Universidad Autónoma de Yucatán

Facultad de Matemáticas

Enfoque Bayesiano del Método Domain para la estimación de la zona de alto potencial de una especie.

Tesis que presenta

L.M. Edith Aracelly Pech Méndez

para obtener el Grado de

Maestro en Ciencias Matemáticas

Mérida, Yucatán **Octubre 2007**

Dedicatoria

A mi madre

Porque luchaste incansablemente sin importar las dificultades enfrentadas día a día para hacerme una mujer preparada. Porque no hay forma de agradecerte todo lo que me das.

Agradecimientos

A mi madre, Gladys Yolanda Méndez Cruz, por su apoyo incondicional día a día, por su comprensión y paciencia durante mis años de estudio.

Al señor Ricardo Velásquez Escalante por sus consejos y apoyo durante mis estudios.

A mis hermanos Oscar Abraham Pech Méndez, Juan Carlos Pech Méndez y a mi hermana Karina Esther Velásquez Méndez por su compañía y apoyo.

A mi asesor de tesis, el Dr. Jorge Armando Argáez Sosa por sus enseñanzas y motivación para seguir creciendo profesionalmente, por su disposición y ayuda incondicional en aclarar mis dudas durante la investigación y por sus observaciones en la elaboración y redacción de este trabajo.

A la M.C. Luci del Carmen Torres Sánchez por sus valiosas observaciones revisando este trabajo y por sus enseñanzas en el salón de clases.

A la Dra. Celene Marisol Espadas Manrique por su valiosa colaboración en las aplicaciones de este trabajo y por sus sugerencias para mejorar la redacción del mismo.

A mis maestros, quienes siempre me han enseñado algo. A todos, mi mayor reconocimiento y gratitud.

A mis amigos Fadelli Pérez Dzul, Abigail Barroso Quiab, Heidy Cecilia Escamilla Puc, Ángel Uh Zapata, Oscar Muñoz Carballo, Julio Ché Uicab, Reymundo Itzá Balám y Neyfis Solís Baas por su amistad, apoyo y por todos los momentos compartidos durante mis estudios.

A la Facultad de Matemáticas por el soporte institucional proporcionado durante mis estudios de maestría.

A CONACYT por haberme otorgado una beca por 11 meses durante mis estudios de maestría.

Y a todas aquellas personas que de una u otra forma colaboraron en la realización de este trabajo.

Índice general

Capítulo 1

Introducción

Una de las tareas importantes de la humanidad es la conservación de la biodiversidad. Para lograr esto es necesario frenar la destrucción del medio ambiente, proteger a las especies de flora y fauna amenazadas y procurar el uso adecuado de los recursos naturales. Una manera de evitar la destrucción del ambiente es contar con zonas protegidas. De este modo las zonas no serán dañadas o destruidas por actividades humanas y permitirán la conservación de las especies presentes en ellas. Es por esto que ha surgido la necesidad de proponer métodos que permitan estimar las zonas que poseen condiciones favorables para el establecimiento de especies, zonas que en este trabajo serán denominadas *zonas de alto potencial*.

En la actualidad existen varios métodos diseñados para determinar las zonas de alto potencial de una especie que usan solamente registros de presencia de la misma, entre los cuales podemos citar: Bioclim (Busby, 1991), Domain (Carpenter, Gillison y Winter, 1993), Floramap (Jones y Gladkov, 1999), GARP (Genetic Algorithm for Rule-set Prediction, Stockwell y Noble, 1991; Stockwell y Peters, 1999), Modelo Mezcla (Argáez, Christen, Nakamura y Soberón, 2005) y Maxent (Phillips, Anderson y Schapire, 2006), los cuales se describirán en la siguiente sección. Los resultados obtenidos utilizando estos métodos permiten la ubicación de zonas donde la presencia de la especie es potencialmente posible. Por otro lado, el conocer las zonas de alto potencial de presencia de una especie facilita los planes de restauración, asi como los inventarios biológicos.

En este trabajo se utilizará para estimar las zonas de alto potencial el método Domain, (Carpenter, Gillison y Winter, 1993) ya que es uno de los métodos más simples de entender y fácil de programar. Además este método ha proporcionado resultados buenos en diversas aplicaciones de ecología (Pallaris, 1998; Espadas, Durán y Argáez, 2003; Montenegro, 2006).

El método Domain asigna un valor de similitud a cada uno de los sitios objetivo, que en general son los nodos de una retícula que cubre la zona de estudio de la especie, basándose en la proximidad de las variables climáticas de cada sitio con las variables climáticas de los sitios de presencia. El valor que se asigna a los sitios objetivos se define utilizando la métrica de Gower (Gower, 1971). Tal valor y la

determinación de un umbral, elegido según el conocimiento del usuario, permitirá definir las zonas de alto potencial de la especie bajo consideración. No obstante, una desventaja del método se encuentra precisamente en la elección del umbral, ya que su determinación la define el experto sin recurrir a métodos o algoritmos númericos formales.

Una metodología que justifica estadísticamente la elección del umbral en el método Domain fue propuesta por Argáez (1996), la cual usa variables climáticas de sitios de presencia registrados y se basa en un método de remuestreo denominado Jackknife (Efron, 1982). Sin embargo esta metodología no considera la información que posee el experto respecto a la especie, la cual se tiene en muchos casos y no se utiliza. En la práctica los expertos utilizan frecuentemente su conocimiento acerca de las regiones de establecimiento y no establecimiento de la especie, para corregir a mano las zonas de alto potencial obtenidas.

El objetivo de este trabajo es postular un umbral en el método Domain utilizando, además de las variables climáticas de los sitios de presencia, el conocimiento del experto. Para lograr esto se usará el enfoque de Estadística Bayesiana y las ideas de remuestreo planteadas por Argáez (1996).

Empleando el enfoque bayesiano se podrá incluir la información del experto en una distribución probabilística llamada distribución posterior. Esta distribución contendrá toda la información acerca del umbral desconocido y permitirá realizar inferencias con respecto al umbral.

Al igual que en los trabajos citados, en este trabajo se supone que se cuenta con registros legítimos de presencia de la especie de los sitios visitados en la colecta. No se tienen registros de ausencia debido a que no es posible registrar la ausencia de la especie sólo por el hecho de no encontrarla en el momento de la visita. Además es importante notar que los registros obtenidos pueden incluir la presencia de uno o más ejemplares de la especie bajo consideración en algunos sitios.

1.1. Métodos Existentes

En esta sección se describirán brevemente los métodos citados que permiten estimar zonas de alto potencial de una especie. Una característica de estos métodos es que sólo utilizan registros de sitios de presencia de la especie y sus correspondientes variables climáticas. Es importante conocer las variables climáticas de los sitios ya que son portadoras de información acerca del clima que favorece la presencia de la especie. Algunos ejemplos de variables climáticas que se consideran son: temperatura, humedad y precipitación. Por lo general en los estudios de distribución de especies se emplean variables climáticas porque son fáciles de obtener.

Para facilitar la comprensión de este trabajo, se definirán algunos conceptos importantes y se postulará la notación utilizada a lo largo del escrito. El primer

concepto que se empleará es el de *zona de alto potencial*, que es aquella zona que posee condiciones favorables para el establecimiento de la especie. Otro concepto que se usará frecuentemente es el de *sitio objetivo*, que es aquel sitio en el que se desea evaluar si la especie bajo estudio puede establecerse.

En la práctica se considera una retícula regular que cubra la región en la cual se desea evaluar la presencia de la especie. Cada nodo definido por la retícula es considerado como sitio objetivo y en cada uno se evalúa la presencia de la especie usando algún método que estime zonas de alto potencial.

Se denotará por X_0 a un sitio objetivo y por X_1,X_2,...,X_n a n sitios de presencia de la especie bajo estudio. En la práctica estos sitios X_i están definidos por coordenadas que representan su ubicación geográfica. Por ejemplo X_i puede denotar la longitud y latitud del sitio en cuestión.

Se supondrá que cada sitio X_i tiene asociado un vector de p variables climáticas conocido y que será denotado por $\overrightarrow{X_i}$. Así para un sitio objetivo X_0 de interés se tendrá:

$$\overrightarrow{X_0} = (a_{01}, a_{02}, \ldots, a_{0p}),$$

y de manera análoga se tendrán los correspondientes vectores para los sitios de presencia:

$$\begin{aligned} \overrightarrow{X_1} &= (a_{11}, a_{12}, \ldots, a_{1p}) \\ \overrightarrow{X_2} &= (a_{21}, a_{22}, \ldots, a_{2p}) \\ &\vdots \\ \overrightarrow{X_n} &= (a_{n1}, a_{n2}, \ldots, a_{np}). \end{aligned}$$

1.1.1. Método *Bioclim*

El método Bioclim (Busby, 1991) es uno de los primeros métodos utilizados para estimar zonas de alto potencial. Este método localiza los nodos X_0 de la retícula que cumplan que cada componente de su vector asociado $\overrightarrow{X_0}$ se encuentre contenido en un intervalo definido por ciertos valores determinados con base en las variables climáticas de los sitios de presencia con que se cuenta. Sean

$$\begin{aligned} m_j &= \min_{1 \leq i \leq n} \{a_{ij}\} = \text{mín}\{a_{1j}, \ldots, a_{nj}\} \\ M_j &= \max_{1 \leq i \leq n} \{a_{ij}\} = \text{máx}\{a_{1j}, \ldots, a_{nj}\}, \end{aligned} \tag{1.1}$$

que son respectivamente el mínimo y el máximo valor de la j-ésima variable climática de los sitios X_i de la muestra.

Sean $m_{j,5}$ y $M_{j,95}$ los percentiles del 5 % y 95 % de la $j-$ésima variable climática. Ahora con estos valores, se definen los intervalos $[m_j, M_j]$ y $[m_{j,5}, M_{j,95}]$, con los cuales se construyen los volúmenes M-dimensionales

$$V = \prod_{j=1}^{p} [m_j, M_j]$$

$$V' = \prod_{j=1}^{p} [m_{j,5}, M_{j,95}].$$

Es claro que tanto V como V' se encuentran definidos en el campo de las variables climáticas y que $V' \subseteq V$.

Un nodo arbitrario X_0 se clasifica como de alto potencial para la presencia de la especie si ocurre que $\overrightarrow{X_0} \in V'$, mientras que se clasifica como de alto potencial marginal si $\overrightarrow{X_0} \in V \backslash V'$. Si ocurre que $\overrightarrow{X_0} \notin V$ el nodo X_0 no se considera de alto potencial para la presencia de la especie.

Es conocido que Bioclim tiende sobrestimar las zonas de alto potencial. Esto se debe a que si un sitio de presencia X_i, $i \in \{1, 2, ..., n\}$ que posea un vector de variables climáticas con una componente que difiera marcadamente de la correspondiente componente de los vectores $\overrightarrow{X_j}$, $j \neq i$, inducirá que V posea mayor volumen, lo que producirá que un mayor número de vectores $\overrightarrow{X_0}$ pertenezca a V. Una ventaja de este método radica en que es fácil de implementar y de entender.

1.1.2. Método *Domain*

El método Domain (Carpenter, Gillison y Winter 1993) asigna un valor de similitud a cada sitio objetivo basándose en la proximidad de las variables climáticas del sitio objetivo con las variables climáticas de los sitios de presencia registrados.

Considere $X_1, X_2, ..., X_n$ los sitios de presencia registrados de la especie y sus correspondientes vectores de variables climáticas $\overrightarrow{X_1}$, $\overrightarrow{X_2}$,..., $\overrightarrow{X_n}$.

La distancia entre las variables de un sitio objetivo X_0 y las variables de un sitio de presencia X_i con $i \in \{1, 2, ..., n\}$ se calcula utilizando la métrica de Gower (Gower, 1971), la cual es dada por:

$$d(\overrightarrow{X_0}, \overrightarrow{X_i}) = \frac{1}{p} \sum_{j=1}^{p} \frac{|a_{0j} - a_{ij}|}{M_j - m_j}, \tag{1.2}$$

donde M_j y m_j se definen como se mencionó anteriormente en la expresión (1.1).

La cantidad $d(\overrightarrow{X_0}, \overrightarrow{X_i})$ representa una distancia promedio entre los vectores $\overrightarrow{X_0}$ y $\overrightarrow{X_i}$. Es importante notar que d no representa una distancia geográfica entre los sitios, sino un valor que relaciona las variables climáticas de los sitios involucrados. La métrica de Gower definida en (1.2) puede aplicarse a dos sitios arbitrarios con tal que se tengan las variables climáticas de ellos.

Se puede observar que la distancia d proporciona un valor entre 0 y 1 para sitios objetivos X_0 que cumplan que cada componente a_{0j} de $\overrightarrow{X_0}$ se encuentre dentro del intervalo $[m_j, M_j]$ para $j = 1, 2, ..., p$, el cual es determinado por las componentes de

los vectores $\overrightarrow{X_1}$, $\overrightarrow{X_2}$, ..., $\overrightarrow{X_n}$. Por otra parte, si algún sitio objetivo X_0 posee componentes a_{0j} de $\overrightarrow{X_0}$ que se encuentren fuera del intervalo $[m_j, M_j]$ entonces la distancia $d(\overrightarrow{X_0}, \overrightarrow{X_i})$ arrojará un valor mayor que 1.

Continuando con la descripción de Domain, usando la métrica de Gower se define el *valor de similitud* de X_0 (un sitio objetivo) con respecto a los sitios de presencia registrados de la siguiente manera:

$$u_{x_0} = \max_{1 \leq i \leq n} \{R(\overrightarrow{X_0}, \overrightarrow{X_i})\} = \max\{R(\overrightarrow{X_0}, \overrightarrow{X_1}), R(\overrightarrow{X_0}, \overrightarrow{X_2}), ..., R(\overrightarrow{X_0}, \overrightarrow{X_n})\}, \tag{1.3}$$

donde

$$R(\overrightarrow{X_0}, \overrightarrow{X_i}) = 1 - d(\overrightarrow{X_0}, \overrightarrow{X_i}).$$

Note que debido a que la distancia d puede tomar valores mayores que 1 para sitios que tienen componentes fuera del intervalo $[m_j, M_j]$, al calcular los valores $R(\overrightarrow{X_0}, \overrightarrow{X_i})$ se podrían obtener valores negativos. Sin embargo esto no representa un problema ya que el valor de similitud para un sitio objetivo X_0 es obtenido calculando el máximo de las variables $R(\overrightarrow{X_0}, \overrightarrow{X_i})$. De este modo si en el conjunto a partir del cual se obtendrá el valor de similitud de un sitio, se tienen tanto valores positivos como negativos de las correspondientes R's, el valor de similitud obtenido será una cantidad positiva.

Finalmente para determinar los sitios de alto potencial se elige un umbral $u \in [0, 1]$, que en general es seleccionado por el usuario. Tal umbral puede interpretarse como el mínimo valor de similitud que la especie tolera para establecerse en cierto lugar. Usando el umbral u, se clasifica a un sitio objetivo X_0 como:

$$\begin{array}{c} \text{Sitio de alto potencial, si } u_{x_0} \geq u, \\ \text{Sitio de bajo potencial, si } u_{x_0} < u. \end{array}$$

Se ha reportado que la metodología Domain sufre menos de sobrestimación que Bioclim (Carpenter, Gillison y Winter, 1993). Intuitivamente esto se debe a que, al usar Domain, un nodo X_0 será clasificado de alto potencial solamente si se cuenta con un sitio observado de presencia cuyo vector de variables sea similar a $\overrightarrow{X_0}$. Se puede observar que si un sitio de presencia X_i posee valores atípicos en su correspondiente vector de variables $\overrightarrow{X_i}$, entonces la región de alto potencial obtenida contendrá también aquellos nodos X_0 cuyo vector de variables se encuentre lo suficientemente cercano de $\overrightarrow{X_i}$, mientras que si se utiliza el método Bioclim considerando el nodo X_i con variables climáticas atípicas, el volumen V obtenido será mayor, lo que producirá que un mayor número de sitios X_0 sean clasificados como de alto potencial.

1.1.3. Método *FloraMap*

En el método conocido como *FloraMap* (Jones y Gladkov, 1999), se calcula la probabilidad de que el vector $\overrightarrow{X_0}$ pertenezca a una distribución normal multivariada,

la cual es determinada por los vectores $\overrightarrow{X_1}, ..., \overrightarrow{X_n}$ observados en los sitios de presencia.

Este método utiliza variables climáticas fijas para realizar el análisis, las cuales corresponden a mediciones mensuales de precipitación total, temperatura promedio y rangos de temperaturas, y se cuenta con un total de 36 variables. Ya que las variables se encuentran correlacionadas se realiza un análisis de componentes principales con las mediciones de las variables en los sitios de presencia. El usuario determina el número de componentes principales, r, que serán consideradas, las cuales se utilizan para construir una distribución normal r variada. Finalmente para un sitio objetivo X_0 se calcula la probabilidad de que el vector de variables se encuentre entre el origen de la distribución normal r-variada y el correspondiente vector $\overrightarrow{X_0}$ en cuestión.

Para implementar este método se ha diseñado un software que posee la desventaja de que las variables climáticas están determinadas como parte del software. Otra desventaja radica en los supuestos que se tienen para la implementación de este método. Como se comentó, se debe suponer que las variables poseen una distribución normal r-variada.

1.1.4. Método *GARP*

El algoritmo conocido como *GARP* (Genetic Algorithm for Rule-set Prediction, Stockwell y Noble, 1991; Stockwell y Peters, 1999) utiliza los sitios de presencia con que se cuenta y de los nodos de la retícula genera una muestra aleatoria de sitios que se denominan pseudoausencias. De manera iterativa busca correlaciones no aleatorias entre los sitios de presencia y los sitios de pseudoausencia, y mediante 4 diferentes subalgoritmos genera diferentes tipos de reglas de la forma: Si se cumple (Condición 1, Condición 2,...) entonces (Predicción). Las reglas son optimizadas por medio de un algoritmo genético que en cada iteración permite modificar las reglas definidas. Al final del algoritmo se obtiene un conjunto de reglas (en general de 20 a 50) con las que se determina las zonas de alto potencial. Es conocido que el algoritmo GARP puede ser inestable en algunas aplicaciones y producir soluciones subóptimas. Más aún, es conocido que si se ejecuta el algoritmo varias veces, utilizando los mismo parámetros iniciales y el mismo conjunto de datos, se puede obtener un resultado distinto en cada ejecución (Stockman, Beamer y Bond, 2006).

1.1.5. Método del Modelo Mezcla

Argáez *et al.* (2005) introduce una metodología para estimar el área de alto potencial de especies suponiendo que, además de los sitios de presencia y las variables climáticas, se cuenta con una estimación del denominado sesgo espacial. Además supone que se conoce la probabilidad de detectar un ejemplar en un sitio en el que se encuentra presente. Todas las variables climáticas se asumen medidas en escala discreta y se considera un modelo mezcla de modelos multinomiales. Usando inferencia Bayesiana, proponen un modelo Dirichlet como distribución *a priori*, resultando que la distribución posterior puede aproximarse como una mezcla de distribuciones

Dirichlet. Para cada nodo de la retícula, se estima la probabilidad de presencia de la especie bajo estudio. La desventaja de este método radica en que requiere de más elementos para estimar la probabilidad de presencia de una especie. En particular, la estimación del sesgo espacial es motivo de investigación (Fernández, 2005). Por todo lo anterior, la implementación de este método no es muy sencilla.

1.1.6. Método Maxent

Maxent (Phillips, Anderson y Schapire, 2006) estima una distribución de probabilidad objetivo encontrando la distribución de probabilidad de máxima entropía, es decir, aquella que es más dispersa o más cercana a la distribución uniforme, sujeto a un conjunto de restricciones que representan la información incompleta acerca de la distribución objetivo. La información disponible consiste en un conjunto de variables climáticas y las restricciones se proponen con base en el valor esperado empírico de cada una de las variables. Los valores empíricos se obtienen del conjunto de sitios de presencia de la especie bajo estudio. Este método no es fácil de implementar, y su funcionamiento utiliza un modelo exponencial para probabilidades, que puede producir valores predichos grandes fuera del rango de las variables climáticas de la región de estudio.

Los métodos descritos proporcionan un mapa que genéricamente se denomina *de alto potencial*, cada uno de los cuales se encuentra medido en unidades diferentes. Bioclim proporciona un mapa categórico, cuyo número de categorías es determinado por el experto, Domain proporciona un mapa de valores de similitud, FloraMap un mapa de probabilidad de pertenencia a una distribución normal multivariada, GARP proporciona un mapa binario y Modelo Mezcla y Maxent, un mapa de probabilidad. Así, no es posible de manera directa comparar los resultados obtenidos.

1.2. Organización del Trabajo

Este trabajo se organizará de la siguiente forma: En el Capítulo 2, se abordará brevemente el tema de la Estadística Bayesiana y se enunciarán algunos resultados importantes de la Teoría de Valores Extremos. Los dos temas serán utilizados para sustentar la distribución que permitirá hacer inferencia del umbral desconocido en el método Domain.

En el Capítulo 3 se propone una forma de estimar el umbral u en Domain. Se utilizarán los registros de presencia de la especie para construir nuevos valores, que serán los datos y que proporcionarán información importante para la selección del umbral. Tales valores serán modelados con una distribución Gumbel, que como se verá, es sustentada a partir de uno de los resultados importantes de la Teoría de Valores Extremos: El Teorema de tipos para extremos. Más aún, en este mismo capítulo, se propondrá una forma de elicitar los parámetros de la distribución *a priori* usando información dada por el experto. Por último se mostrará la expresión de la distribución posterior encontrada.

En el Capítulo 4 se presentarán aplicaciones a dos casos reales. La primera aplicación será para especie *Dictyanthus aeneus Woodson*, que es una especie endémica de la Península de Yucatán. La segunda aplicación será para la especie *Ateles geoffroyi*, conocida comúnmente como mono araña. Los datos utilizados en cada aplicación contienen variables climáticas de los sitios de presencia de la especie. Los sitios de presencia utilizados para la primera especie se obtuvieron del herbario de CICY y los sitios de la segunda se obtuvieron de la publicación Mapping primate populations in the Yucatan Peninsula, Mexico: A first assessment (Serio-Silva, Rico-Gray, y Ramos-Fernández, 2006). La información y datos en las dos aplicaciones fueron proporcionados por la Dra. Celene Espadas del Centro de Investigación Científica de Yucatán (CICY).

Para cada especie bajo estudio se obtendrá la distribución posterior, siguiendo la metodología propuesta en el Capítulo 3 y se calculará la moda de la distribución resultante. La moda será el valor propuesto para el umbral, ya que es el que tiene la máxima probabilidad en la distribución posterior. Como resultado adicional, en este trabajo se obtendrá un intervalo de máxima credibilidad posterior para el umbral. Además del mapa que se obtendrá usando el valor estimado del umbral, se emplearán los extremos del intervalo de credibilidad para generar las zonas de alto potencial de Domain, los cuales permitirán tener un escenario conservador (con el límite inferior del intervalo) y un escenario menos conservador (con el límite superior del intervalo). Finalmente, con el fin de comparar el método propuesto en este trabajo, se obtendrán para cada una de las especies, las zonas de alto potencial usando el umbral 0.9 y el umbral propuesto por Argaéz (1996), respectivamente.