MODELOS BAYESIANOS PARA MODELAR DISTRIBUCIONES DE ESPECIES CON REGISTROS DE SOLO PRESENCIAS

[BAYESIAN MODELS FOR SPECIES DISTRIBUTION MODELLING WITH ONLY-PRESENCE RECORDS]

[a]Bartolo de Jesús Villar Hernández, [b]Sergio Pérez Elizalde.

Colegio de Postgraduados, Campus Montecillo, C.P. 56230, Texcoco, Edo. de México, México. Emails: [a]bartolo.villar@colpos.mx, [b]sergiop@colpos.mx
**Corresponding author*

SUMMARY

One of the central issues in ecology is the study of geographical distribution of species of flora and fauna through Species Distribution Models (SDM). Recently, scientific interest has focused on presence-only records. Two recent approaches have been proposed for this problem: a model based on maximum likelihood method (*Maxlike*) and an inhomogeneous poisson process model (*IPP*). In this paper we discussed two bayesian approaches called *MaxBayes* and *IPPBayes* based on *Maxlike* and *IPP* model, respectively. To illustrate these proposals, we implemented two study examples: (1) both models were implemented on a simulated dataset, and (2) we modeled the potencial distribution of genus *Dalea* in the Tehuacan-Cuicatlán biosphere reserve with both models, the results was compared with that of *Maxent*. The results show that both models, *MaxBayes* and *IPPBayes*, are viable alternatives when species distribution are modeled with only-presence records. For simulated dataset, *MaxBayes* achieved prevalence estimation, even when the number of records was small. In the real dataset example, both models predict similar potential distributions like *Maxent* does.

Key words: only-presence records; species distribution models; ocurrence probability; inhomogeneous poisson procces; Maxlike; Maxent; bayesian approach.

RESUMEN

Uno de los temas centrales en ecología es el estudio de la distribución geográfica de especies tanto de flora como de fauna a través de Modelos de Distribución de Especies (MDE). Recientemente el interés científico se ha centrado en aquellos registros de solo presencias. Dos enfoques recientes se han propuesto para este problema: un modelo basado en el método de máxima verosimilitud (*Maxlike*) y un modelo de proceso Poisson no homogéneo (*IPP*). En este trabajo se discuten dos enfoques bayesianos denominados *MaxBayes* e *IPPBayes* construidos en base a los anteriores. Para ilustrar dichas propuestas, se implementaron dos ejemplos de estudio: (1) se implementaron ambos modelos en un conjunto de datos simulados, y (2) se modeló la distribución potencial del género *Dalea* en la reserva de la biosfera Tehuacán-Cuicatlán con ambos modelos, los resultados se compararon con los obtenidos mediante *Maxent*. Los resultados indican que ambos modelos aquí propuestos, constituyen alternativas viables cuando se modelan distribuciones de especies con registros de solo presencias. En el caso de datos simulados, *MaxBayes* logra estimar la prevalencia aún cuando el número de registros es pequeño. En el ejemplo con datos reales, ambos modelos predicen patrones de distribución similares a *Maxent*.

Palabras clave: registros de solo presencia; modelos de distribución de especies; probabilidad de ocurrencia; proceso Poisson no homogéneo; Maxlike; Maxent; enfoque bayesiano.

INTRODUCCIÓN

Uno de los temas centrales en ecología es el estudio de la distribución geográfica de especies tanto de flora como de fauna. Hoy en día, se han estado desarrollado modelos cuyo objetivo a grandes rasgos es modelar la distribución espacial de las especies de interés. Estos Modelos de Distribución de Especies (MDEs) se han utilizado para diversos propósitos, por ejemplo, han sido aplicados para estudiar la

propagación de especies intrusas (Thuiller *et al*., 2005), para investigar los impactos del cambio climático en la extinción de ciertas especies (Thomas *et al*., 2004), para conocer la diversidad biológica de una zona en particular (Graham *et al*., 2006), por citar solo algunos. En todas las aplicaciones de los MDEs, el problema central es utilizar la información de donde las especies han sido observadas (y donde no) y asociar ésta información con un conjunto de covariables medioambientales para determinar la probabilidad [o algún índice proporcional a ésta] de que una determinada especie pueda estar presente o no en sitios no muestreados (Latimer *et al*., 2006). Generalmente, el área de interés se divide en una malla de celdas del mismo tamaño, donde la elección del tamaño de las celdas quedará determinado según la resolución deseada por el investigador, y la probabilidad de presencia se generalizará para toda la celda.

Recientemente el interés científico se ha centrado en aquellos registros que provienen de herbarios, museos y colecciones privadas. Estos registros de *solo presencias* no provienen de un muestreo sistemático y en la mayoría de los casos presentan sesgo muestral, dado que fueron colectados cerca de carreteras, poblados o áreas de interés específicas (Fithian y Hastie, 2013). A la par del interés de los ecólogos de estudiar este tipo de datos, se han propuesto modelos estadísticos que abordan en menor o mayor medida el problema, por ejemplo, el implementado en el popular software *MaxEnt* (Phillips *et al*., 2004), y generalizaciones del modelo logístico. En la literatura científica de los últimos años, *Maxent* es el software más citado. Su amplia utilización se explica en parte por su facilidad de uso ya que funciona como una *caja negra* donde las únicas entradas que necesita el software son las ubicaciones georeferenciadas de los puntos de ocurrencia asociadas a un conjunto de covariables medioambientales. También se proporciona un archivo donde se especifica el mismo grupo de covariables correspondientes al *background* (una muestra aleatoria de ubicaciones provenientes de toda el área de interés). Aunado a lo anterior, en la mayoría de los trabajos que hacen uso de *Maxent* se ha hecho una incorrecta interpretación de la *salida logística* interpretando dicha salida como una estimación de la probabilidad de ocurrencia. Se ha ignorado el hecho de que *Maxent*, al tratar de aproximar la probabilidad de presencia, se asume que la prevalencia (proporción de sitios ocupados a través de toda el área de interés) es 0.5.

Dos enfoques recientes se han propuesto para abordar el problema de modelar distribuciones de especies con registros de *solo presencias*. El primero de ellos es el modelo *Maxlike* propuesto por Royle *et al*. (2012) con el que la probabilidad de ocurrencia (ψ) puede calcularse mediante el método de máxima verosimilitud. Para ello, con *Maxlike* se asume que los registros provienen de un muestreo aleatorio y que la probabilidad de detección es constante en la zona de interés. Otro propuesta para registros de *solo presencias* es un proceso Poisson no homogéneo (IPP) propuesto por (Fithian y Hastie, 2013; Warton y Shepherd, 2010) que modela la intensidad de ocurrencia, no la probabilidad de ocurrencia.

En el presente trabajo se proponen dos metodologías, en el marco de inferencia bayesiana, que se han denominado *MaxBayes* e *IPPBayes*. Estas dos alternativas se han construido a partir del modelo conceptual de *Maxlike* y el modelo *IPP*, respectivamente. Para ilustrar de forma práctica dichas propuestas, se implementaron dos ejemplos de estudio: (1) se simularon registros de presencia-ausencia mediante un ensayo *Bernoulli* en una zona ficticia de 10,000 celdas en donde la prevalencia fue de 0.38, de las cuales una vez descartadas las ausencias, se muestreo aleatoriamente las presencias que se utilizaron para ajustar los modelos y estimar sus respectivos parámetros, y en el caso de *MaxBayes* comparar la prevalencia estimada contra la real, y (2) se utilizaron registros de presencia del género *Dalea* provenientes de la zona de la reserva de la biosfera Tehuacán-Cuicatlán y se compararon las distribuciones potenciales arrojadas de ambos modelos con el software *Maxent*.

Los resultados indican que ambos modelos aquí propuestos, constituyen alternativas viables cuando se modelan distribuciones de especies con registros de *solo presencias*. En el caso del ejemplo con datos simulados y distribuciones *a priori* no informativas para los parámetros en el modelo *MaxBayes*, el modelo calcula una prevalencia muy cercana a la real, aún cuando el número de presencias es pequeño. Dicha estimación puede estar más cercana a la real en caso de utilizar distribuciones *a priori* informativas. Para los datos del género *Dalea*, tanto *MaxBayes* como el modelo *IPPBayes* predicen patrones de distribución potencial similares al obtenido con el software *Maxent*, aunque dicha similitud es más acentuada en el caso de *MaxBayes* e *IPPBayes*. La ventaja de *MaxBayes* sobre *Maxent*, es que el primero estima la prevalencia y por tanto también estima la probabilidad de ocurrencia, mientras que *Maxent* a través de su salida logística estima un índice que informa acerca de qué tan idóneo es un sitio para albergar a la especie con respecto a otros, y no es una probabilidad. Por otro lado *IPPBayes*, estima la intensidad de ocurrencia, es decir, el número esperado de especímenes por unidad de área y, cuando se utilizan distribuciones *a priori* no informativas para los parámetros, dicha intensidad es relativa al número de presencias utilizadas para ajustar el modelo.

MATERIALES Y MÉTODOS

Métodos para el análisis de datos de *solo presencias*

Método basado en la máxima verosimilitud (*Maxlike*).

Según Royle *et al*. (2012), cuando se toman muestras aleatorias de celdas $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, donde $\boldsymbol{\mathcal{X}}$ representa el conjunto de todos los valores posibles de $\boldsymbol{x}$, es decir, $\boldsymbol{x_1, \dots, x_N}$ y como registros $\boldsymbol{y_1, \dots, y_N}$, y posteriormente se descartan aquellas celdas donde no hay registro de la especie, entonces la prevalencia puede estimarse bajo el enfoque de máxima verosimilitud. La característica principal en los datos de *solo presencias* es que la variable $\boldsymbol{Y}$ ya no es aleatoria, debido a que $\boldsymbol{y = 1}$ con probabilidad 1 para todas las observaciones. En lugar de ello se asume que $\boldsymbol{X}$ es una $\boldsymbol{v.a.}$, y que el conjunto de $\boldsymbol{n}$ ubicaciones de presencia $\boldsymbol{x_1, \dots, x_n}$ son los datos sobre los cuales se basa la inferencia. Los valores de $\boldsymbol{x}$ que aparecen en la muestra representan el sesgo en la selección sobre el conjunto de todos los valores posibles $\boldsymbol{\mathcal{X}}$ favoreciendo aquellos donde $\boldsymbol{y = 1}$ (Royle *et al*., 2012).

En la notación de Royle *et al*. (2012), $\pi()$ representa la distribución de propabilidad de x , y $\psi()$ representa la distribución de probabilidad de y. La verosimilitud está basada en la distribución de probabilidad condicional de x para los cuales $y = 1$, es decir, $\pi(x|y = 1)$. Aplicando el teorema de Bayes se tiene

$$\pi(x|y = 1) = \frac{\psi(y=1|x)\pi(x)}{\psi(y=1)}. \qquad (1)$$

La ecuación (1) está expresada en términos del espacio geográfico, sin embargo, puede expresarse en términos del espacio medioambiental Z (donde Z es la $v.a.$, que denota a la variable medioambiental) siempre que las z sean muestreadas aleatoriamente. La distribución de probabilidad $\pi(x)$ es la que describe los resultados posibles de la variable aleatoria x (pixel identidad). Supóngase que el espacio de valores de x es discreto y que tiene M elementos únicos equiprobables, y por lo tanto, $\pi(x) = 1/M$. Por otra parte, $\psi(y = 1|x)$ es la probabilidad de que $y = 1$ condicionada por x y que se denomina como *probabilidad de ocurrencia*. Note que $\psi(y = 1)$ es la probabilidad marginal de que un pixel albergue a la especie, que por definición es $\psi(y = 1) = \sum_{x\in\mathbb{X}} \psi(y = 1|x)\pi(x)$, que es el **promedio** espacial de probabilidad de ocurrencia y que en la literatura se denomina *prevalencia* (Royle *et al*., 2012).

La Verosimilitud

Note que $\psi(y_i = 1|x_i)$ depende de algunos parámetros $\boldsymbol{\beta}$ asociados a las covariables medioambientales, por lo que podemos escribirla como $\psi(y_i = 1|x_i; \boldsymbol{\beta})$, por lo tanto, $\pi(x_i|y_i = 1) = \psi(y_i = 1|x_i; \boldsymbol{\beta})\pi(x_i)/\sum_{x\in\mathcal{X}} \psi(y_i = 1|x_i; \boldsymbol{\beta})\pi(x)$; donde $\pi(x_i)$ es constante y por tanto se cancela del numerador y del denominador resultando en

$$\pi(x_i|y_i = 1) = \frac{\psi(y_i=1|x_i;\boldsymbol{\beta})}{\sum_{x\in\mathcal{X}} \psi(y_i=1|x_i;\boldsymbol{\beta})} \qquad (2)$$

La función de verosimilitud de $\boldsymbol{\beta}$ dada una observación x_i dentro del conjunto de datos de *solo presencias* se basa en $\pi(x_i|y_i; \boldsymbol{\beta})$ considerada como una función de los parámetros $\boldsymbol{\beta}$. Por lo tanto, para una muestra de datos de *solo presencias* $x_1, \dots, x_n$ la función de verosimilitud es

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\psi(y_i=1|x_i;\boldsymbol{\beta})}{\sum_{x\in\mathcal{X}} \psi(y_i=1|x_i;\boldsymbol{\beta})} \qquad (3)$$

El denominador de la ecuación (3) corresponde a la probabilidad marginal de ocurrencia a través del área de estudio, que se calcula sumando sobre todos los elementos de $x \in \mathcal{X}$ donde $\mathcal{X}$ (o sobre una muestra aleatoria de $\mathcal{X}$ llamada *background* y denotada como $\mathfrak{B}$) corresponde al espacio de valores de x.

Enfoque bayesiano (*MaxBayes*)

Puede construirse un enfoque bayesiano del modelo de Royle *et al*. (2012) asignando una distribución *a priori* para los parámetros $\boldsymbol{\beta}$ del modelo. Dicha distribución puede ser informativa, si se dispone de conocimiento de expertos, o bien, no informativa. Por ejemplo, en muchas ocasiones se dispone de escasos registros de presencias, lo que dificulta la estimación de la prevalencia en el modelo *Maxlike*, sin embargo, en la mayoría de los casos el investigador tiene una idea acerca de en qué rango de valores se encuentra dicha prevalencia, lo cual puede reflejarse en una distribución *a priori* para β_0 que represente dicho conocimiento.

Reescribiendo la ecuación (3) en términos del espacio medioambiental $\mathbf{x}(\mathbf{z})$ y suponiendo que $\boldsymbol{\beta}$ se distribuye *a priori* como una normal multivariada , es decir $\boldsymbol{\beta} \sim NM(\boldsymbol{\beta_0}, \mathbf{V_0})$, donde $\boldsymbol{\beta_0}$ y $\mathbf{V_0}$ corresponden a la media y la covarianza *a priori*, respectivamente; esto es, dichos hiperparámetros cuantifican el estado de conocimiento sobre los parámetros $\boldsymbol{\beta}$ antes de observar los datos. Aplicando el teorema de Bayes e ignorando los términos que no involucran a $\boldsymbol{\beta}$, la distribución *a posteriori* queda expresada proporcionalmente como

$$p(\boldsymbol{\beta}|y_i = 1, \mathbf{x}(\mathbf{z})) \propto \prod_{i=1}^{n} \frac{\psi(y_i=1|\mathbf{x}(\mathbf{z_i});\boldsymbol{\beta})}{\sum_{x\in\mathcal{X}} \psi(y_i=1|\mathbf{x}(\mathbf{z_j});\boldsymbol{\beta})} \times exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta_0})'\mathbf{V_0}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta_0})\right\} \quad (4)$$

donde una forma natural de modelar $\psi(y_i = 1|\mathbf{x}(\mathbf{z});\boldsymbol{\beta})$ es mediante la función liga *logit*, expresándola entonces como

$$\begin{aligned} logit\,(\psi(y_i = 1|\mathbf{x}(\mathbf{z});\boldsymbol{\beta}) &= \log\frac{\psi(y_i = 1|\mathbf{x}(\mathbf{z});\boldsymbol{\beta})}{1-\psi(y_i = 1|\mathbf{x}(\mathbf{z});\boldsymbol{\beta})} \\ &= \mathbf{x}(\mathbf{z})'\boldsymbol{\beta}. \end{aligned}$$

Proceso Poisson no Homogéneo (*IPP*)

Otro modelo estadístico que aborda el problema de hacer inferencia con datos de solo presencias es un proceso Poisson no homogéneo propuesto en el contexto de los MDEs por Warton y Shepherd (2010) y ampliado por Fithian y Hastie (2013). En el contexto del modelo *IPP*, en lugar de modelar la probabilidad de ocurrencia, se modela la *intensidad* de ocurrencia; esto es, la cantidad correspondiente al número esperado de especímenes por unidad de área.

Sea D el área geográfica de interés, típicamente un conjunto en $\mathbb{R}^2$. Asociado a cada ubicación geográfica $z \in D$ se tiene un vector $x(z)$ de variables medioambientales medidas o estimadas. El conjunto de datos de solo presencias consiste de n_1 ubicaciones de avistamientos $z_i \in D$ para $i = 1,2,\dots,n_1$, además de n_0 observaciones del *background* z_i para $i = n_1+1,\dots,n_1+n_0$ (generalmente un grid regular o muestra aleatoria uniforme de D).

El modelo *IPP* es un modelo simple para un conjunto de puntos aleatorios Z que caen dentro algún dominio D y puede definirse por su función de intensidad como $\lambda: D \to [0,\infty)$ que representa la verosimilitud de que un punto caiga dentro o cerca de z. Para cualquier subconjunto $A \subseteq D$, al integrar $\lambda(z)dz$ se obtiene el número de registros de presencias en A y se expresa como $\Lambda(A) = \int_A \lambda(z)dz$ donde la única restricción es que la integral sea finita, $\Lambda(D) < \infty$.

Existen dos formas de expresar a un modelo *IPP* con intensidad λ. El primer enfoque consiste en que del número de puntos es una variable aleatoria Poisson con media $\Lambda(D)$ y condicionada sobre el número de puntos, sus ubicaciones son independientes e idénticamente distribuidas (*i.i.d*) con densidad $p_\lambda(z) = \lambda(z)/\Lambda(D)$. El otro enfoque consiste en pensar al modelo *IPP* como un límite continuo de un modelo de conteo Poisson independiente para una discretización muy fina de D. Si $N(A) = \#(Z \cap A)$ es el número de puntos que caen en el conjunto A, entonces $N(A) \sim Poisson(\Lambda(A))$. En el caso de un dominio finito y discreto $D = \{z_1, z_2, \dots, z_m\}$, el modelo IPP se reduce a un modelo Poisson discreto con $N(z_i) \sim Poisson(\lambda(z_i))$. En este sentido el modelo *IPP* puede verse como una discretización muy fina (es decir, de celdas muy pequeñas) de D (Fithian y Hastie, 2013).

Función de verosimilitud del modelo *IPP*

Sea $\lambda(z_i)$ la intensidad en el punto z_i, que limita el número esperado de presencias por unidad de área en función de k variables explicativas, ésta puede modelarse como una función log-lineal, $\lambda(z_i) = e^{\alpha+\beta' x(z)}$, donde $x(z)$ representa a las covariables medioambientales. La log-verosimilitud en términos de la muestra de presencias en el modelo *IPP* se expresa como

$$l(\alpha,\beta,\mathbf{y}) = \sum_{i:y_i=1}(\alpha+\beta' x(z_i)) - \int_D e^{\alpha+\beta' x(z)}dz - \log n_1!. \quad (5)$$

Note que en la ecuación ((5)), para cualquier $\hat{\beta}$, $\hat{\alpha}$ juega el rol de constante de normalización que garantiza que $\lambda(z)$ integre a n_1, esto es, el total de registros de presencias, y que por tanto, si n_1 no es de interés para el investigador tampoco lo será $\hat{\alpha}$. Según Fithian y Hastie (2013) al diferenciar (5) con respecto a α se obtiene que $n_1 = \int_D e^{\alpha+\beta' x(z)}dz = \Lambda(D)$, el número de registros utilizados para ajustar el modelo. Según Fithian y Hastie (2013), cuando no sea posible evaluar analíticamente la integral en (5), ésta puede evaluarse numéricamente con base en el *background*. Por tanto, una aproximación de (5) es:

$$l(\alpha,\beta,\mathbf{y}) \approx \sum_{i:y_i=1}(\alpha+\beta' x(z_i)) - \frac{|D|}{n_0}\sum_{i:y_i=0} e^{\alpha+\beta' x(z_i)} - \log n_1! \quad (6)$$

donde $|D| = \int_D 1dz$ representa el área total de la región de estudio. Los puntos del *background* pueden ser tanto una muestra uniforme de D o bien un grid regular.

Enfoque bayesiano (*IPPBayes*)

De manera análoga que en el modelo *MaxBayes*, puede construirse un enfoque bayesiano del modelo *IPP* para datos de solo presencias. El enfoque adoptado hace que D (área de interés) sea un espacio discreto compuesto por una rejilla de celdas muy pequeñas del espacio geográfico de interés. Resulta natural asignar a $\boldsymbol{\beta}$ una distribución *a priori* normal multivariada, es decir, $\boldsymbol{\beta} \sim NM(\boldsymbol{\beta_0}, \mathbf{V_0})$, donde al igual que en el modelo *MaxBayes*, $\boldsymbol{\beta_0}$ y $\mathbf{V_0}$

corresponden a la media y la covarianza *a priori*, respectivamente. Recuerde que α es el parámetro asociado al número de presencias en el área de estudio (abundancia) y que en muchas ocasiones el experto tiene una idea de en qué rango de valores se encuentra dicha cantidad. Este conocimiento puede reflejarse asignado una distribución *a priori* informativa para α.

Tomando como función de verosimilitud el antilogaritmo de (6) tal que $\boldsymbol{\beta} = (\alpha, \beta)'$ se tiene que

$$L(\boldsymbol{\beta}|\mathbf{y}) \approx \frac{1}{n_1!}\, exp\left(-\frac{|D|}{n_0}\sum_{i:y_i=0} e^{\mathbf{x}(\mathbf{z_i})'\boldsymbol{\beta}}\right)\prod_{i:y_i=1} e^{\mathbf{x}(\mathbf{z_i})'\boldsymbol{\beta}}. \quad (7)$$

Aplicando la regla de Bayes e ignorando aquellos términos que no involucren a $\boldsymbol{\beta}$ se tiene que la distribución *a posteriori* de $\boldsymbol{\beta}$ es proporcional a

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto exp\left(-\frac{|D|}{n_0}\sum_{i:y_i=0} e^{\mathbf{x}(\mathbf{z_i})'\boldsymbol{\beta}}\right)\prod_{i:y_i=1} e^{\mathbf{x}(\mathbf{z_i})'\boldsymbol{\beta}} \times exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta_0})'\mathbf{V_0}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta_0})\right\} \quad (8)$$

Note que el cálculo de momentos, marginales y otras cantidades de interés a partir de (8) no puede realizarse analíticamente, por lo que es factible la implementación de algún algoritmo de simulación de cadenas de Markov Monte Carlo (MCMC, por sus siglas en Inglés), por ejemplo, el algoritmo de Metrópolis-Hastings cuya teoría pude consultarse en Chib y Greenberg (1995).

Estudio de caso

Para ilustrar de forma práctica los métodos estadísticos descritos en el presente trabajo se implementaron ambos modelos en un conjunto de datos simulados, y en un conjunto de datos reales. En el segundo caso, los resultados de *MaxBayes* e *IPPBayes* se compararon con los de *Maxent*.

Simulación de datos

Para la simulación se contempló un zona de interés compuesta por 10,000 celdas. Se consideraron dos covariables $x(z_1)$ y $x(z_2)$ para las cuales se simularon 10,000 valores provenientes de una distribución normal estándar ($x(z_1)\sim N(0,1)$ y $x(z_2)\sim N(0,1)$). La probabilidad de presencia $\psi(x)$ se calculó a partir del $\log\left(\frac{\psi(x)}{1-\psi(x)}\right) = -2 + 2 * x(z_1) - 2 * x(z_2)$. Los registros de presencia-ausencia ($y \in \{0,1\}$) para cada celda se calcularon como un ensayo *Bernoulli* con probabilidad ψ ($y_i \sim Ber(1,\psi)$). La proporción de sitios ocupados o prevalencia en este ejemplo de simulación fue de 0.38. Para implementar el modelo *MaxBayes* se tomaron muestras aleatorias de tamaño 2,000 , 1,000 y 100 del subconjunto de celdas ocupadas con la finalidad de comparar las estimaciones de los parámetros y la prevalencia al variar los registros de presencias. El *background* se definió como el conjunto de datos correspondientes a las covariables en las 10,000 celdas.

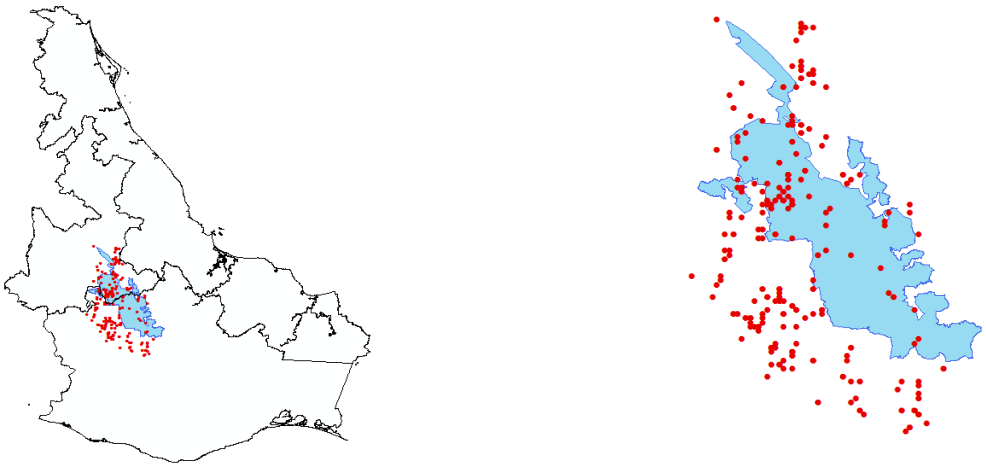
Para implementar el modelo *MaxBayes* se partió de la ecuación (4), asignando para $\boldsymbol{\beta}$ una distribución normal con media cero y varianza grande, para reflejar el desconocimiento *a priori*. Ésto es, $\boldsymbol{\beta}\sim NM(\mathbf{0}, \mathbf{V_0})$, donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ son los parámetros asociados al intercepto y a las covariables $x(z_1)$ y $x(z_2)$ simuladas, y

$$\mathbf{V_0} = \begin{bmatrix} 10^5 & 0 & 0 \\ 0 & 10^5 & 0 \\ 0 & 0 & 10^5 \end{bmatrix}. \quad (8)$$

Para implementar el modelo *IPPBayes*, se asignó la misma distribución *a priori* para $\boldsymbol{\beta}$ que en modelo *MaxBayes*, utilizando $n_1 = 2000$, $n_1 = 1000$ y $n_1 = 100$ registros de presencias para mostrar como afecta el número de presencias en el parámetro α y como consecuencia las intensidades de ocurrencia estimadas.

Datos de género *Dalea*

Como ejemplo con datos reales se eligió el género *Dalea* cuyos datos constan de 301 presencias (Ver figura 1b) provenientes de la Reserva de la Biosfera Tehuacán-Cuicatlán, disponibles en el portal de la CONABIO como parte del proyecto Q014 (http://www.conabio.gob.mx/remib/doctos/remibnodosdb.html). El género *Dalea* se considera endémico de la zona que comprende a la Provincia Florística del Valle de Tehuacán-Cuicatlán (Méndez *et al*., 2004), cuyos límites abarcan los estados de Oaxaca y Puebla. La zona total de estudio (*landscape*) se consideró a los estados de Veracruz, Oaxaca y Puebla. La información de las covariables medioambientales fue descargado del portal http://www.worldclim.org/bioclim que corresponde a la base de datos global BIOCLIM. Las covariables medioambientales utilizadas fueron la precipitación anual (Pp), la altitud sobre el nivel del mar (Alt), la temperatura media anual (Tmedia), el rango de la temperatura anual (RangoT), además de la latitud (Lat) y la longitud (Lon).



(a) Zona de estudio. (b) Presencias *Dalea*.

(c)

Figura 1: Zona de estudio y Reserva Cuicatlán-Tehuacán.

Para la implementación de cada uno de los modelos (*Maxent*, *MaxBayes* e *IPPBayes*), se dividió el área total de estudio en una rejilla regular de $(30\ arcseg \times 30\ arcseg) \approx (1\ km \times 1\ km)$ por celda. Se extrajeron los valores de cada una de las covariables medioambientales asociadas al centroide de cada celda, los cuales se utilizaron para formar el *background* o conjunto de datos en toda el área de estudio. En ninguno de los modelos implementados en este apartado se abordó el problema del posible sesgo en los registros de presencia. Las covariables medioambientales fueron estandarizadas siguiendo la recomendación de Royle *et al*. (2012). En el caso de *Maxent* se utilizó el valor por defaúlt para la prevalencia $\tau = 0.5$ recomendado por Elith et al. (2011).

Para ajustar el modelo *MaxBayes* se partió de la ecuación (4), asignando para $\boldsymbol{\beta}$ una distribución normal con media cero y varianza grande, para reflejar el desconocimiento *a priori*. Esto es, $\boldsymbol{\beta} \sim NM(\mathbf{0}, \mathbf{V_0})$, donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)'$ son los parámetros asociados al intercepto y a las covariables medioambientales, y

$$\mathbf{V_0} = \begin{bmatrix} 10^5 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 10^5 \end{bmatrix}.$$

En el caso del modelo *IPPBayes*, se partió de la ecuación (8). Se especificó la misma distribución *a priori* para $\boldsymbol{\beta}$ utilizada en *MaxBayes*. El *intensidad de ocurrencia* se modelo como $\lambda(z_i) = exp\,(\mathbf{x}(\mathbf{z_i})'\boldsymbol{\beta})$ donde $\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_6)$ es el vector de coeficientes que incluye al intercepto α.

Simulación de la distribución *a posteriori* mediante MCMC

Las distribuciones *a posteriori* derivadas de (4) y (8) no pueden integrarse en forma analítica por lo que se utilizó el algoritmo Metrópolis-Hastings mediante el paquete *MHadaptive* (Chivers, 2012) de R (R Core Team, 2013). En ambos ejemplos de aplicación, tanto en *MaxBayes* como en *IPPBayes* se simularon 50,000 valores de las distribución *a posteriori*, tomando como *burnIn* a los primeras 25,000 iteraciones. Posteriormente se calcularon los estimadores bayesianos bajo pérdida $0-1$ para cada componente de $\boldsymbol{\beta}$, esto es, la moda de los valores simulados después del periodo de calentamiento.

El procedimiento anterior se repitió en ambos ejemplos en tres ocasiones, proporcionando diferentes valores de inicio para cada cadena, lo anterior con el fin de medir la convergencia de las cadenas a la distribución estacionaria. Dicha convergencia se midió mediante la prueba de Gelman y Rubin (1992) implementada en el paquete *coda* (Plummer *et al*., 2006) de R.

RESULTADOS Y DISCUSIÓN

Datos de simulación

En el Tabla (1) se presenta la información correspondiente al modelo *MaxBayes*. El número de registros n, para ajustar el modelo en este ejemplo de simulación fue de 2,000, 1,000 y 100. La tabla contiene las estimaciones de los parámetros $\boldsymbol{\beta}$ con sus respectivos intervalos de máxima probabilidad *a posteriori* (HPD). Note que la estimación de β_0

asociado al intercepto en el modelo *MaxBayes* es muy similar aún variando n, y cercana al valor real $\beta_0 = -1$, sin embargo, la incertidumbre asociada a la estimación del parámetro crece, lo que se refleja en intervalos HPD con mayor longitud. Un comportamiento similar lo presentan las estimaciones para β_1 y β_2 , donde las estimaciones correspondientes se acercan al valor real ($\beta_1 = 2$ y $\beta_2 = -2$), y los intervalos HPD tienen mayor amplitud a medida que n disminuye. Recuerde que este ejemplo, se han utilizado distribuciones *a priori* no informativas, por lo que la inferencia del modelo bayesiano se basa en los datos y que por tanto, hereda las propiedades asintóticas de los estimadores de máxima verosimilitud (EMV) y a medida que n crece, la incertidumbre asociada a los parámetros disminuye.

La prevalencia estimada por *MaxBayes* se ilustra en la figura (2). Note que a medida que el n incrementa, el valor estimado de la prevalencia se acerca más la prevalencia real (línea azúl en 2). Recuerde que a través de β_0 *MaxBayes* estima la prevalencia, por lo que en el caso de muestras pequeñas, cualquier información de expertos en relación a la especie de interés que ayude a identificar la prevalencia, debe hacerse a través de una distribución *a priori* informativa sobre todo para el parámetro β_0.

Por otra parte, en la Tabla (2) se presenta la información del modelo *IPPBayes*. Note que la estimación del parámetro α es quien tiene mayor variación a medida que n_1 lo hace, mientras que las estimaciones para β_1 y β_2 son menores. Al disminuir el número de los registros, los intervalos HPD tienen mayor amplitud en virtud de que se han utilizado distribuciones *a priori* no informativas para α, β_1 y β_2 y la inferencia bajo las distribuciones *a posteriori* se basan casi por completo en la verosimilitud. Como ya se mencionó, en el modelo *IPPBayes*, α garantiza que $n_1 \approx \sum_D e^{\mathbf{x(z_i)}'\hat{\boldsymbol{\beta}}}$, es decir, el número de registros utilizados para ajustar el modelo. Note que a medida que n_1 disminuye, la estimación de α es menor, ajustando en cada caso las intensidades de ocurrencia hacía abajo. En este sentido el modelo *IPPBayes* bajo distribuciones *a priori* no informativas proporciona intensidades de ocurrencia relativas al tamaño de registros utilizados para ajustar el modelo.

Tabla 1: Resumen de *MaxBayes* para distintos n (simulación).

Variable	Parámetro	Estimación	HPD Inf	HPD Sup
		$n = 2000$		
Intercepto	β_0	-0.88	-1.06	-0.70
$x(z_1)$	β_1	1.87	1.73	2.16
$x(z_2)$	β_2	-1.86	-2.19	-1.75
		$n = 1000$		
Intercepto	β_0	-0.96	-1.13	-0.62
$x(z_1)$	β_1	1.84	1.66	2.26
$x(z_2)$	β_2	-1.79	-2.29	-1.68
		$n = 100$		
Intercepto	β_0	-1.05	-1.61	0.52
$x(z_1)$	β_1	1.97	1.35	4.60
$x(z_2)$	β_2	-1.96	-3.69	-1.02

.

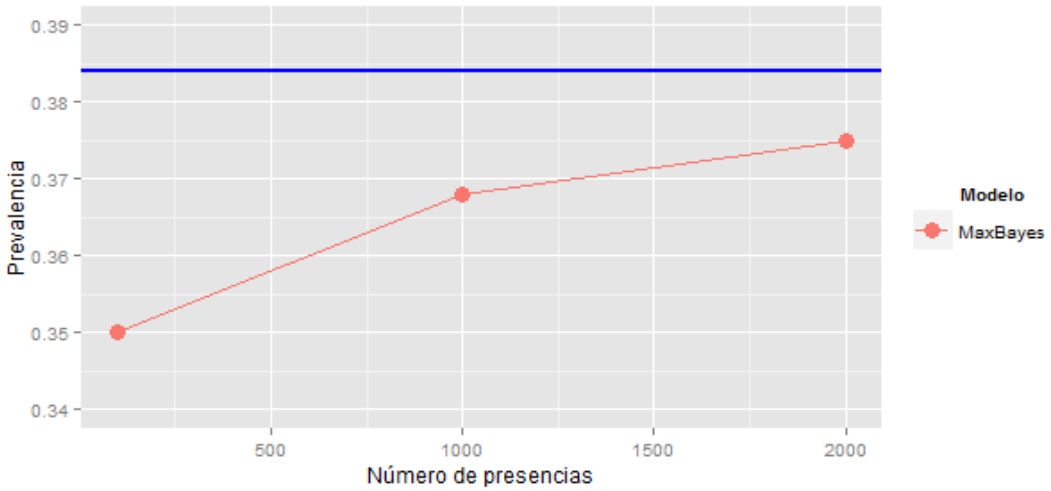


Figura 2: Prevalencia estimada por *MaxBayes* en función del tamaño de presencias.

En muchos estudios en relación a los MDEs, el investigador lleva años investigando una determinada especie de interés, por lo que posee información en relación a ésta, como por ejemplo, un estimado de la abundancia. Dicha información puede emplearse por medio de distribuciones *a priori* informativas para $\boldsymbol{\beta}$ en el modelo *IPPBayes*. Esto representa una ventaja del modelo *IPPBayes* con respecto a su contraparte frecuentista.

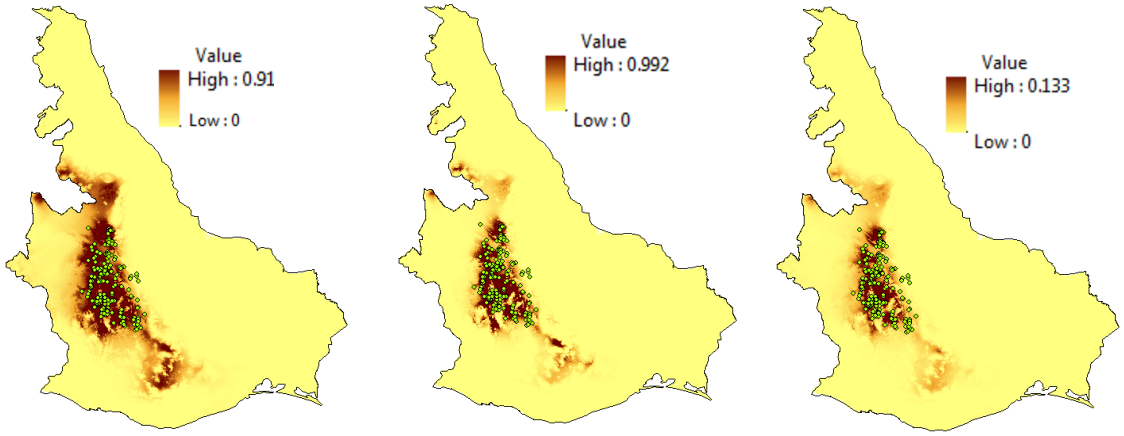
La prueba de convergencia de de Gelman y Rubin (1992) que se aplicó a las muestras simuladas de la distribución *a posteriori* en ambos casos de estudio, tanto para *MaxBayes* como el modelo *IPPBayes*, indicó que dichas cadenas convergieron a la distribución estacionaria.

Género *Dalea*

En las figuras (3a-3b) se observan los mapas de probabilidad de presencia obtenidas con *MaxBayes* y la salida logística de *Maxent*, respectivamente. En la Tabla (3) se resumen las estimaciones de $\boldsymbol{\beta}$ tanto para el modelo *MaxBayes* como el modelo *Maxent*. Note que los signos asociados a cada estimador son iguales en ambos modelos, lo que nos da cuenta en qué sentido las covariables afectan la presencia del género estudiado. También se resumen los intervalos de máxima probabilidad *a posteriori* (HPD) asociadas a la estimación de cada parámetro en el modelo *MaxBayes*.

Tabla 2: Resumen de *IPPBayes* para distintos n_1 (simulación).

Variable	Parámetro	Estimación	HPD Inf	HPD Sup
		$n_1 = 2000$		
Intercepto	α	-1.94	-2.00	-1.89
$x(z_1)$	β_1	0.58	0.54	0.63
$x(z_2)$	β_2	-0.57	-0.61	-0.53
		$n_1 = 1000$		
Intercepto	α	-2.65	-2.72	-2.56
$x(z_1)$	β_1	0.60	0.52	0.65
$x(z_2)$	β_2	-0.55	-0.62	-0.49
		$n_1 = 100$		
Intercepto	α	-4.97	-5.24	-4.72
$x(z_1)$	β_1	0.72	0.48	0.88
$x(z_2)$	β_2	-0.48	-0.69	-0.30



Salida logística de *Maxent*. Probabilidad de ocurrencia *MaxBayes*. Intensidad de ocurrencia IPPBayes.

Figura 3: Distribución potencial del género *Dalea* obtenidos mediante los modelos *Maxent*, *MaxBayes* e *IPPBayes*.

Tabla 3: Resumen de *MaxBayes* y *Maxent* para distintos tamaños de muestra.

n = 301		MaxBayes			Maxent
Variable	**Parámetro**	**Estimación**	**HPD Inf**	**HPD Sup**	**Estimación**
Intercepto	β_0	-10.39	-12.26	-9.49	–
Altitud	β_1	-11.90	-13.17	-8.38	-19.45
Tmedia	β_2	-11.05	-12.22	-8.21	-27.38
Pp	β_3	-5.35	-6.57	-4.46	-24.95
RangoT	β_4	4.34	3.08	4.92	10.08
Lon	β_5	1.33	-0.27	2.48	0.74
Lat	β_6	-1.40	-2.26	-0.71	-5.49

Tal como se aprecia en la Figura (3), tanto *Maxent* como *MaxBayes* proporcionan el mismo patrón de predicción, aunque como señala Royle *et al*. (2012), *Maxent* tiende a subestimar la presencia de la especie en aquellas áreas donde se le ha observado, mientras que sobreestima para zonas donde no se encuentran registros de la misma. El hecho de que *Maxent* aparentemente subestime la probabilidad de presencia (a través de la salida logística) en aquellas zonas con registros y sobrestime aquellas donde no existen, se debe básicamente a dos razones; la primera obedece al hecho de que *Maxent* asume que aquellos sitios con condiciones típicas para la especie tienen probabilidad de 0.5 (a través de τ), y la segunda razón obedece al hecho de que se utilizan diferentes funciones de enlace para ψ, en el caso de *MaxBayes* se utiliza el modelo logístico $\psi(y_i = 1|x(z_i), \beta_0, \boldsymbol{\beta}) = e^{\beta_0 + \boldsymbol{\beta} x(z_i)}/(1 + e^{\beta_0 + \boldsymbol{\beta} x(z_i)})$, mientras que *Maxent* utiliza un modelo log-lineal de la forma $\psi(y_i = 1|x(z_i), \boldsymbol{\beta}) = e^{\boldsymbol{\beta} x(z_i)}$. Como acertadamente señala Merow y Silander (2014), la principal diferencia entre estas dos funciones es el intercepto β_0 que incluye *MaxBayes*, el cual define la prevalencia esperada en el área de estudio.

Por otra parte, en la figura 3 se ilustra el mapa de intensidades de ocurrencia del género *Dalea* mientras que en la Tabla (4) se resume las estimaciones de los parámetros del modelo *IPPBayes*, también se incluyen los intervalos de máxima probabilidad *a posteriori* respectivos.

Tabla 4: Resumen del modelo *IPPBayes*

Variable	***Parámetro***	***Estimación***	***HPD inf***	***HPD sup***
Intercepto	α	-12.99	-13.97	-11.43
Altitud	β_1	-3.95	-5.31	-2.97
Tmedia	β_2	-4.31	-5.48	-3.73
Pp	β_3	-4.73	-5.26	-3.55
RangoT	β_4	1.92	1.47	2.42
Lon	β_5	-0.26	-1.17	0.84
Lat	β_6	-1.37	-2.02	-0.86

Como ya se señaló anteriormente, el parámetro α únicamente garantiza que al integrar $\lambda(z)$ sobre todo D, nos de como resultado n_1, es decir, integre al total de registros de presencias que en nuestro caso es $n_1 = 301$ (para nuestro caso de estudio, dado que hemos discretizado D , implica que $n_1 \approx \sum_D e^{\mathbf{x}(\mathbf{z_i})'\hat{\boldsymbol{\beta}}}$).

Recuerde que en el modelo *IPPBayes* lo que se modela es la intensidad de ocurrencia de la especie por unidad de área. Note que el concepto de intensidad de ocurrencia está íntimamente ligada al concepto de probabilidad de presencia, dado que en aquellos sitios donde la intensidad de ocurrencia es mayor corresponde a las máximas probabilidades asignadas por *MaxBayes*, y como se observan en la figura 3, el modelo *IPPBayes* proporciona la misma distribución potencial del género *Dalea* que el modelo *MaxBayes*. Es importante destacar que las intensidades de ocurrencia calculadas el modelo *IPPBayes* son intensidades de ocurrencia *relativas* ya que los registros utilizados en el modelo conforman una fracción de la población de la especie de interés, sin embargo, constituye una forma alternativa para modelar la distribución potencial de la especie.

Recomendaciones

En ambos modelos propuestos en este trabajo, es factible incluir un término que represente la dependencia espacial entre celdas vecinas del grid, concretamente en la función de enlace de cada modelo. En el caso del modelo *MaxBayes*, la presencia/ausencia de la especie puede asociarse con la presencia/ausencia en ubicaciones vecinas por lo que la función de enlace quedaría expresada como $logit\ (\psi(y_i = 1|\mathbf{x}(\mathbf{z}); \boldsymbol{\beta}) = \mathbf{x}(\mathbf{z})'\boldsymbol{\beta} + \rho_i$. De forma similar, en el modelo *IPPBay* la intensidad de ocurrencia se afectaría por ρ_i a través de $\lambda(z_i) = exp\ \{\alpha + \beta' x(z) + \rho_i\}$. El término ρ_i puede modelarse mediante un modelo normal condicional autoregresivo (CAR).

CONCLUSIONES

Los resultados indican que ambos modelos aquí propuestos, *MaxBayes* e *IPPBayes*, constituyen alternativas viables cuando se modelan distribuciones de especies con registros de *solo presencias*. Ambos modelos permiten incorporar conocimiento *a priori* en relación a las especies de interés que pueden resultan en predicciones más acordes a la naturaleza estudiada, sobre todo cuando el investigador cuenta con escasos registros de presencia, como suele ser en la mayoría de los casos. Lo anterior es una mejora sustancial con respecto a los modelos *Maxlike* e *IPP*.

En el caso del ejemplo con datos simulados y distribuciones *a priori* no informativas para los parámetros de *MaxBayes*, éste es un modelo que aproxima muy bien la prevalencia aún cuando el número de presencias es pequeño. Dicha estimación puede ser mejor cuando se utilizen distribuciones *a priori* informativas. Para los datos del género *Dalea*, tanto *MaxBayes* como el modelo *IPPBayes* predicen patrones de distribución potencial similares al obtenido con el software *Maxent*, aunque dicha similitud es más acentuada en el caso de *MaxBayes* e *IPPBayes*. La ventaja de *MaxBayes* sobre *Maxent*, es que el primero estima la prevalencia y por tanto también estima la probabilidad de ocurrencia, *Maxent*

por el contrario solo proporciona un índice que indica que tan idóneo es el sitio para albergar a la especie con respecto a otros y generalmente ese índice sobrestima la presencia de la especie en sitios donde no existen registros, mientras que subestima para aquellas zonas donde la especie ha sido registrada. Por otro lado *IPPBayes*, estima la intensidad de ocurrencia, es decir, el número esperado de especímenes por unidad de área, y cuando se utilizan distribuciones *a priori* no informativas para los parámetros, dicha intensidad es relativa al tamaño de presencias utilizadas para ajustar el modelo.

REFERENCIAS

Chib, S. y Greenberg, E. 1995. Understanding the Metropolis-Hasting Algorithm. The American Statistician, 49(4), 227–335.

Chivers, C. 2012. MHadaptive: General Markov Chain Monte Carlo for Bayesian Inference using adaptive Metropolis-Hastings sampling. R package version 1.1-8.

Elith, J., Phillips, S. J., Hastie, T., Dukiv, M., Chee, Y. E. y Yates, C. J. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions, 17, 43–57.

Fithian, W. y Hastie, T. 2013. Finite-sample equivalence in statistical models for presence-only data. Annals of Applied Statistics, 7, 1917–1939.

Gelman, A. y Rubin, D. 1992. Inference from iterative simulation using multiple sequences. Statistical Sciencie, 7, 457–511.

Graham, C. H., Moritz, C. y Williams, S. E. 2006. Habitat history improves prediction of biodiversity in rainforest fauna. The National Academy of Sciences of the USA, 1, 632–636.

Latimer, A. M., Wu, S., Gelfan, A. E. y Silander, J. A. J. 2006. Building statistical models to analyze species distributions. Ecological Applications, 16(1), 33–50.

Méndez, L., Ortiz, E. y Villaseñor, J. (2004). Las Magnoliophyta endémicas de la porción xerofítica de la provincia florística del Valle de Tehuacán-Cuicatlán, México. Anales del Instituto de Biología. UNAM. Serie Botánica, 751, 87–104.

Merow, C. y Silander, J. A. 2014. A comparison of Maxlike and Maxent for modelling species distributions. Methods in Ecology and Evolution, 5, 215–225.

Phillips, S., Dudik, M. y Schapire, R. 2004. A Maximum Entropy Approach to Species Distribution Modeling. Preceedings of the Twenty-Firts International Conference on Machine Learning, 1–8.

Plummer, M., Best, N., Cowles, K. y Vines, K. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News, 6, 1, 7–11.

R Core Team 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Royle, J., Chandle, R. B., Yackulic, C. y Nichols, J. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling speciesdistributions. Methods in Ecology and Evolution, 3, 545–554.

Thomas, C., Cameron, A., Green, R., Bakkenes, M., Beaumont, L.J., Collignham Y.C., Erasmus, B.F.N., de Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.a., Townsend Peterson, A., Phillips, O.L. y Williams, S. 2004. Extinction risk from climate change. Nature 427, 145–148.

Thuiller, W., Richardson, D.M., Pysek, P., Midgley, G.F., Hughes, G.O., Rouget, R. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasion at a global scale. Global Change Biology, 11, 2234–2250.

Warton, D. y Shepherd, L. 2010. Poisson Point Process Models solve the "Pseudo-absence problem for presence-only data in ecology". The Annals of Applied Statistics, 4, 1383–1402.

Submitted April 28, 2013 – Accepted August 10, 2015