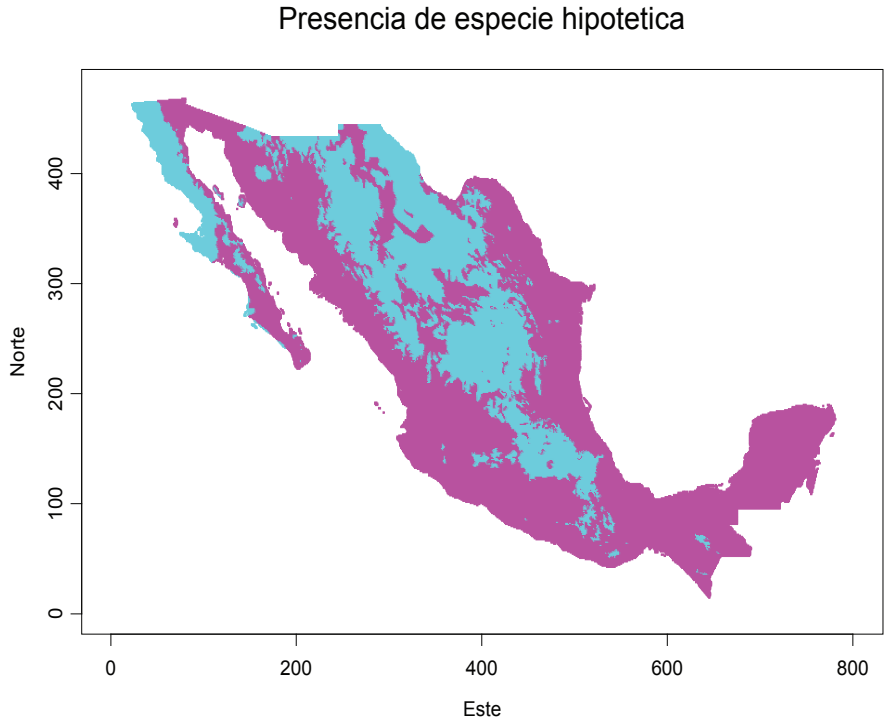


**Figura 1.** Presencia de especie hipotética en la República Mexicana (tono claro).

Los datos de las diferentes submuestras descritas en la sección anterior se utilizaron como información para construir mapas de presencia utilizando el método de algoritmos genéticos de clasificación (GARP), el de envolturas climáticas (FLORAMAP) y el modelo autologístico jerárquico descrito en este reporte. Se utilizó una imagen base de la República Mexicana dividida en pixeles cuadrados con tamaño de 0.04 grados por lado, con margen izquierdo (Este) en 118°W y margen inferior (Sur) en 14°N, dando una resolución de $800 \times 475$. Las covariables

utilizadas en el análisis fueron mapas digitalizados de isotermalidad, isoyetalidad, elevación y vegetacion potencial de Rzedowski con la misma resolución y georeferencia que el mapa base. En todos los casos, los pixeles correspondientes a cobertura oceánica y a cuerpos de agua fueron codificados como datos faltantes a fin de eliminarlos del análisis.

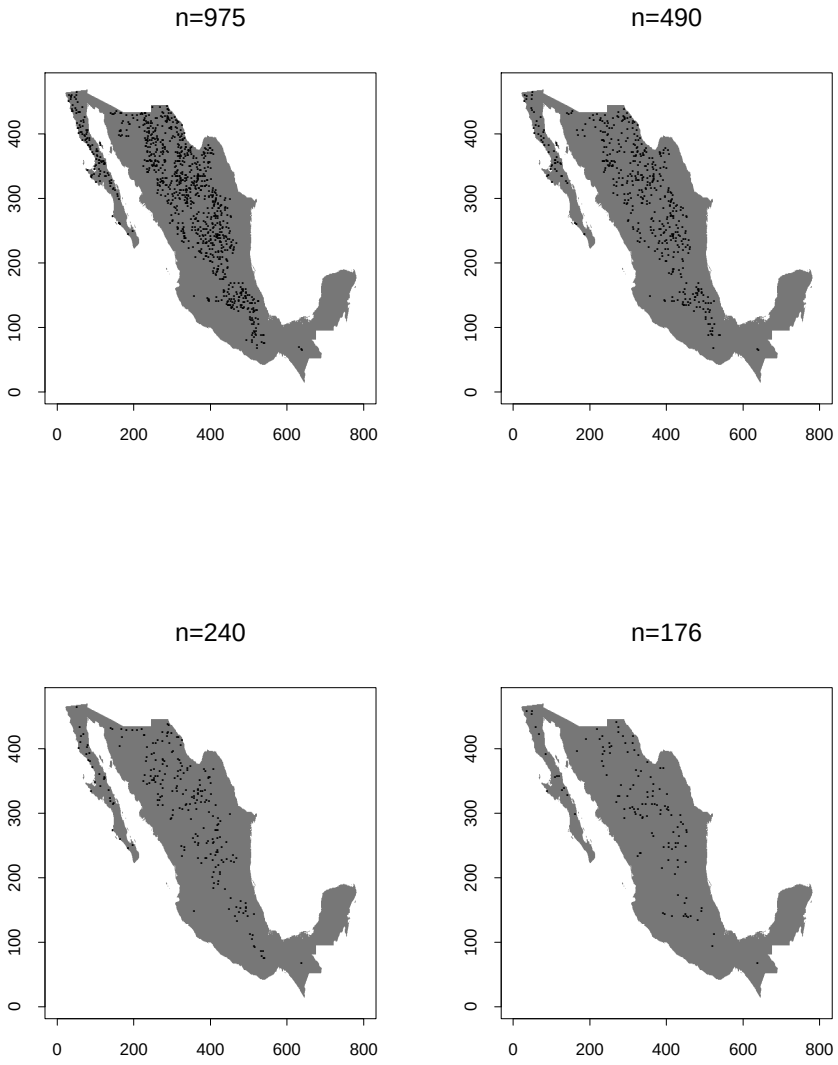


**Figura 2.** Distribución geográfica de las pseudo-observaciones.

Para comparar el desempeño de los métodos conocidos como GARP y FLORAMAP se hicieron corridas con cada método utilizando los datos de presencia mostrados en la figura 2. En el caso de GARP se utilizo el paquete DESKTOP GARP proporcionado por la CONABIO y una versión del algoritmo GARP implementada en sistema LINUX (LINUX GARP). Debido a la forma en que opera este programa, el cual

guarda un mapa de presencia para cada iteración lo cual ocupa mucho espacio de almacenamiento, solo se hicieron $J = 100$ iteraciones del algoritmo. Para cada pixel, la probabilidad de presencia se estimó con la proporción de veces que $x_{ij} = 1$, es decir,

$$p_i = \frac{\sum_{j=1}^{100} I_{[x_{ij}=1]}}{100}$$

Para el modelo autologístico jerárquico se hicieron 10000 iteraciones del algoritmo MMCM descrito en la sección anterior. Como se mencionó anteriormente, en este método se debe dejar pasar un cierto número de iteraciones para permitir al algoritmo converger a las distribuciones condicionales completas y garantizar que las simulaciones consecuentes corresponden al proceso que se pretende estimar. En nuestro caso, se consideró que después de 5000 simulaciones se había alcanzado este objetivo, por lo que en la construcción del mapa de probabilidad de presencia solo se utilizaron los resultados de las últimas 5000 simulaciones. Cada corrida con 10000 simulaciones tomó menos de 40 minutos en una procesador PENTIUM III. La probabilidad de presencia bajo el modelo autologístico se estimo utilizando la ecuación (8).

### 3.1 Comparación de la capacidad clasificatoria de los métodos autologístico, GARP y FLORAMAP. Curva de operación del receptor

Los mapas de probabilidad de presencia para con los cuatro algoritmos bajo comparación con N=975 y N=176 pseudo-observaciones se presentan en las figuras 3 y 4. En ambas figuras los tonos claros corresponden a alta probabilidad de presencia. Para N=976 se puede observar cierta similitud entre el patrón de probabilidades y el mapa de presencia verdadera de la figura 1, aunque es aparente que los 4 métodos sobreestiman el área en donde la especie probablemente está presente. Con N=176 pseudo observaciones el patrón de probabilidades de presencia se parece más al mapa de la figura 1, excepto para el mapa de probabilidades obtenido con Desktop GARP, el cual sigue sobreestimando el área de presencia probable.

Para un nivel fijo de $\pi$ siempre se corre el riesgo de tomar una mala decisión, de tal modo que existe el riesgo tanto de decidir que la especie esta presente cuando en realidad $x_i = 0$ o decidir que la especie esta ausente cuando el valor verdadero de $x_i$ es igual a uno. Tanto en el mapa de probabilidad construído con el modelo autologístico como con el obtenido con GARP se corren ambos riesgos, de modo que la comparación de la calidad del mapa construído con cada método debe hacerse en término de estos dos tipos de error. Una manera adecuada de hacer

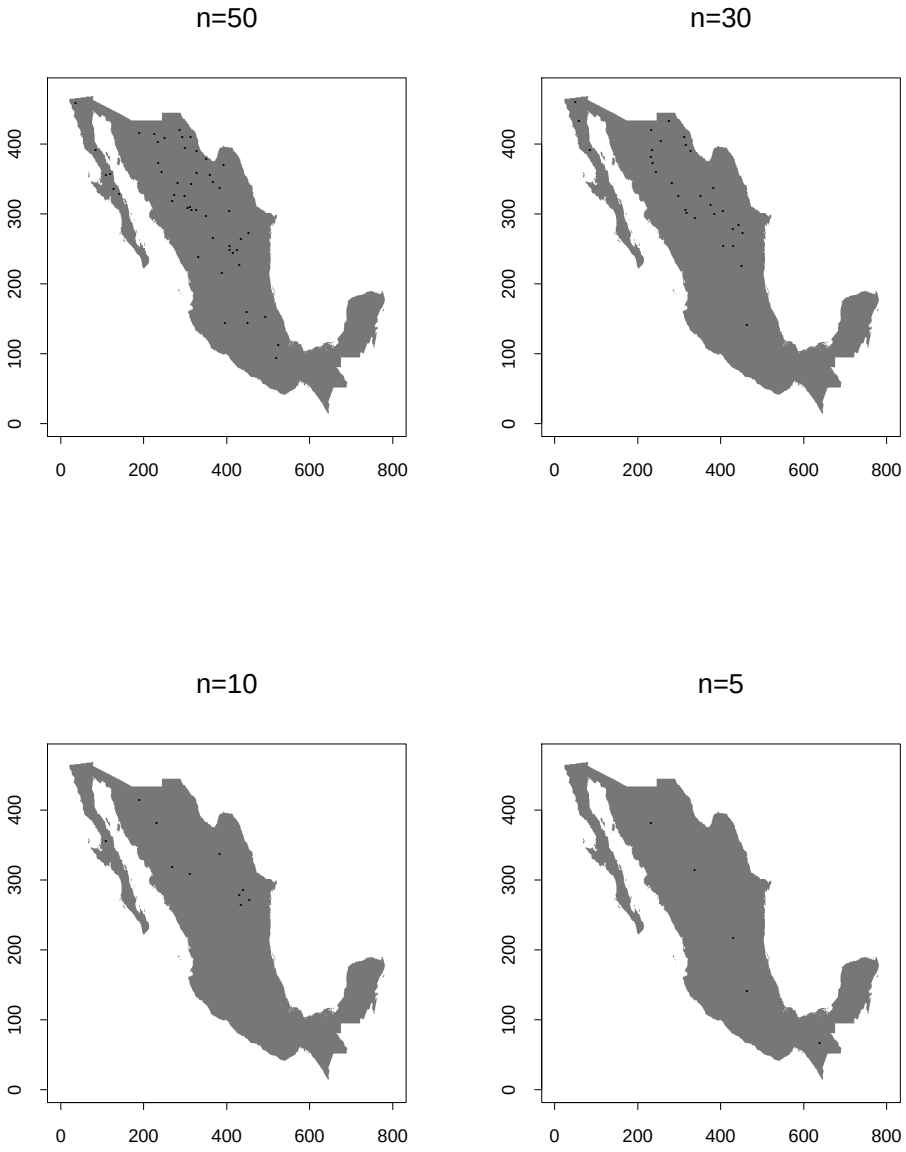


**Figura 3.** Distribución geográfica de las pseudo-observaciones.

esta comparación es mediante la curva característica de operación del receptor (CCOR) (Egan, 1975), la cual se obtiene graficando la taza de falso positivo (TFP) vs la taza de verdadero positivo (TVP) obtenidas con diferentes valores de $\pi$ La taza de falso positivo y de verdadero positivo se definen como

$$TFP(\pi) = \frac{\sum_{i=1}^{N} I_{[p_i>\pi]} I_{[x_i=0]}}{\sum_{1=1}^{N}(1-x_i)}, \qquad TVP(\pi) = \frac{\sum_{i=1}^{N} I_{[p_i>\pi]} I_{[x_i=1]}}{\sum_{1=1}^{N} x_i}.$$

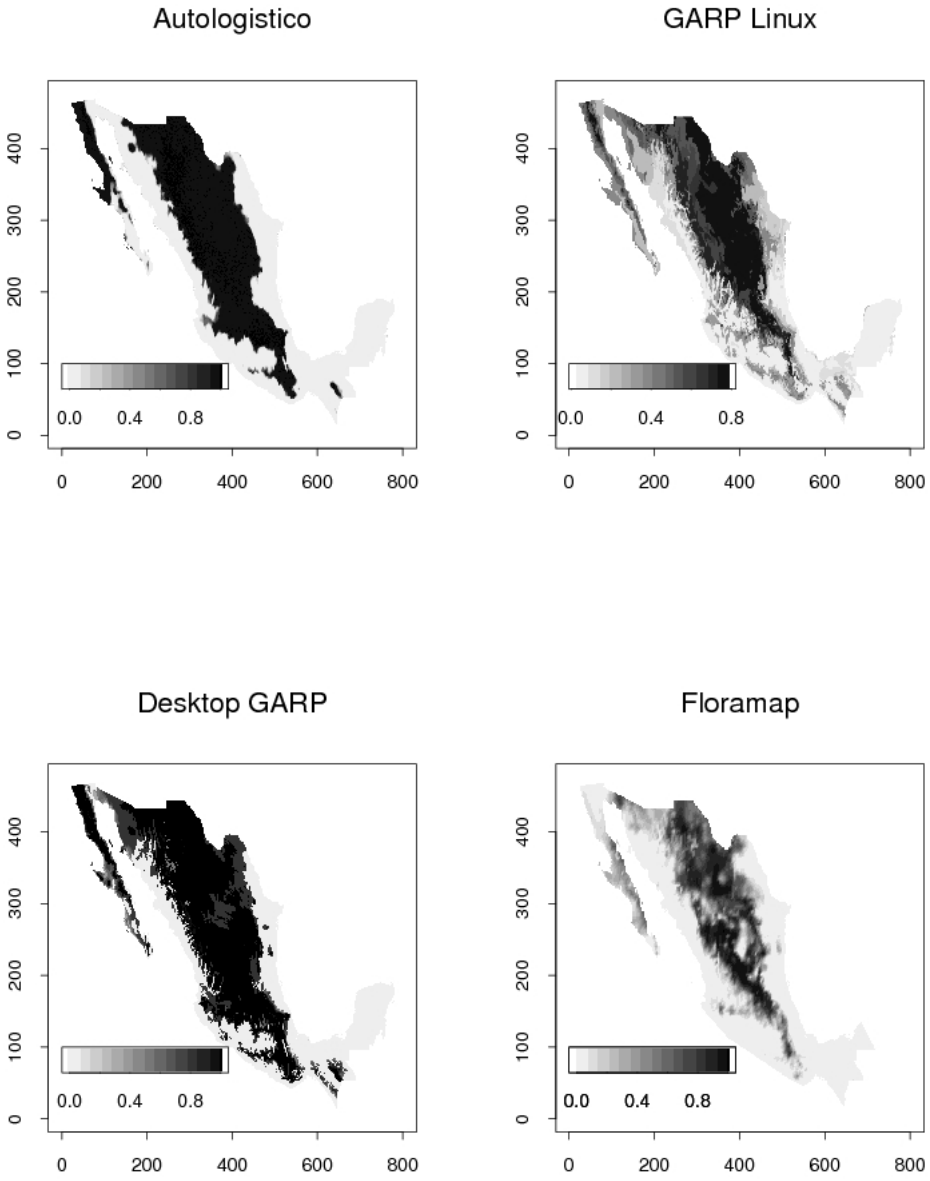


**Figura 4.** Mapas de probabilidad para N=975.

Para cada valor del umbral de clasificación $\pi$ la TVP representa la proporción de pixeles clasificados con presencia de la especie y que que $x_{it} = 1$, es decir, la proporción de clasificaciones positivas correctas. Por su parte, TFP es la proporción de pixeles clasificados con ausencia de la especie cuando la especie está presente, esto es, TFP representa la proporción de clasificaciones negativas erróneas. Nótese que TVP y TFP son calculados a partir de muestras y que por lo tanto están sujetas a variabilidad muestral. A la gráfica de TFP *vs* TVP se le conoce

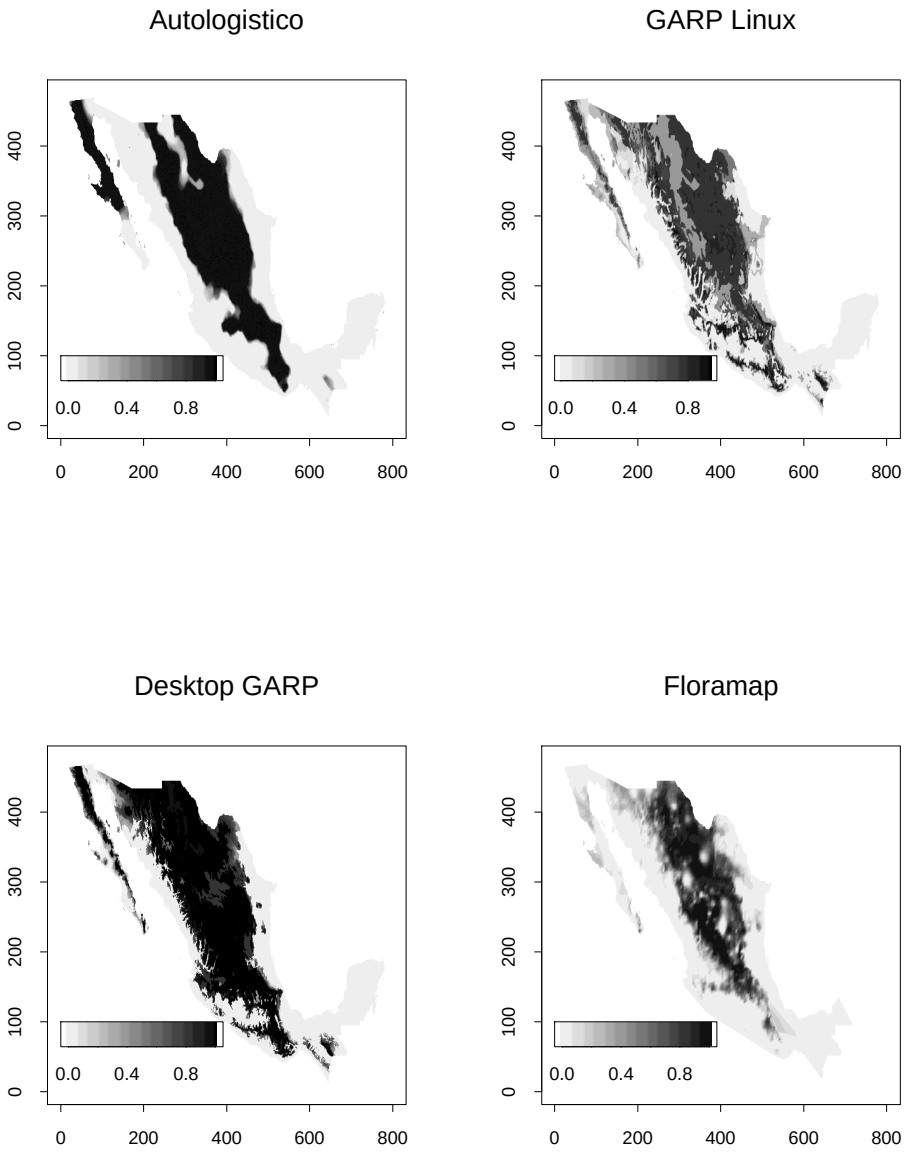


**Figura 5.** Mapas de probabilidad para N=50.

como curva de operación del receptor, término heredado de su orígen en el análisis de imágenes de radar en la segunda guerra mundial. La curva de operación ideal une los puntos $(0,0)$, $(0,1)$ y $(1,1)$ y cuando la clasificación de los pixeles se hace completamente al azar, la curva de operación es la bisectriz. Esto permite comparar los resultados obtenidos con los 4 métodos aquí analizados, ya que un buen desempeño debe verse reflejado en una curva de operación alejada de la bisectriz y cercana a la curva ideal. El área bajo la curva en el intervalo $(0,1)$ es

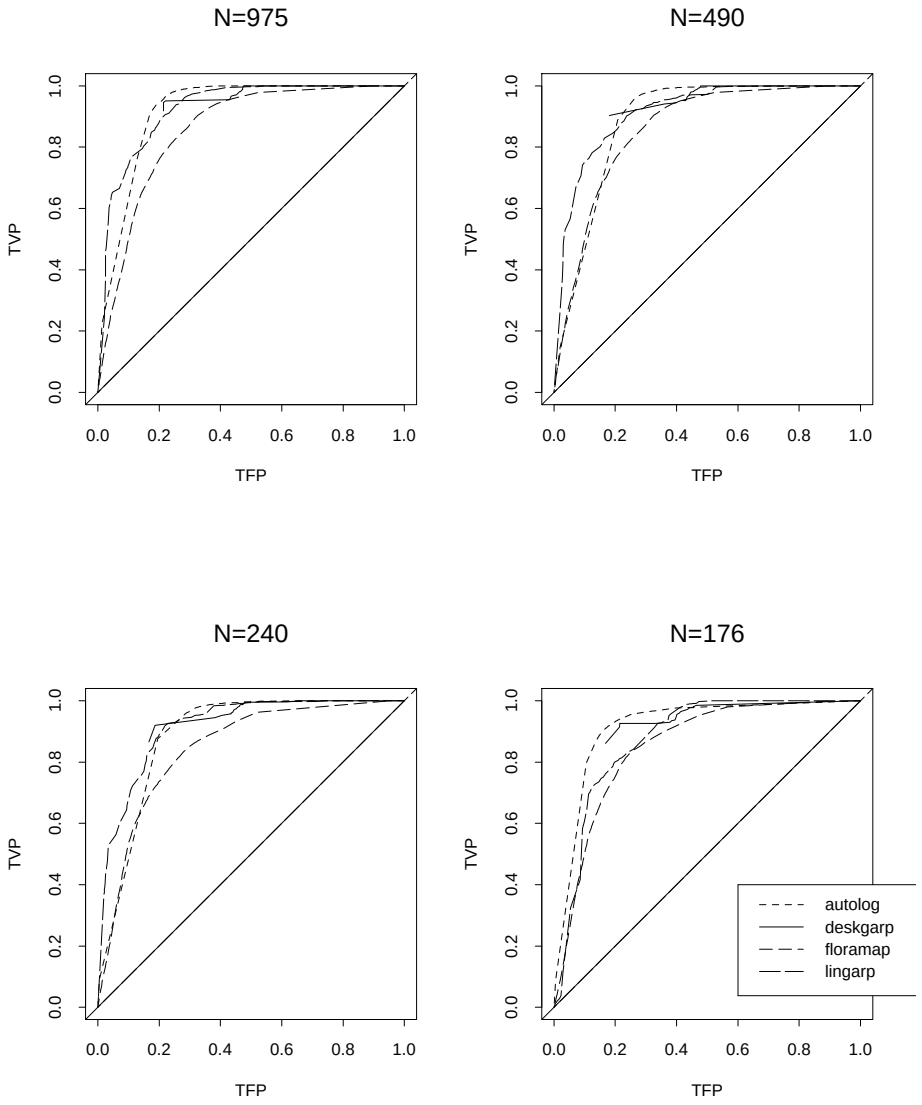


**Figura 6.** Curvas de operación del receptor para los 4 métodos comparados con diferente número de pseudo-observaciones.

la medida aceptada de desempeño de un método clasificatorio (Egan, 1975).

La curva de operación característica para los 4 métodos con tamaño de muestra 975, 490, 240 y 176 se presenta en la figura 5. Estos números de pseudo-observaciones son relativamente grandes comparados con el tamaño de la mayoría de las bases de datos en aplicaciones reales. Como se puede apreciar de la figura, el desempeño global de los 4 métodos es muy similar y debido a la presencia de variabilidad muestral es muy probable una prueba de diferencia entre los valores del área bajo la curva para los posibles pares de métodos (LINUXGARP-autologístico

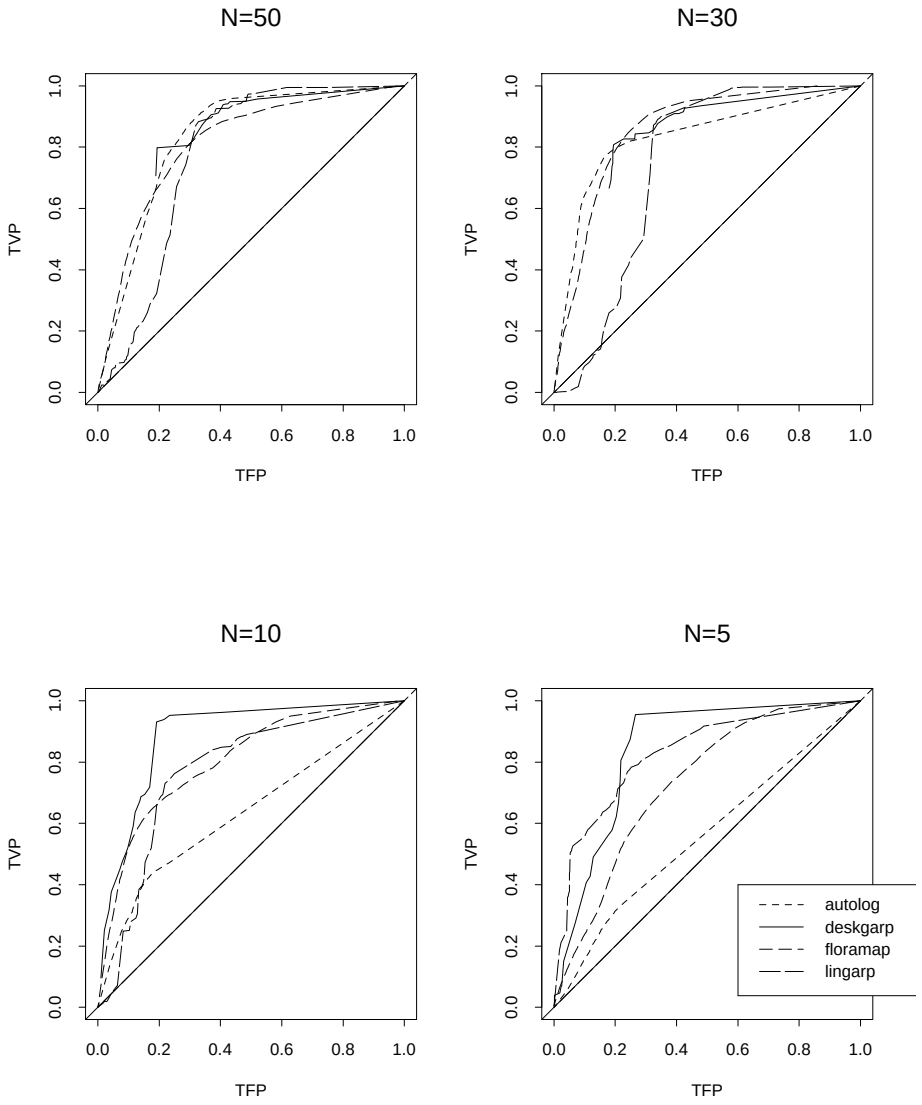


**Figura 7.** Curvas de operación del receptor para los 4 métodos comparados con diferente número de pseudo-observaciones.

por ejemplo) resultara no significativa estadísticamente hablando. No es claro como construir estas pruebas apareadas ya que se desconoce la distribución exacta de el área bajo la curva, aunque una posibilidad es la construcción de pruebas basadas en el método Monte Carlo o alguna modificación del Bootstrap. Este tópico requiere mayor investigación. Cabe hacer notar la existencia de una fuente adicional de variabiliad asociada a FLORAMAP y a LINUXGARP, ya que estos métodos trabajan con pixeles de menor resolución a los empleados por los otros dos métodos. Para poder comparar las curvas TFP-TVP con LINUXGARP y FLORAMAP fue necesario transformar su resolución a pixeles

de 4×4 km mediante interpolación lineal, por lo que la varianza asociada a la probabilidad de presencia estimada bajo la resolución fina es mas elevada (Patil, *et al* 2001).

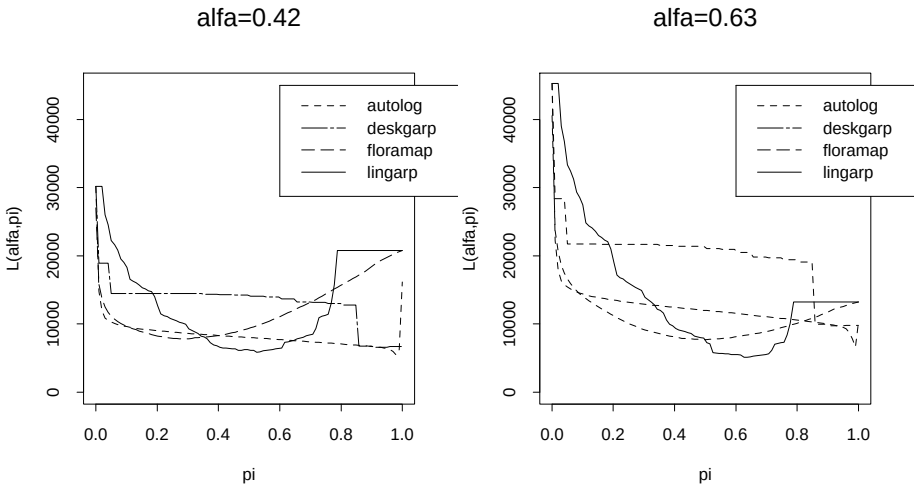


**Figura 8.** Función de pérdida para los diferentes métodos con N=975.

En la figura 6 se presentan las curvas de operación cuando el número de registros de presencia de la especie hipotética fue 50, 30, 10 y 5. Con excepción de la curva obtenida con FLORAMAP que mantiene una forma casi constante, las otras tres se acercan gradualmente a la bisectríz a medida que el número de registros decrece, haciendo evidente la sensibilidad de los algoritmos GARP y autologístico a la proporción de registros con respecto al tamaño del área de estudio. Por su parte, el poco cambio en la forma de la curva con FLORAMAP para cualquier número de registros de presencia sugiere que el algoritmo usado por este dos método es mas rígido en el sentido de que los resultados dependen más en los datos climáticos y no en las posiciones geográficas donde los datos fueron observados. La decisión sobre si la rigidez de un algoritmo es o no más deseable que la sensibilidad requiere de investigación en mayor detalle y esta fuera del alcance de este proyecto.

### 3.2 Comparación de algoritmos de clasificación bajo una función de pérdida

Otro posible enfoque para comparar entre los algoritmos de clasificación GARP, FLORAMAP y autologístico es proponer una función de pérdida basada en el costo de los errores de mala clasificación. Puesto que cada pixel puede ser clasificado como «ausencia» ($x_i = 0$) o «presencia» ($x_i = 1$), existen dos tipos de error:

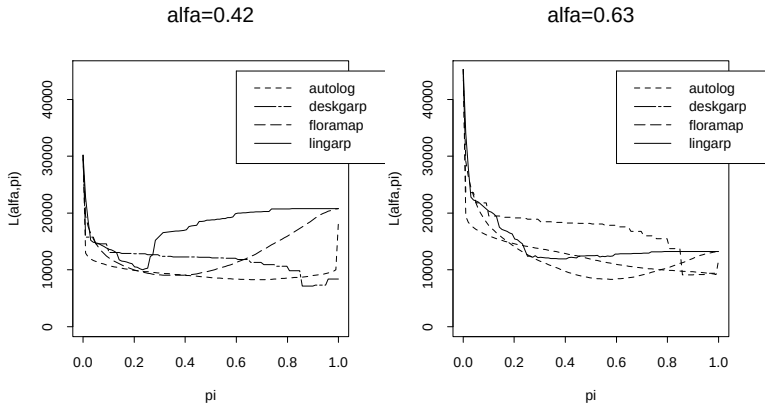


**Figura 9.** Función de pérdida para los diferentes métodos con N=50.

- $E_{01} = (p_i > \pi | x_i = 0) = I_{[p_i > \pi]} I_{[x_i = 0]}$, (clasificar una ausencia como presencia).
- $E_{10} = (p_i < \pi | x_i = 1) = I_{[p_i < \pi]} I_{[x_i = 1]}$, (clasificar una presencia como ausencia).

Suponiendo que $E_{01}$ y que $E_{10}$ tienen asociados costos $\alpha$ y $\phi$ respectivamente, $\phi = 1 - \alpha$ y que tanto $\alpha$ como $\phi$ son independientes de la localidad, entonces el costo total del error $E_{01}$ es

$$C_{01} = \alpha \sum_i I_{[p_i > \pi]} I_{[x_i = 0]},$$

y similarmente, para el error $E_{10}$ el costo total es

$$C_{10} = \phi \sum_i I_{[p_i < \pi]} I_{[x_i = 1]}.$$

Nótese que los costos dependen de $\mathbf{x}$ y de $\pi$, es decir, del verdadero estado de la naturaleza y del criterio de decisión. Una función de pérdida adecuada en este caso es la función de pérdida cuadrática

$$\Lambda(\mathbf{x}, \pi) = \sqrt[2]{C_{01}^2 + C_{10}^2}.$$

La figura 7 y 8 muestran la función de pérdida $\Lambda(\mathbf{x}, \pi)$ para cada algoritmo de clasificación cuando el número de registros fue 975 y 50 respectivamente, para dos valores de $\alpha$. Cuando $\alpha = 0.42$, el costo de clasificar ausencias como presencias es menor que el costo de clasificar presencias como ausencias, mientras que cuando $\alpha = 0.63$ se da el caso inverso.

Para $N = 975$ el mínimo de $\Lambda(\mathbf{x}, \pi)$ se obtiene con el modelo autologístico cuando $\alpha = 0.42$ y con LINUXGARP cuando $\alpha = 0.63$, aunque la localización de este valor crítico de la función de pérdida es distinta en ambos casos. Por otro lado, cuando $N = 50$ el mínimo de

$\Lambda(\mathbf{x}, \pi)$ se obtiene con el modelo autologístico cuando $\alpha = 0.42$ y con DESKTOPGARP cuando $\alpha = 0.63$. Al igual que en la figura 7, la localización del mínimo en ambos valores de $\alpha$ se da a distinto nivel del criterio de decisión $\pi$. Cabe hacer notar que ninguna función de pérdida es uniformemente menor que las demás, ya que en algunas regiones del eje $\pi$ alguno de los algoritmos se comporta mejor que los otros en cuanto a $\Lambda(\mathbf{x}, \pi)$ se refiere. Sin embargo la forma cuadrática de $\Lambda(\mathbf{x}, \pi)$ nos garantiza la existencia de un mínimo dentro del eje $\pi$, permitiendo la catalogar el desempeño de los diferentes algoritmos de clasificación en función del número de registros de presencia de la especie hipotética. El lector no debe perder de vista que el análisis comparativo tanto en término de la curva de operación como de la función de pérdida solo es posible hacerlo si se conoce el verdadero estado de la naturaleza $\mathbf{x}$, por lo que un análisis de este tipo difícilmente puede hacerse con especies biológicas verdaderas.

## 4. Conclusiones

El problema de construcción de mapas de distribución de especies biológicas a partir de información altamente fragmentada es complejo debido a los factores involucrados en la selección de hábitat de los organismos. Dado que los diferentes algoritmos que se compararon ponderan de manera distinta tanto a las observaciones como a la información adicional en forma de covariables, es natural esperar discrepancias entre los resultados.

Se encontró que las discrepancias entre los mapas de probabilidad de presencia con los diferentes algoritmos de clasificación son ignorables si la proporción de registros de presencia con respecto al tamaño del área de estudio es moderado (975/108000), y que esas discrepancias se acentúan al bajar dicha proporción. Los algoritmos varían en cuanto a sensibilidad a la proporción antes mencionada y no es claro que es preferible, si la sensibilidad mostrada por el modelo autologístico y GARP o la excesiva robustez mostrada por FLORAMAP. Este es un tópico que requiere mas investigación.

## Bibliografía

[1] N. H. Augustin, M. Mugglestone y S. Buckland, «An autologistic model for the spatial distribution of wildlife», *Journal of Applied Ecology*, vol. 33, 1996, 339–347.

[2] J. Besag, «Spatial interaction and the statistical analysis of lattice systems (with discussion)», *Journal of the Royal Statistical Society Series B*, vol. 40, 1974, 147–174.

[3] ———, «On the analysis of dirty pictures (with discussion)», *Journal of the Royal Statistical Society Series B*, vol. 48, 1986, 259–302.

[4] S. Buckland y D. Elston, «Empirical models for the spatial distribution of wildlife», *Journal of Applied Ecology*, vol. 30, 1993, 478–495.

[5] J. Busby, «Bioclim a bioclimatic and prediction system», en *C.R. Margules & M.P. Augustin Biological Conservation: Cost Effective Biological Surveys and Analysis*, East Melbourne, Australia: CSIRO Publications, 1991, 64–68.

[6] J. Egan, *Signal detection theory and roc analysis*, New York: Academic Press, 1975.

[7] D. Moore, B. Lees y S. Davey, «A new method for predicting vegetation distributions using decision tree analysis in a geographic information system», *Environmental Management*, vol. 15(1), 1991, 59–71.

[8] P. Osborne y B. Tigar, «Interpreting bird atlas data using logistic models: an example from lesotho, southern africa», *Journal of Applied Ecology*, vol. 29, 1992, 55–62.

[9] G. Patil, R. Brooks, W. Myers, D. Rapport y C. Taillie, «Ecosystem health and its measurement at landscape scale: Toward the next generation of quantitative assessments», *Ecosystem Health*, vol. 7(4), 2001, 307–316.

[10] A. Peterson, L. Ball y K. Cohon, «Predicting distributions of mexican birds using ecological niche modelling methods», *Ibis*, vol. 144, 2002, 27–32.

[11] D. Stockwell, «Lbs: Bayesian learning system for rapid expert system development», *Expert Systems With Applications*, vol. 6, 1993, 137–147.

[12] D. Stockwell, S. Davey, J. Davis y I. Noble, «Using induction of decision trees to predict greater glider density», *AI Applications in Natural Resource Management*, vol. 4(4), 1990, 33–43.

[13] D. Stockwell y I. Noble, «Induction of sets of rules from animal distribution data: a robust and informative method of data analysis», *Mathematics and Computers in Simulation*, vol. 33, 1992, 385–390.