CIMAT
BIBLIOTECA

CIMAT

Centro de Investigación en Matemáticas, A.C.

Estimación de la Probabilidad de Presencia de una Especie con base en Mediciones de covariables y Diseño de Zonas para Proteger Especies

TESIS

que para obtener el grado de
DOCTOR EN CIENCIAS

con orientación en
Probabilidad y Estadística

P R E S E N T A

Jorge Armando Argáez Sosa

Guanajuato, Gto., octubre de 2003

C I M A T
BIBLIOTECA
018339

Agradecimientos

Agradezco a mis asesores, Dr J. Andrés Christen Gracia y Dr. Miguel Nakamura Savoy por la dirección de esta tesis.

Al Dr. Juan José Fernández Durán, Dr. Rafael Durán García, Dr. Carlos Díaz Ávalos y Dr. Enrique Villa Diharce, por la revisión de este trabajo y ser mis sinodales.

A la C. Dr. Celene Espadas Manrique, por sus valiosos comentarios y toda su ayuda a lo largo del desarrollo de esta tesis.

A Paulino Simá Polanco, por su colaboración en los casos de estudio.

Al CICY, por todo el apoyo y facilidades que me otorgó para realizar mis estudios de doctorado.

Al CIMAT, por proporcionarme las instalaciones necesarias para la feliz culminación de mi trabajo de tesis.

Al CONACYT, por apoyarme con la beca número 115344.

A mis hermanos Elizabeth Argáez Sosa, Silvia Argáez Sosa y Javier Vallado Flores. Gracias por ser apoyo en todo momento.

A mis suegros, José H. Cruz Canché y Rilma C. Reyes Méndez, por sus oraciones y palabras de ánimo.

A mis amigos Ramón y Lilia. Gracias por los momentos que me han permitido compartir con ustedes.

Contenido

Introducción

Los mapas conocidos como *Mapas de Habitación Potencial* o simplemente *Mapas de Presencia Potencial* de una especie han surgido como una herramienta relevante para contestar preguntas biológicas y ecológicas acerca de la especie y permiten planear acciones con el fin de que los recursos naturales se utilicen de manera apropiada. En un mapa de habitación potencial de una especie se observan las zonas en las que dicha especie puede establecerse con mayor probabilidad. Estas zonas, que se interpretan como las zonas de alto potencial de establecimiento de la especie, se relacionan con el área de distribución de la especie, que es una de las expresiones fundamentales de su ecología e historia evolutiva (Udvardy, 1969; Brown, Stevens y Kaufman, 1996; Gaston y Blackburn, 2000). El conocimiento de las zonas de alto potencial de establecimiento permite abordar preguntas científicas que aumentan el conocimiento acerca de una especie (Peterson, Soberón y Sánchez-Cordero, 1999) y también es útil para abordar problemas prácticos, como estimar las rutas de entrada de especies denominadas invasoras (Soberón, Golubov y Sarukhán, 2001), evaluar el daño que podría ocasionar la entrada de especies-plaga a una región (Sánchez-Cordero y Martínez-Meyer, 2000) y otras.

El hecho de conocer las regiones donde una especie de interés es capaz de establecerse con mayor probabilidad, es decir, los mapas de presencia potencial, permite planear y realizar actividades relacionadas con el estudio, cuidado, manejo, etc. de la especie. En este contexto, un problema al que se enfrentan las agencias dedicadas al estudio y cuidado de la biodiversidad es determinar cuál región (o regiones) es adecuada para declararla zona protegida. No existe una definición estándar de lo que significa proteger una región. Por ejemplo, el Sistema Nacional de Áreas Naturales Protegidas postula que:

> "Un área natural protegida es aquella zona en la que los ambientes originales no han sido significativamente alterados por la actividad humana o que requieran ser restauradas, y que hayan quedado sujetas a cualquiera de los regímenes de protección."

Por su parte, la Unión Internacional para la Conservacion de la Naturaleza y Recursos Naturales (IUCN, por sus siglas en inglés) propone la siguiente definición:

> "Una región protegida es una porción de tierra o mar especialmente dedicada para la protección y mantenimiento de diversidad biológica y de los recursos naturales y culturales asociados, y su manejo a través de medios legales o cualquier otro medio adecuado efectivo."

En esta tesis se abordan dos problemas que surgen de acuerdo con lo presentado en párrafos anteriores. El primero es encontrar el denominado mapa de establecimiento potencial de una especie de interés, con base en sitios de presencia de la especie y en valores de covariables. El segundo problema consiste en determinar una zona que se propondrá para proteger especies de interés, con base en la probabilidad de presencia de cada especie en sitios particulares de interés. En la formulación de ambos problemas se supone que la región bajo estudio se encuentra cubierta por una retícula regular. Para abordar el primer problema se asume, además, que para cada nodo de la retícula se conoce el valor de un conjunto de covariables físicas y/o climáticas, medidas u observadas en dicho nodo. Los datos con que se cuenta consisten de un conjunto de sitios en los que la especie bajo estudio ha sido localizada.

Debido a que la zona de estudio se supone cubierta por una retícula, el primer problema que se aborda (Capítulo 1), encontrar el mapa de establecimiento potencial de una especie, consiste en estimar la probabilidad de presencia de la especie en cada nodo de la retícula. Al desplegar en una gráfica la probabilidad obtenida en cada nodo se genera el mapa de probabilidades de presencia.

Existen varios métodos diseñados para abordar este problema, ninguno de los cuales se encuentra formulado formalmente, lo que produce que dichos métodos posean desventajas claramente identificadas. La primera desventaja que se observa es que cada método produce resultados cuya interpretación difiere de la interpretación de los resultados de los otros métodos. En otras palabras, no se cuenta con una manera formal de evaluar la presencia potencial de una especie. Debido a la falta de sustento estadístico en la formulación de los métodos existentes, ninguno de ellos produce alguna medida que cuantifique la certidumbre que se tiene para el potencial.

Los métodos existentes no identifican, y por tanto no utilizan, algunos elementos evidentes que en esta tesis se utilizan para formular un modelo estadístico para abordar este problema. Uno de los elementos que no consideran esos métodos es el hecho de que los sitios de presencia se encuentran localizados, principalmente, cerca de carreteras y asentamientos humanos. Este hecho, denominado *sesgo espacial*, se relaciona con la distribución heterogénea que se observa en los sitios de presencia.

Otro elemento que no consideran formalmente los métodos existentes es el conocimiento que un experto versado en la especie puede aportar con respecto a las zonas de establecimiento y no establecimiento de una especie bajo estudio. En la práctica, el conocimiento de un experto se utiliza *a posteriori*, para corregir errores evidentes observados en el mapa de habitación potencial que resulta. En esta tesis se identifica una manera en la que un experto

puede aportar su conocimiento, y éste se utiliza formalmente en el proceso de inferencia.

El problema descrito se aborda considerando de manera explícita los elementos que se identifican y que no se utilizan en los otros métodos. Además, se propone una forma de calificar la probabilidad de presencia obtenida en cada nodo, por medio de un mapa paralelo denominado *mapa de certidumbre*.

El supuesto que se adopta para inferir la probabilidad de presencia de la especie en un nodo es que dicha probabilidad es igual a la probabilidad de que la especie seleccione el vector de covariables presente en el nodo para establecerse. Así, el parámetro de interés es un vector de probabilidades, en el que cada componente es la probabilidad de presencia de la especie en una posible combinación de covariables. Para la formulación del modelo se postula que un registro de presencia en un nodo particular es producto de tres factores: la presencia de la especie en el nodo, la visita al nodo y la detección de la especie.

El modelo que se propone surge del hecho de que sólo se cuenta con registros de presencia, y los datos con que se cuenta para modelar son los vectores de covariables observados en los sitios de presencia. Ya que cada covariable se asume medida en escala discreta, los datos con que se cuenta pueden resumirse por medio de un vector de conteos, en el que cada componente representa el número de sitios de presencia que poseen alguna combinación específica de valores de covariables.

La consideración de un modelo con base en todas las posibles combinaciones de valores de covariables produce un problema de estimación con datos escasos, pues el número de sitios de presencia es usualmente pequeño en comparación con el número total de combinaciones de valores de covariables posibles. Para evitar esto se propone considerar una reducción en la dimensionalidad del parámetro de interés. Para esto, se considera un modelo mezcla, en el que cada término corresponde a una pareja de covariables. El parámetro de interés para cada pareja será un vector de probabilidades, en el que cada componente es la probabilidad de presencia de la especie en una posible combinación de valores de la pareja de covariables. Así, se consideran todas las parejas de covariables para construir el modelo, lo que en general produce una reducción en la dimensión del parámetro de interés.

Para cada pareja de covariables se considera un modelo multinomial, en el que cada celda corresponde a una posible combinación de valores de la pareja de covariables. La probabilidad de observar una presencia de la especie en una celda se postula como el producto de la probabilidad de presencia de la especie, la probabilidad de visitar el vector de covariables del nodo y la probabilidad de detectar a la especie.

Para hacer la inferencia acerca de la probabilidad de presencia, la distribución *a priori* se postula como el producto de distribuciones *a priori* correspondientes a cada una de las parejas de covariables. Ya que el interés es realizar inferencias acerca de vectores de probabilidades, se propone una distribución Dirichlet para cada pareja de covariables.

Para postular (elicitar) los parámetros de cada una de las distribuciones Dirichlet, se utiliza la información que posee el experto (Capítulo 3). En esta tesis se identifica una

manera en la que el experto puede aportar la información que posee acerca de las regiones de presencia y/o ausencia de la especie, por medio de mapas denominados mapas *a priori* de presencia y ausencia. Utilizando esta información se propone un mecanismo para postular (elicitar) los valores de los parámetros de la distribución Dirichlet. Con estos valores a la mano, se procede a realizar la inferencia acerca de la probabilidad de presencia.

Utilizando el Teorema de Probabilidad Total, se obtiene que la probabilidad de presencia de la especie en un nodo está dada por una mezcla, donde cada componente es la probabilidad predictiva de presencia de la especie en el nodo, para una pareja de covariables. Al evaluar la probabilidad de presencia de la especie en cada nodo de la retícula se obtiene el mapa de probabilidades de presencia de la especie.

Además de estimar la probabilidad de presencia, se propone una medida que califica la certidumbre de la probabilidad que se obtiene en cada nodo. La certidumbre se calcula a partir de las distribuciones marginales posteriores. Al desplegar en un mapa los resultados se obtiene el denominado *mapa de certidumbre*, que utilizado junto con el mapa de probabilidades de presencia, permite realizar inferencias con respecto a las regiones de alto potencial de establecimiento de la especie con base en mayor sustento estadístico.

Para estudiar el desempeño de la metodolgía que se propone se diseñó un estudio de simulación de acuerdo con el siguiente esquema. Como primer paso se genera un mapa de probabilidades de presencia de una especie ficticia, con base en la consideración de que cada especie posee un vector de covariables óptimo, en el cual encuentra las condiciones favorables para su establecimiento. Dado el vector de covariables óptimo, se postula que la probabilidad de establecimiento de la especie en un nodo es función de la distancia entre el vector de covariables observado en el nodo y el vector de covariables óptimo. El mapa que se obtiene se considera el mapa real de probabilidades de establecimiento de la especie.

Con base es este mapa se generan sitios de presencia, de acuerdo con el mecanismo detectado bajo el cual ocurren las presencias. Esos sitios se utilizan para ejecutar los métodos existentes, incluyendo el que se propone en esta tesis, y cada mapa que resulte se compara con el mapa real de probabilidades. Este esquema permite evaluar, de manera cualitativa, el desempeño de cada método. Se observó que la metodología que se propone en esta tesis arroja resultados adecuados en cada caso analizado. Los resultados observados aportan evidencia de que la metodología es robusta al sesgo espacial presente en los sitios de presencia, y también es robusta a sitios de presencia localizados lejos de las zonas de alta probabilidad de presencia de la especie.

Con respecto al segundo problema (Capítulo 2), decidir una zona para proteger especies, en la literatura se proponen algunas formas de proceder, principalmente utilizando el enfoque de programación lineal. En ellas se supone que se conocen algunos elementos que en la práctica no se encuentran fácilmente disponibles, o bien, cuya postulación restringe marcadamente el conjunto de soluciones. Por ejemplo, se supone que se conoce el conjunto de nodos de presencia y de ausencia de cada especie considerada, es decir, que se conoce el

área de distribución de la especie, lo que en la práctica es en sí un problema de investigación, precisamente relacionado con el problema que se aborda en el Capítulo 1 de esta tesis. Otro supuesto que comúnmente se adopta es que se conoce el número de nodos que se desea proteger. Bajo este supuesto, la solución que se propone no puede contener más nodos que el número especificado, lo que restringe marcadamente el conjunto de soluciones posibles.

Desde el punto de vista ecológico, existen algunos hechos relevantes que no se consideran en los métodos propuestos. Uno de éstos es el hecho de que las especies bajo estudio pueden encontrarse en diferente situación con respecto a la urgencia que se tenga por protegerlas. Otro hecho que no se considera en esos trabajos es que algunas especies pueden ser, en algún sentido, más valiosas que otras. En otras palabras, no se considera un *valor biológico* para cada una de las especies. La consideración de estos elementos permite obtener zonas para proteger a las especies con base en información relevante acerca de las mismas.

En esta tesis se aborda este problema utilizando el enfoque de la Teoría de Decisiones, bajo el cual es posible considerar los elementos que los métodos existentes no utilizan. Bajo este enfoque se deben definir tres elementos básicos: *el espacio de acciones*, *el espacio de estados de la naturaleza* y una *función de pérdida*. El espacio de acciones conjunta todas las posibles decisiones que pueden tomarse en el contexto del problema planteado. El espacio de estados de la naturaleza conjunta los posibles resultados acerca del fenómeno estudiado. La función de pérdida asigna un valor a cada pareja *acción-estado de la naturaleza*, el cual se interpreta como la pérdida que se obtiene si se ejecuta la *acción* y ocurre el *estado de la naturaleza* especificado. Bajo este esquema, la decisión (o acción) que se tome será aquella que minimice la pérdida esperada de la función de pérdida.

Para abordar este problema se procede primeramente a considerar el caso en el que se debe tomar la decisión en un nodo particular y se estudia una sola especie. La función de pérdida para este caso se define en términos de cantidades mediante las cuales se incluyen los factores detectados y que otros enfoques no consideran. Por ejemplo, una de las cantidades que define la función de pérdida se interpreta como el valor biológico de la especie.

La función de pérdida propuesta para este caso se utiliza para abordar el problema de considerar varias especies a la vez en un solo nodo. La función de pérdida se define como la suma de funciones de pérdidas correspondientes a cada especie en particular. Esta función de pérdida considera la posibilidad de que algunas especies se postulen como más urgentes para ser protegidas que otras.

Seguidamente se aborda el caso en el que se desea seleccionar una región (conjunto de nodos) que se propondrá para proteger. En la práctica existen algunos hechos que conducen a considerar dos restricciones sobre la región que se propondrá para proteger. La primera restricción surge por el hecho de que, en general, se cuenta con un presupuesto establecido para determinar una región para proteger. Esta restricción se involucra en el espacio de decisiones, considerando como espacio de decisiones aquellas regiones (conjuntos de nodos) cuyo costo sea menor que el presupuesto con que se cuenta. Para esto se supone que se

cuenta con una cantidad para cada nodo, la cual se interpreta como el costo de dicho nodo.

La segunda restricción se relaciona con la posible preferencia que los usuarios pueden tener por proteger regiones que no sean altamente fragmentadas. Al considerar esta restricción será posible proponer una solución obtenida bajo la perspectiva de un debate que existe en la literatura. Este debate surge de la pregunta ¿es preferible seleccionar una región conexa para proteger, la cual posea cierta área, o es preferible proteger varias regiones de menor área, que en conjunto posean la misma área que la región conexa? En esta tesis, en lugar de aportar evidencia en favor de una u otra opción del mencionado debate, se propone una función de pérdida que permite explorar ambas opciones. Para considerar la segunda restricción, denominada *restricción por conexidad*, la función de pérdida que se propone, que es una generalización de la función de pérdida propuesta en el caso de proteger un solo nodo, incluye un término adecuado que permite al usuario imponer diferente importancia a la preferencia por zonas conexas. Por medio de este término será posible explorar diferentes opciones con respecto al debate mencionado. Las cantidades que definen las funciones de pérdida se determinan utilizando información de un experto con respecto a las especies y por medio de información adicional con que se cuenta de la región bajo estudio (Capítulo 3).

Ya que la región que se proponga para proteger es conformada por un conjunto de nodos, el espacio de soluciones para el caso general es, en principio, el conjunto potencia de los nodos de la retícula. La decisión que se tome será el conjunto de nodos que posea la menor pérdida esperada. Aunque en esta tesis se considera el espacio de decisiones restringido por la restricción por presupuesto, la cardinalidad del espacio de soluciones es tal que no es posible considerar todos los conjuntos para evaluar la pérdida esperada de cada uno y seleccionar el conjunto de nodos que produce la menor pérdida. Por lo tanto, se recurre a algoritmos de búsqueda para obtener una solución. Se propone considerar dos algoritmos de búsqueda de manera conjunta, lo que permite aprovechar las ventajas de ambos para encontrar una solución.

Para observar el funcionamiento de la metodología que se propone se efectuó un ejercicio de simulación, en el que se consideran 3 especies de interés y se postulan diferentes valores para las cantidades que definen la función de pérdida. En el ejercicio de simulación se observó que las cantidades de las que depende la función de pérdida son relevantes para determinar la región que se propondrá para proteger, dependiendo de la situación de las especies con respecto a la urgencia que se tenga por protegerlas y a su valor biológico. Una especie a la que se imponga mayor importancia para ser protegida, o bien, que se considere como poseedora de mayor valor biológico, tendrá mayor influencia en la zona que se propondrá para proteger.

Las ideas que se proponen en esta tesis para abordar los dos problemas descritos se utilizan en el Capítulo 4 para estudiar especies reales sobre una región de estudio. Primeramente se estudia una sola especie (Sección 4.1), para la que se estima el mapa de probabilidades de presencia de la especie y con base en este mapa, se encuentra una zona que se propondrá para

proteger, bajo diferentes condiciones. Posteriormente se encuentra una zona para proteger a las especies (Sección 4.2). En ambos casos se exploran diferentes escenarios para encontrar las regiones a proteger.

Capítulo 1

Estimación de la Probabilidad de Presencia

En este capítulo se aborda el primer problema descrito en la introducción: dado un conjunto de sitios (localidades) en los que una especie de interés ha sido observada y un vector de M covariables medidas u observadas en cada uno de esos sitios, estimar la probabilidad de presencia de la especie en sitios de interés, en los que también se cuenta con los valores de las M covariables. Si en particular se estima esa probabilidad en cada nodo de una retícula se genera el mapa de probabilidades de presencia sobre la región de estudio.

Los métodos actualmente disponibles con los que se aborda este problema son: *Domain* (Busby, 1991), *Bioclim* (Carpenter, Gillison y Winter, 1993), *FloraMap* (Jones y Gladkov, 1999) y *GARP* (Genetic Algorithm for Rule-set Prediction, Stockwell y Noble, 1991; Stockwell y Peters, 1999; Peterson y Cohoon, 1999; Peterson, Stockwell y Kluza, 2002), con los que se obtiene un mapa que genéricamente se denomina *mapa de establecimiento potencial*, o simplemente, *mapa de potencial*.

Para aplicar cualquiera de dichos métodos, los datos fundamentales consisten en un conjunto de sitios donde la especie bajo estudio ha sido localizada. Estos datos se denominan *registros (o sitios) de presencia*. En esta tesis se asume que sólo se cuenta con legítimos registros de presencia. Si también se contara con legítimos registros de ausencia, este problema podría abordarse utilizando diversos enfoques, tales como modelos lineales generalizados (Austin, 2002), modelos autologísticos (Pettitt, Weir y Hart, 2002) o Kriging (Heagerty y Lele, 1998).

Al estudiar la forma de proceder de los métodos mencionados resulta evidente que (1) no cuentan con sustento estadístico formal, (2) no existe una definición formal de lo que es establecimiento potencial, (3) no proporcionan alguna medida de la precisión del resultado que generan como potencial y (4) no identifican (y por lo tanto no utilizan) formalmente posible información que expertos pueden aportar con respecto a las áreas de habitación de las especies. La falta de una definición formal del concepto de *potencial* se percibe al observar

que cada uno de los métodos mencionados proporciona un resultado cuya interpretación es diferente a la interpretación de los resultados obtenidos con los otros métodos. Así, mientras Bioclim proporciona un mapa categórico, Domain proporciona un mapa de valores de similaridad, FloraMap produce un mapa descrito como "mapa de probabilidades de que un vector de covariables pertenezca a una distribución normal multivariada" y GARP proporciona un mapa binario. Con respecto a (3), la falta de sustento estadístico de cada método no permite dotar al valor de potencial que generan de una medida de precisión.

Con respecto a (4), cuando se aplica alguno de los algoritmos mencionados la opinión o conocimiento de los expertos (quizá uno solo) se utiliza para corregir errores evidentes del mapa de potencial que se obtiene. Al observar algún mapa de potencial los expertos frecuentemente reducen las superficies obtenidas, utilizando su conocimiento acerca de la especie, sin recurrir a mecanismos explícitos o a criterios formales. En este capítulo se propone un método que considera los puntos (1)-(4) en la modelación y en el proceso de inferencia.

Un hecho que es relevante en la modelación es la identificación de los elementos que intervienen en el proceso con el que se obtienen los sitios de presencia. El primer elemento surge de observar que los sitios de presencia típicamente ocurren agrupados alrededor de carreteras o cerca de asentamientos humanos. Este hecho motiva a considerar la existencia de un concepto, denominado *sesgo espacial*, el cual se asocia con la distribución geográfica heterogénea observada en los sitios de presencia. Ya que para cada sitio de presencia se conocen los valores de M covariables, cualquier distribución geográfica de un conjunto de sitios sobre la región de estudio induce una distribución de puntos en el espacio de covariables. Los puntos en el espacio de covariables pueden ser no uniformemente distribuidos, por lo que la noción de *sesgo en las covariables* puede también estar presente. El sesgo en las covariables determina la probabilidad de que un sitio con un conjunto de covariables específico sea físicamente examinado para evaluar la presencia de la especie.

Una vez que se considera un sitio como visitado, se identifica la noción de la llamada *detectabilidad* de la especie, que surge de la observación de que aún si un sitio donde la especie se encuentra presente es físicamente examinado, la especie puede no ser detectada. La detectabilidad se interpreta como la probabilidad de observar la presencia de la especie dado que se encuentra presente en el nodo visitado.

Conjuntando los elementos descritos, la *probabilidad de observación* se refiere a la probabilidad de registrar la presencia de la especie en un sitio, una vez que la probabilidad de presencia, el sesgo en las covariables y la detectabilidad han sido consideradas. Así, un registro de presencia ocurre cuando la especie se encuentra presente en un nodo, el nodo es visitado por observadores y éstos detectan la presencia de la especie. La concepción de que un sitio de presencia es generado por este mecanismo constituye la base del modelo que se propone en este capítulo (Sección 1.1).

Un supuesto relevante en el que se basa la construcción del modelo es que la probabilidad

de presencia de una especie en un nodo s es igual a la probabilidad de que el vector de covariables observado en dicho nodo sea seleccionado por la especie para establecerse. Si $\mathbf{V} = (V_1, \ldots, V_M)$ representa el vector aleatorio de valores de covariables seleccionados por la especie cuando se hace presente en un nodo $s \in R$, la cantidad de interés es $P\{\mathbf{V} = \mathbf{e}(s)\}$, donde $\mathbf{e}(s)$ es el vector de covariables observado en s. Para generar el mapa de probabilidades de presencia se deberá evaluar $\theta(f) = P(\mathbf{V} = f)$ para todo f posible. Así, el parámetro de interés es el vector $\boldsymbol{\theta} = (\theta(f))_{f \in F}$, donde F denota el conjunto de todos los vectores de covariables posibles sobre R. La probabilidad de presencia de la especie en un nodo s será entonces la cantidad $\theta(\mathbf{e}(s))$.

Los datos con que se cuenta en los sitios de presencia se resumen por medio de un vector de conteos, denotado por $\mathbf{C}$, que contiene información acerca del número de sitios de presencia en los que se observa cada vector de covariables posible. Ya que el número de sitios de presencia es usualmente menor que el conjunto de vectores de covariables posibles, la consideración de un modelo que incluya todos los elementos del vector $\mathbf{C}$ dará lugar a un problema de estimación con datos escasos. Para evitar esto se propone considerar las $C_{M,2}$ parejas de covariables como objetos de estudio, como se explica en los siguientes párrafos. En general esta forma de modelar produce una reducción en la dimensión del parámetro de interés.

Para una pareja de covariables $J = (a, b)$ fija, con $1 \leq a < b \leq M$, se definen los correspondientes elementos $\mathbf{V}_J = (V_a, V_b)$, $\mathbf{e}_J(s) = (e_a(s), e_b(s))$ y $\theta_J(g) = P(\mathbf{V}_J = g)$, donde g es un posible vector de covariables para la pareja J. Sea $C_J(g)$ el número de sitios de presencia tales que $\mathbf{e}_J(s_i) = g$. Bajo este esquema, el vector que resume la información con que se cuenta es $\mathbf{C}_J = (C_J(g))_{g \in F_J}$, mientras que el parámetro de interés es $\boldsymbol{\theta}_J = (\theta_J(g))_{g \in F_J}$, donde F_J es el conjunto de todos los valores posibles para la pareja J. Al considerar todas las parejas de covariables, los datos con que se cuenta se resumen mediante el vector $\mathbf{C}' = (\mathbf{C}_J)_{J \in G}$ y el parámetro de interés es el vector $\boldsymbol{\theta}' = (\boldsymbol{\theta}_J)_{J \in G}$, donde G es el conjunto de las $C_{M,2}$ parejas de covariables.

Para cada pareja de covariables se propone un modelo multinomial para el vector de conteos $\mathbf{C}_J$. La observación de un dato en una celda de esta multinomial es producto de los elementos descritos en párrafos anteriores (probabilidad de presencia, sesgo en covariables y detectabilidad). Con base en estos elementos se postula un modelo mezcla, el cual considera todas las parejas de covariables. Cada componente de la mezcla es precisamente un modelo multinomial con parámetro $\boldsymbol{\theta}_J$.

Para realizar la inferencia acerca de la probabilidad de presencia se involucra la información *a priori* con que se cuenta (Sección 1.2), utilizando el enfoque de la teoría Bayesiana. Bajo este enfoque se postula una *distribución a priori* para el parámetro de interés, la cual depende a su vez de ciertos parámetros. La información dada por el experto se utiliza para especificar valores para los parámetros que definen esta distribución (Capítulo 3). Como es usual cuando se consideran modelos mezcla desde el enfoque Bayesiano, la distribución *a*

priori se formula como el producto de las distribuciones *a priori* postuladas para los vectores $\boldsymbol{\theta}_J$. Para cada $\boldsymbol{\theta}_J$ se propone una distribución Dirichlet, que es comúnmente utilizada para modelar vectores de probabilidades, como en el caso que nos ocupa. La interpretación que poseen los parámetros de la distribución Dirichlet permiten proponer un mecanismo de elicitación sencillo con el cual utilizar la información del experto para fijar cantidades particulares para los parámetros *a priori* (Sección 3.2).

Para estimar la probabilidad de presencia de la especie en un nodo (Sección 1.2) se utiliza el Teorema de Probabilidad Total, que produce que dicha probabilidad este dada por la suma ponderada de probabilidades predictivas de presencia de cada pareja de covariables en dicho nodo. Por su parte, la probabilidad predictiva de presencia para cada pareja de covariables se infiere con base en la distribución marginal posterior del parámetro de interés.

Además de estimar la probabilidad de presencia en un nodo, en este capítulo se introduce una manera de dotar al estimador de una medida de precisión (Sección 1.3), utilizando la distribución marginal posterior de cada pareja y la idea del modelo mezcla. Ya que se obtiene una medida de precisión para cada nodo de la retícula, es posible desplegar un segundo mapa, que se denomina *mapa de certidumbre*. El uso de este mapa conjuntamente con el mapa de probabilidades de presencia, permitirá realizar inferencias con respecto a las áreas de alto potencial de establecimiento de la especie con base en mayor sustento estadístico.

La distribución marginal posterior que resulta no corresponde a alguna expresión conocida, por lo que para estimar las cantidades de interés, la probabilidad de presencia y una medida de certidumbre, debe recurrirse a métodos de simulación de Monte Carlo vía Cadenas de Markov (MCMC) (Sección 1.4). En esta tesis se utiliza el algoritmo conocido como Metropolis-Hastings, que se implementó postulando una distribución de propuesta independiente de la familia Dirichlet. Sin embargo, con base en la génesis con la que se obtienen los sitios de presencia, se encontró una distribución alterna que aproxima a la distribución posterior exacta y permite obtener las cantidades de interés por medio de cálculos cerrados. Esta aproximación resulta útil para un usuario no versado en la implementación de algoritmos de simulación MCMC.

Para explorar el funcionamiento del método que se propone, se realizó un experimento de simulación (Sección 1.5) de acuerdo con el siguiente esquema. Se postula que cada especie posee un vector de covariables *ideal*, el cual determina las condiciones óptimas para su establecimiento. Para simular un mapa de probabilidades de establecimiento se procede a fijar el vector de covariables ideal. Con base en éste, se postula que la probabilidad de presencia de la especie en un nodo s se obtiene como función de la distancia entre el vector de covariables del nodo, $\mathbf{e}(s)$, y el vector ideal. Al desplegar las probabilidades generadas para cada nodo de una retícula se obtiene el denominado *mapa real de probabilidades de presencia*. Con base en las probabilidades obtenidas, se generan sitios de presencia siguiendo el esquema identificado con el que éstas ocurren, es decir, considerando la probabilidad de presencia de la especie, el sesgo espacial (que induce el sesgo en las covariables) y la

detectabilidad. A los sitios de presencia que se generan se les aplica el método que se propone y los métodos alternos. Los mapas de potencial que se obtienen se comparan con el mapa real de probabilidad de presencia, con lo que se observa de manera cualitativa el funcionamiento de cada uno de los métodos.

Los experimentos de simulación realizados aportan evidencia de que el método propuesto es robusto al hecho de que los sitios de presencia se encuentran geográficamente sesgados, es decir, localizados en su mayoría cerca de carreteras y/o asentamientos humanos. También se obtuvo evidencia de que el método no es afectado por sitios de presencia que se encuentran geográficamente lejos de la región simulada de alta probabilidad de presencia. Estos sitios no producen una zona con alta probabilidad de presencia de la especie alrededor de ellos, a diferencia de los métodos alternos.

1.1 Modelo Estadístico

Para formalizar las ideas sobre las cuales se basa la modelación, se introduce la siguiente notación. Sea R el conjunto de nodos determinados por la retícula regular que cubre la región de interés. Sea $\mathbf{e}(s) = (e_1(s), \ldots, e_M(s))$ el vector de covariables observado en $s \in R$, el cual es conformado por los valores de M covariables medidas u observadas en s. Todas las covariables se asumen categóricas, o bien, que se encuentran medidas en escala discreta, por lo que si alguna covariable se encuentra medida en escala continua, se procederá a clasificarla en categorías. Así, se tiene que $e_k(s) \in \{1, ..., R_k\}$, $1 \leq k \leq M$, donde R_k representa el número de clases o categorías de la k-ésima covariable. Denótese por $F = \{1, ..., R_1\} \times \cdots \times \{1, ..., R_M\}$ al conjunto de todas las posibles configuraciones de las covariables sobre R.

Sea $u(s)$ una variable aleatoria binaria que toma el valor 1 si la especie se encuentra presente en el nodo s y el valor 0 si no lo está. El mapa de probabilidades de presencia para la especie sobre R es la probabilidad $P\{u(s) = 1\}$, como función de s. Esta probabilidad es precisamente la que se define como probabilidad de presencia de la especie en un nodo particular s, y es la cantidad de interés en este estudio.

Una noción fundamental que se considera en este capítulo consiste en que la presencia de la especie en un nodo es determinada por los valores de las covariables del nodo y no por la posición geográfica del mismo. En otras palabras, se postula que la especie decide establecerse en s si el correspondiente $\mathbf{e}(s)$ satisface ciertas condiciones propias de la especie. Sea $\mathbf{V} = (V_1, \ldots, V_M)$ el vector aleatorio de valores de covariables tácitamente seleccionadas por la especie cuando se hace presente en un nodo. Así, la suposición fundamental que justifica la inferencia de las denominadas zonas de alto potencial de establecimiento de la especie, con base en las covariables observadas en sitios de presencia reportados, puede

describirse mediante la igualdad

$$P\{u(s) = 1\} = P\{\mathbf{V} = \mathbf{e}(s)\}, \tag{1.1}$$

que postula que la probabilidad de presencia de la especie en s es igual a la probabilidad de que el vector de covariables de s sea seleccionado por la especie para hacerse presente en el nodo.

Los sitios de presencia con que se cuenta son resultado de visitas realizadas a sitios sobre la región de interés. Sea N el número total de visitas realizadas durante el período de tiempo considerado en el estudio. Estas N visitas dan lugar a los n registros de presencia con que se cuenta, algunos de los cuales pueden ocurrir en un mismo nodo.

Debido a que sólo se cuenta con registros de presencia de la especie, la información que es posible observar y estudiar se encuentra contenida en los vectores de covariables de los sitios $s_1, ..., s_n$. Estos sitios corresponden a localidades geográficas exactas donde ejemplares de la especie han sido localizados. El vector de covariables que se asigna a cada s_i es el vector de covariables del nodo de R mas cercano a s_i. Así, los datos con que se cuenta son en realidad los vectores $\mathbf{e}(s_1), ..., \mathbf{e}(s_n)$.

Sea $C(f)$ el número de elementos de la muestra tales que $\mathbf{e}(s_i) = f$, $f \in F$, $1 \leq i \leq n$, y sea $\mathbf{C} = \{C(f)\}_{f \in F}$ el vector de todos los conteos observados, los cuales se asumen ordenados de acuerdo con el orden lexicográfico de F. El vector $\mathbf{C}$ resume los datos observados, por lo que cualquier intento de modelado deberá estar enfocado en describir el comportamiento probabilístico de $\mathbf{C}$. En la práctica, es común que el número de sitios de presencia con que se cuenta sea pequeño con respecto a $|F|$, por lo que muchos de los elementos de $\mathbf{C}$ serán de hecho cero. Por esta razón, un modelo paramétrico que considere todos los miembros de F dará lugar a un problema de estimación con datos escasos (*sparse data*). Para evitar esto, se propone una reducción en la dimensionalidad de los parámetros, la cual se basa en la consideración de interacciones de parejas de covariables, como se explica a continuación.

Sea G el conjunto de todos los pares de índices (a, b), $1 \leq a < b \leq M$. Para reducir notación, sea $J = (a, b)$ un par genérico de G. Sea $\mathbf{e}_J(s) = (e_a(s), e_b(s))$ el vector conformado por los valores de la pareja J observados en s. Es decir, $\mathbf{e}_J(s)$ contiene los valores de la a-ésima y b-ésima coordenadas del correspondiente $\mathbf{e}(s)$. Sea $F_J = \{1, \ldots, R_a\} \times \{1, \ldots, R_b\}$ el conjunto de todas las posibles combinaciones de valores para la pareja J. Para un elemento $g \in F_J$, sea $C_J(g)$ el número de sitios en la muestra que satisfacen $\mathbf{e}_J(s_i) = g$, y sea $\mathbf{C}_J = \{C_J(g)\}_{g \in F_J}$ el vector de todos los conteos obtenidos de esta manera.

Simplificando la suposición (1.1), postulando que la presencia de la especie es determinada sólo por el vector $\mathbf{V}_J = (V_a, V_b)$ correspondiente a la pareja $J = (a, b)$ en lugar de todo el vector $\mathbf{V}$, se postula que

$$P\{u(s) = 1 \mid J\} = P\{\mathbf{V}_J = \mathbf{e}_J(s) \mid J\}. \tag{1.2}$$

Para $g \in F_J$, sea $\theta_J(g) = P(\mathbf{V}_J = g \mid J)$ y $\boldsymbol{\theta}_J = \{\theta_J(g)\}_{g \in F_J}$. El vector $\boldsymbol{\theta}_J$ especifica la densidad de $\mathbf{V}_J$. En otras palabras, $\boldsymbol{\theta}_J$ especifica la probabilidad de presencia de la especie en cada $g \in F_J$. De la expresión (1.2) se observa que la probabilidad de presencia de la especie en s es igual a la probabilidad de que el vector $\mathbf{V}_J$ asuma el valor $\mathbf{e}_J(s)$, por lo que se tiene que $\boldsymbol{\theta}_J$ es, de hecho, el parámetro de interés para la pareja J.

Para conjuntar los parámetros de interés correspondientes a las parejas de covariables se introduce la notación $\boldsymbol{\theta}' = (\boldsymbol{\theta}_J)_{J \in G}$. De la misma manera, $\mathbf{C}' = (\mathbf{C}_J)_{J \in G}$ conjunta los vectores de conteos observados en los sitios de presencia para las parejas de covariables. Bajo esta notación, el parámetro de interés es $\boldsymbol{\theta}'$. En seguida se formalizan los conceptos explicados en la introducción de este capítulo, los cuales permiten proponer un modelo, el cual se formula considerando el concepto de parejas de covariables.

Sea $\delta(s)$ la probabilidad de que el nodo s sea examinado para evaluar la presencia de la especie, durante el período de tiempo considerado en el estudio. Este concepto es el denominado *sesgo espacial*. El sesgo espacial induce lo que se ha denominado *sesgo en las covariables*, que para cada pareja $J \in G$ y $g \in F_J$, se denota por $v_J(g)$. Esta cantidad se interpreta como la probabilidad de que por lo menos un nodo que posea el vector $g \in F_J$ para la pareja de covariables J sea físicamente examinado para evaluar la presencia de la especie. El sesgo espacial y el sesgo en las covariables se suponen relacionadas mediante la expresión

$$v_J(g) = 1 - \prod_{\{s \in R : \mathbf{e}_J(s) = g\}} \{1 - \delta(s)\}. \tag{1.3}$$

La expresión (1.3) denota implícitamente que los nodos que posean un valor fijo de covariables para la pareja J son visitados independientemente y propone una forma de obtener el sesgo en las covariables a través de la postulación del sesgo espacial. La manera en la que el sesgo espacial y el sesgo en las covariables se encuentran relacionados puede definirse de otra manera. En esta tesis se adopta la relación (1.3).

Aunque la relación (1.3) puede no ser estrictamente cierta, el supuesto de independencia de las visitas efectuadas a nodos con valores de covariables fijos para una pareja de covariables no parece ser muy restrictivo. Cuando se realizan expediciones de colecta, en general no se examinan intencionalmente nodos que posean un mismo vector de covariables para evaluar la presencia de la especie. Aún cuando el muestreo (y/o posterior visita) de los nodos se realice de manera sistemática, los vectores de covariables correspondientes a los nodos visitados no serán necesariamente visitados de manera sistemática.

Sea ahora $d(s)$ la detectabilidad de la especie en un nodo $s \in R$. La cantidad $d(s)$ se interpreta como la probabilidad de detectar un ejemplar de la especie dado que se encuentra presente en un nodo s visitado. La detectabilidad es una propiedad inherente de la especie, es decir, no depende de los nodos de la retícula sino de consideraciones propias de la especie.

Por esta razón, en lo sucesivo se supone que la detectabilidad es constante sobre la región de estudio y se denota simplemente por d. Con respecto a este supuesto, la expresión (1.2) también postula que la especie tiende a estar presente en aquellos nodos donde el vector $\mathbf{e}_J(s)$ "se parece" a los valores más probables de $\mathbf{V}_J$, por lo que es sensato suponer que la detectabilidad no depende estrictamente de la posición geográfica de s. En caso de que la detectabilidad no se suponga constante sobre la región de estudio, se deberán realizar algunas modificaciones evidentes en las expresiones que se introducen a lo largo de este capítulo. Por ejemplo, será necesario postular una relación análoga a la expresión (1.3) para obtener la correspondiente detectabilidad en las covariables, la cual se interpretará como la probabilidad de detectar la presencia de la especie en el vector de covariables g. La inferencia cuando la detectabilidad no se asume constante sobre la región no se aborda en esta tesis.

En resumen, los elementos relevantes que intervienen en el proceso de detectar la presencia de la especie en un nodo son: la probabilidad de que la especie se encuentre presente en el nodo visitado, la probabilidad de visitar el nodo y la probabilidad de detectar la presencia de la especie, dado que ésta se encuentra presente en el nodo visitado.

Conjuntando las consideraciones realizadas en párrafos anteriores, si $o(s)$ es una variable binaria que toma el valor 1 si se observa un ejemplar de la especie en el nodo s y el valor 0 de otro modo, se obtiene que la probabilidad de registrar la presencia de la especie en s está dada por

$$P\{o(s) = 1 \mid J\} = P\{\mathbf{V}_J = \mathbf{e}_J(s) \mid J\}\, v_J\{\mathbf{e}_J(s)\}\, d. \tag{1.4}$$

Esta igualdad resume la génesis con que ocurren los sitios de presencia. En esta formulación la probabilidad de presencia de la especie, $P\{\mathbf{V}_J = \mathbf{e}_J(s) \mid J\}$, no será identificable sin antes especificar las cantidades $v_J\{\mathbf{e}_J(s)\}$ y d. Se asume por lo tanto que $v_J\{\mathbf{e}_J(s)\}$ se encuentra determinada exactamente mediante la especificación de $\delta(s)$ para cada $s \in R$ y se calcula utilizando la expresión (1.3). Por su parte la cantidad $\delta(s)$ se supone espacialmente uniforme o generada a partir de alguna información adicional con que se cuente.

En el esquema resumido en la igualdad (1.4), lo que es aleatorio y observado es el vector de covariables para la pareja J en un sitio de presencia, es decir, una realización de $\mathbf{V}_J$, y no o_s, el cual se encuentra fijo en el valor 1 como consecuencia del diseño (sólo se cuenta con registros de presencia). Incorporando la parametrización propuesta y usando las expresiones (1.2) y (1.4) se obtiene

$$P\{o(s) = 1 \mid \boldsymbol{\theta}_J, J\} = P\{\mathbf{V}_J = \mathbf{e}_J(s) \mid \boldsymbol{\theta}_J, J\}\, v_J\{\mathbf{e}_J(s)\}\, d. \tag{1.5}$$

Con base en esta igualdad y en las consideraciones realizadas, se han sentado las bases para postular el modelo. Suponga por el momento que la cantidad N, el número de sitios visitados que dan lugar a los n registros de presencia, es conocido. Si los N sitios examinados

se consideran independientes (si $\boldsymbol{\theta}_J$ se supone variable aleatoria puede asumirse que los N sitios son intercambiables, supuesto más débil que el de independencia; ver Bernardo y Smith, 1994, pp. 167–171), cada nodo muestreado puede conceptualizarse como aleatoriamente clasificado en una celda de $R_a R_b + 1$ posibles. Las primeras $R_a R_b$ celdas se encuentran etiquetadas de acuerdo con los posibles valores $g \in F_J$. Un nodo que se rotule con etiqueta g indica que se cuenta con un registro de la presencia de la especie en algún sitio s_i tal que $\mathbf{e}_J(s_i) = g$. La última celda corresponde a las $N - n$ visitas que resultaron en $o(s) = 0$. Esta última celda contiene información del número de nodos en los que no se observó la presencia de la especie, pero no contiene información de los vectores de covariables en los que no se observó la presencia de la especie.

De la expresión (1.5) se deduce que la probabilidad de que un nodo sea clasificado en la celda etiquetada por g es $\theta_J(g)\, v_J(g)\, d$. La conceptualización de que cada una de las N visitas realizadas puede clasificarse en una celda de $R_a R_b + 1$ posibles, corresponde a la formulación estándar de un experimento multinomial, por lo que si $\mathbf{c}_J = \{c_J(g)\}_{g \in F_J}$ es un vector de conteos tal que $\sum_{g \in F_J} c_J(g) \leq N$, entonces es posible postular el modelo

$$P\left(\mathbf{C}_J = \mathbf{c}_J \mid \boldsymbol{\theta}_J, J\right) = \tau_J \left\{1 - \sum_{g \in F_J} \theta_J(g) v_J(g) d\right\}^{N - \sum_{g \in F_J} c_J(g)} \prod_{g \in F_J} \{\theta_J(g) v_J(g) d\}^{c_J(g)} \quad (1.6)$$

para $\mathbf{c}_J$, donde $\tau_J = N! \left\{\prod_{g \in F_J} c_J(g)!\right\}^{-1} \left[\{N - \sum_{g \in F_J} c_J(g)\}!\right]^{-1}$ es la constante de normalización.

Con base en la idea de considerar las parejas de covariables, si $\mathbf{c}$ es un vector de conteos tal que $\sum_{f \in F} c(f) \leq N$, se postula el siguiente modelo mezcla para el vector $\mathbf{C}$:

$$P(\mathbf{C} = \mathbf{c} \mid \boldsymbol{\theta}') = \sum_{J \in G} \pi(J)\, \{k_J(\mathbf{c}_J, N)\}^{-1}\, P(\mathbf{C}_J = \mathbf{c}_J \mid \boldsymbol{\theta}_J, J), \quad (1.7)$$

donde la constante $k_J(\mathbf{c}_J, N)$ es el número de configuraciones diferentes del vector $\mathbf{c}$, con $\sum_{f \in F} c(f) \leq N$, que dan lugar al mismo vector de conteos $\mathbf{c}_J$. La constante $k_J(\mathbf{c}_J, N)$ no depende de $\boldsymbol{\theta}_J$ y puede calcularse por medio de argumentos de tipo combinatorio: dado un vector $\mathbf{c}_J$ tal que $\sum_{g \in F_J} c_J(g) \leq N$, deberá obtenerse el número de vectores $\mathbf{c}$, con $\sum_{f \in F} c(f) \leq N$, que dan lugar al mismo $\mathbf{c}_J$.

Por su parte, el conjunto $\{\pi(J),\ J \in G\}$, con $\sum_{J \in G} \pi(J) = 1$, representa la función de masa de probabilidad postulada sobre G. Los valores $\pi(J)$ se obtendrán con base en conocimiento del experto, como se explica en la Sección 3.1.2. En la siguiente proposición se demuestra que el modelo (1.7) representa una legítima distribución de probabilidad para los vectores $\mathbf{C}$ tales que $\sum_{f \in F} C(f) \leq N$.

Proposición 1 *El modelo mezcla dado por la expresión (1.7) define una distribución de probabilidad para los vectores* $\mathbf{c}$ *tales que* $\sum_{f \in F} c(f) \leq N$, *es decir, se satisfacen las siguientes*

condiciones:

1. $P(\mathbf{C} = \mathbf{c} \mid \boldsymbol{\theta}') \geq 0$, *para toda* $\mathbf{c}$ *tal que* $\sum_{f \in F} c(f) \leq N$.
2. $\sum_{\{\mathbf{c}:\, \sum_{f \in F} c(f) \leq N\}} P(\mathbf{C} = \mathbf{c} \mid \boldsymbol{\theta}') = 1$.

Demostración. Ya que cada $P(\mathbf{C}_J = \mathbf{c}_J \mid \boldsymbol{\theta}_J, J) \geq 0$, por ser P el modelo multinomial, y también $k_J(\mathbf{c}_J, N) > 0$, pues por lo menos se cuenta con un $\mathbf{c}$ que produce el vector $\mathbf{c}_J$, la primera condición se cumple.

Para probar la segunda condición basta seguir la secuencia de igualdades que se presenta a continuación, que se obtiene de manipular de manera adecuada los sumandos:

$$\begin{aligned}
\sum_{\{\mathbf{c}:\, \sum_{f \in F} c(f) \leq N\}} P(\mathbf{C} &= \mathbf{c} \mid \boldsymbol{\theta}') = \sum_{\{\mathbf{c}:\, \sum_{f \in F} c(f) \leq N\}} \sum_{J \in G} \pi(J)\, \{k_J(\mathbf{c}_J, N)\}^{-1}\, P\left(\mathbf{C}_J = \mathbf{c}_J \mid \boldsymbol{\theta}_J, J\right) \\
&= \sum_{J \in G} \pi(J) \sum_{\{\mathbf{m}:\, \sum_{f \in F} m(f) \leq N\}} \sum_{\{\mathbf{c}:\, \mathbf{c}_J = \mathbf{m}\}} \{k_J(\mathbf{c}_J, N)\}^{-1}\, P(\mathbf{C}_J = \mathbf{c}_J \mid \boldsymbol{\theta}_J, J) \\
&= \sum_{J \in G} \pi(J) \sum_{\{\mathbf{m}:\, \sum_{f \in F} m(f) \leq N\}} \sum_{\{\mathbf{c}:\, \mathbf{c}_J = \mathbf{m}\}} \{k_J(\mathbf{m}, N)\}^{-1}\, P\left(\mathbf{C}_J = \mathbf{m} \mid \boldsymbol{\theta}_J, J\right) \\
&= \sum_{J \in G} \pi(J) \sum_{\{\mathbf{m}:\, \sum_{f \in F} m(f) \leq N\}} k_J(\mathbf{m}, N)\, \{k_J(\mathbf{m}, N)\}^{-1}\, P\left(\mathbf{C}_J = \mathbf{m} \mid \boldsymbol{\theta}_J, J\right) \\
&= \sum_{J \in G} \pi(J) \sum_{\{\mathbf{m}:\, \sum_{f \in F} m(f) \leq N\}} P\left(\mathbf{C}_J = \mathbf{m} \mid \boldsymbol{\theta}_J, J\right) \\
&= \sum_{J \in G} \pi(J) = 1.
\end{aligned}$$

■

Una interpretación del modelo (1.7) es probabilística y se basa en la noción de mezcla: se conceptualiza que un ejemplar de la especie selecciona una pareja de covariables J del conjunto G, con probabilidad $\pi(J) \geq 0$. Condicionada a considerar solamente la pareja J que seleccionó, la probabilidad de presencia de la especie en un sitio s está dada por $\theta_J(\mathbf{e}_J(s))$. Este esquema induce una estructura de conteos multinomiales para el vector $\mathbf{C}_J$, lo cual a su vez, induce una estructura de conteos para $\mathbf{C}$. El término $\{k_J(\mathbf{c}_J, N)\}^{-1}$ en la expresión (1.7) implica que se asigna la misma probabilidad de ocurrencia a todos los posibles vectores $\mathbf{C}$, con $\sum_{f \in F} C(f) \leq N$, que producen el mismo vector $\mathbf{C}_J$.

Por su parte, la función de masa de probabilidad $\{\pi(J),\ J \in G\}$ puede conceptualizarse como aquella que resume la idiosincracia de la especie con respecto a la "apreciación" que realiza de un sitio, de acuerdo con los valores de las parejas de covariables presentes en él.

En la postulación del modelo se asume que la cantidad N es conocida. Sin embargo, no es una regla general que se conserve una lista completa de sitios visitados a lo largo del período de tiempo considerado, especialmente cuando se toman en cuenta registros históricos, por lo

que la cantidad N debe considerarse, en principio, desconocida. Sin embargo, para una pareja J fija y dada la expresión (1.5), se espera que el número de sitios de presencia registrados en nodos con vector de covariables $g \in F_J$, es decir la cantidad $C_J(g)$, se encuentre dada por la expresión $C_J(g) = N\,\theta_J(g)\,v_J(g)\,d$, para N grande, de donde se obtiene por despeje que $N\,\theta_J(g) = C_J(g)\,v_J^{-1}(g)\,d^{-1}$. Sumando ambos lados de esta igualdad sobre todos los vectores de covariables posibles para la pareja J, y ya que $\sum_{g\in F_J}\theta_J(g) = 1$, se obtiene que $N \approx N_J = \left\lceil \sum_{g\in F_J} C_J(g)\,v_J^{-1}(g)\,d^{-1} \right\rceil$ para todo J. Una forma simple de proceder es postular $N = |G|^{-1}\sum_{J\in G} N_J$ como aproximación de trabajo en el modelo (1.7), en lugar de considerar N como un parámetro desconocido de ruido, en cuyo caso sería necesario postular una distribución de probabilidad para N.

1.2 Inferencia: Probabilidad Predictiva

En esta sección se propone la manera de hacer inferencias con respecto a la probabilidad de presencia de una especie en cada nodo. Como se discute en la introducción, en la práctica es posible que un experto aporte alguna información con respecto a las regiones de establecimiento y/o no establecimiento de especies de interés sobre la región bajo estudio. Con el fin de involucrar esta información en el proceso de inferencia, se utiliza el enfoque de la estadística Bayesiana.

Siguiendo el procedimiento usual para distribuciones mezcla desde el enfoque Bayesiano, la distribución *a priori* para el parámetro de interés, $\boldsymbol{\theta}' = (\boldsymbol{\theta}_J)_{J\in G}$, se postula como el producto de distribuciones *a priori* $f(\boldsymbol{\theta}_J)$ asignadas a los parámetros $\boldsymbol{\theta}_J$, es decir $f(\boldsymbol{\theta}') = \prod_{J\in G} f(\boldsymbol{\theta}_J)$. Para postular la distribución *a priori* de esta manera se asume que los vectores $\boldsymbol{\theta}_J$, $J \in G$, son independientes. Aunque este supuesto es discutible, la postulación de la distribución *a priori* de esta manera permite realizar las operaciones de manera sencilla. Un argumento que permite tolerar la imposición de ese supuesto es que la distribución posterior que resulta es una mezcla de probabilidades posteriores y no un producto de probabilidades posteriores. Así, suponer independencia no implica supuesto alguno de este tipo para la distribución posterior que resulta.

Ya que el espacio de covariables se ha conceptualizado como el conjunto de todas las parejas de covariables, la ley de probabilidad total produce

$$P\{u(s) = 1 \mid \mathbf{C}'\} = \sum_{J\in G} P\{u(s) = 1 \mid \mathbf{C}', J\}\,\pi(J \mid \mathbf{C}'),$$

donde $P\{u(s) = 1 \mid \mathbf{C}'\}$ representa la probabilidad de presencia de la especie en el nodo s. Por su parte, la cantidad $P\{u(s) = 1 \mid \mathbf{C}', J\}$ representa la probabilidad predictiva de presencia de la especie en el nodo s cuando se considera solamente la pareja J para hacer inferencias. La cantidad $\pi(J \mid \mathbf{C}')$ puede interpretarse como la probabilidad (posterior) que la

especie asigna a la pareja J cuando decide hacerse presente en la región bajo estudio. Un valor alto de $\pi(J \mid \mathbf{C}')$, comparado con los correspondientes valores obtenidos para otras parejas, indica que la especie considera cuidadosamente los valores de esta pareja de covariables al decidir hacerse presente en la región. Por otro lado, un valor bajo de $\pi(J \mid \mathbf{C}')$, comparado con los correspondientes valores de otras parejas de covariables, indica que la especie no considera importante los valores de esta pareja en s al decidir hacerse presente en la región. La cantidad $\pi(J \mid \mathbf{C}')$ aporta información acerca de las parejas de covariables preferidas por la especie.

En la siguiente proposición se presenta un resultado que permitirá calcular las cantidades de interés a través de un valor esperado.

Proposición 2 *Para un nodo arbitrario $s \in R$ se cumple*

$$P\{u(s) = 1 \mid \mathbf{C}', J\} = E[\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}', J],$$

donde $\mathbf{C}' = (\mathbf{C}_J)_{J\in G}$ se define como antes.

Demostración. Utilizando la igualdad (1.2) en la definición de probabilidad predictiva, se tiene

$$\begin{aligned}
P\{u(s) &= 1 \mid \mathbf{C}', J\} = P\{u(s) = 1 \mid \mathbf{C}_J, J\} \\
&= \int P\{V = \mathbf{e}_J(s) \mid \mathbf{C}_J, J\}\, f(\boldsymbol{\theta}_J \mid \mathbf{C}_J, J) d\boldsymbol{\theta}_J \\
&= \int \theta(\mathbf{e}_J(s)) f(\boldsymbol{\theta}_J \mid \mathbf{C}_J, J) d\boldsymbol{\theta}_J \\
&= \int \theta(\mathbf{e}_J(s)) f(\theta(\mathbf{e}_J(s)) \mid \mathbf{C}_J, J) d\theta(\mathbf{e}_J(s)) \\
&= E[\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}_J, J] \\
&= E[\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}', J].
\end{aligned}$$

■

De este hecho se obtiene por sustitución que la probabilidad de presencia de la especie en el nodo s está dada por

$$P\{u(s) = 1 \mid \mathbf{C}'\} = \sum_{J\in G} E[\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}', J]\,\pi(J \mid \mathbf{C}'). \tag{1.8}$$

La cantidad $E[\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}', J]$ se calcula con base en la distribución marginal posterior $f(\theta(e_J(s)) \mid \mathbf{C}_J, J)$ de $\theta(e_J(s))$. Como antes, sea $P(\mathbf{C}_J \mid \boldsymbol{\theta}_J, J)$ el modelo multinomial para la pareja J dado por la expresión (1.6). Ya que el interés es hacer inferencia sobre vectores de probabilidades, se postula una distribución Dirichlet como *a priori* para cada $\boldsymbol{\theta}_J$. Esta

distribución posee la expresión

$$f(\boldsymbol{\theta}_J) = \frac{\Gamma(\alpha_J)}{\prod\limits_{g\in F_J}\Gamma\{\alpha_J(g)\}}\prod_{g\in F_J}\theta_J(g)^{\alpha_J(g)-1}, \tag{1.9}$$

donde $\alpha_J = \sum_{g\in F_J}\alpha_J(g)$, $\alpha_J(g) > 0$. El parámetro de esta distribución es el vector $\boldsymbol{\alpha}_J = (\alpha_J(g))_{g\in F_J}$ y el conocimiento del experto se utiliza para postular (elicitar) valores para estos parámetros. La elicitación de los parámetros de la distribución *a priori* $f(\boldsymbol{\theta}_J)$ y la postulación de valores $\pi(J)$ se presenta en la Sección 3.1.2. La distribución posterior conjunta que resulta para cada J al aplicar el Teorema de Bayes con el modelo multinomial (1.6) y la distribución *a priori* Dirichlet (1.9) es

$$\begin{aligned} f(\boldsymbol{\theta}_J, J \mid \mathbf{C}') \;=\; & \frac{\pi(J)N!\Gamma(\alpha_J)}{(N-n)!\prod\limits_{g\in F_J}c_J(g)!\Gamma\{\alpha_J(g)\}}\left\{1-\sum_{g\in F_J}\theta_J(g)\nu_J(g)d\right\}^{N-n}\times \\ & \prod_{g\in F_J}\theta_J(g)^{c_J(g)+\alpha_J(g)-1}\nu_J(g)^{c_J(g)}, \end{aligned} \tag{1.10}$$

de la cual se deberá obtener la cantidad $E[\theta_J\{g\} \mid \mathbf{C}', J]$ para cada $g \in F_J$, y la cantidad $\pi(J \mid \mathbf{C}')$ para cada $J \in G$. Sin embargo, la distribución marginal $f(\boldsymbol{\theta}_J \mid \mathbf{C}', J)$ no corresponde a alguna expresión conocida, por lo que es necesario recurrir a algoritmos numéricos (MCMC) para estimar las cantidades $E[\theta_J\{g\} \mid \mathbf{C}', J]$. Por su parte, de la distribución marginal $f(J \mid \mathbf{C}', \boldsymbol{\theta}_J)$ se estima la cantidad $\pi(J \mid \mathbf{C}')$. La implementación del algoritmo MCMC se presenta en la Sección 1.4.

Aunque no es difícil implementar un algoritmo numérico, enseguida se propone una forma de proceder que permite obtener de manera aproximada las cantidades involucradas en la expresión (1.8) mediante fórmulas cerradas. Esta forma de proceder surge a partir de la génesis observada con que ocurren los sitios de presencia. En caso de no existir el sesgo en las covariables y postulando que $d = 1$, la cantidad $X_J(g) = N\theta_J(g)$ representa el número de presencias (conteos reales) que se espera observar para el vector de covariables $g \in F_J$ en las N visitas. Por otro lado, de la expresión (1.5) se deduce que el número de sitios de presencia esperados está dado por la expresión $C_J(g) = N\,\theta_J(g)\,v_J(g)\,d$, de donde se obtiene por despeje que $N\,\theta_J(g) = C_J(g)\,[v_J(g)\,d]^{-1}$. Con base en estas consideraciones, se obtiene que la cantidad $X_J^*(g) = C_J(g)\,[v_J(g)\,d]^{-1}$ aproxima al conteo multinomial real $X_J(g)$ correspondiente al vector de covariables $g \in F_J$. Denotando por $\mathbf{X}_J^* = \{X_J^*(g)\}_{g\in F_J}$ al vector que conjunta las aproximaciones a los conteos esperados, se propone utilizar una distribución Dirichlet con parámetros $\mathbf{X}_J^* + \boldsymbol{\alpha}_J$ como distribución posterior para $\boldsymbol{\theta}_J$, la cual

esta dada por la expresión

$$f(\boldsymbol{\theta}_J \mid \mathbf{X}_J^*, J) = \frac{\Gamma(N+\alpha_J)}{\prod\limits_{g\in F_J}\Gamma\{X_J^*(g)+\alpha_J(g)\}}\prod_{g\in F_J}\theta_J(g)^{X_J^*(g)+\alpha_J(g)-1}. \tag{1.11}$$

En esta expresión, la cantidad N es la aproximación postulada al final de la Sección 1.1. La consideración rigurosa de un modelo alternativo de este tipo para las $X_J(g)$'s involucra la identificación de una constante de normalización desconocida, la cual depende del parámetro $\boldsymbol{\theta}_J$. El hecho relevante de la distribución Dirichlet que se propone como aproximación para la distribución posterior exacta para cada J se encuentra en que, examinando las densidades $f(\boldsymbol{\theta}_J \mid \mathbf{C}_J, J)$ y $f(\boldsymbol{\theta}_J \mid \mathbf{X}_J^*, J)$, se observa numéricamente que los correspondientes valores esperados, $E[\theta_J(g) \mid \mathbf{C}_J, J]$ y $E[\theta_J(g) \mid \mathbf{X}_J^*, J]$, son virtualmente iguales para todo $g \in F_J$ y $J \in G$. Este hecho se observa en los mapas de probabilidad de presencia obtenidos utilizando MCMC y la aproximación propuesta, los cuales resultan ser idénticos (Sección 1.5). Utilizando la aproximación Dirichlet dada por (1.11) se obtiene que

$$E[\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{X}_J^*, J] \approx \frac{X_J^*(\mathbf{e}_J(s)) + \alpha_J(\mathbf{e}_J(s))}{N+\alpha_J},$$

donde $X_J^*(\mathbf{e}_J(s)) = C_J(\mathbf{e}_J(s))[\nu_J(\mathbf{e}_J(s))\,d]^{-1}$. Por su parte, la cantidad $\pi(J \mid \mathbf{C}')$ se obtiene integrando la expresión (1.10) con respecto a $\boldsymbol{\theta}_J$, con lo que se obtiene

$$\pi(J \mid \mathbf{C}') \propto \pi(J)\int_{\theta_J} P(\mathbf{C}_J \mid \boldsymbol{\theta}_J, J)\,f(\boldsymbol{\theta}_J)\,d\boldsymbol{\theta}_J.$$

Ya que se cuenta con una distribución $f(\boldsymbol{\theta}_J \mid \mathbf{X}_J^*, J)$ que aproxima a $f(\boldsymbol{\theta}_J \mid \mathbf{C}_J, J)$, por el Teorema de Bayes se tiene que

$$\int_{\theta_J} P(\mathbf{C}_J \mid \boldsymbol{\theta}_J, J)\,f(\boldsymbol{\theta}_J)\,d\boldsymbol{\theta}_J \approx \frac{f(\mathbf{C}_J \mid \boldsymbol{\theta}_J^0, J)\,f(\boldsymbol{\theta}_J^0)}{f(\boldsymbol{\theta}_J^0 \mid \mathbf{X}_J^*)}$$

para algún vector $\boldsymbol{\theta}_J^0 = \{\theta_J^0(g)\}_{g\in F_J}$ fijo. Sustituyendo las expresiones con que se cuenta, se obtiene que

$$\begin{aligned} \pi(J \mid \mathbf{C}') \;\propto\; & \frac{\pi(J)\Gamma(\alpha_J)}{\Gamma(N+\alpha_J)}\prod_{g\in F_J}\frac{\Gamma\{X_J^*(g)+\alpha_J(g)\}\,v_J(g)^{c_J(g)}}{\Gamma\{\alpha_J(g)\}}\times \\ & \left\{1-\sum_{g\in F_J}\theta_J^0(g)v_J(g)\right\}^{N-n}\prod_{g\in F_J}\{\theta_J^0(g)\}^{c_J(g)-X_J^*(g)}. \end{aligned} \tag{1.12}$$

En la práctica, bastará calcular el lado derecho de (1.12) para cada $J \in G$ y normalizar las cantidades que resultan para obtener una aproximación de la cantidad $\pi(J \mid \mathbf{C}')$.

Una ventaja de la aproximación propuesta es la tratabilidad matemática que posee, pues corresponde a una distribución Dirichlet. La distribución exacta debe necesariamente tratarse por medio de métodos numéricos para obtener las cantidades de interés.

Con respecto al vector $\boldsymbol{\theta}_J^0$ que se utiliza para calcular la aproximación de $\pi(J \mid \mathbf{C}')$, $J \in G$, se considera $\theta_J^0(g) = \{X_J^*(g) + \alpha_J(g)\}\{N^* + \alpha_J\}^{-1}$ para cada $g \in F_J$. La cantidad $\theta_J^0(g)$ así definida corresponde al valor esperado de $\theta_J(g)$ obtenido de utilizar la distribución Dirichlet que aproxima a la distribución posterior exacta. Ya que la distribución propuesta como aproximación es buena, cada $\theta_J^0(g)$ es a su vez una buena aproximación para el correspondiente $\theta_J(g)$, por lo que el vector $\boldsymbol{\theta}_J^0$ es una buena aproximación del verdadero $\boldsymbol{\theta}_J$.

Una vez que se calcula la probabilidad predictiva de presencia para cada $s \in R$, se procede a observar los resultados en un mapa, como se explica enseguida. Se propone asignar un nivel de gris o de color a cada valor de probabilidad. Para esto se considera una partición arbitraria $I_j = ((j-1)/r, j/r]$, $1 \leq j \leq r$, del intervalo $[0,1]$ y una escala de gris o de color asociada a dicha partición. Cada nodo se despliega con el color asociado al intervalo I_{j_s} para el que $P\{u(s) = 1 \mid \mathbf{C}'\} \in I_{j_s}$. En esta tesis se utiliza $r = 10$, es decir, el intervalo $[0,1]$ se divide en 10 subintervalos de longitud .1, y se utiliza una escala de gris para desplegar los mapas de probabilidad de presencia.

1.3 Certidumbre para la Probabilidad de Presencia

En los problemas de estimación no es solamente relevante obtener un estimador de la cantidad de interés, sino que también es relevante dotar al estimador de alguna medida de la certidumbre (precisión). Por ejemplo, un intervalo de probabilidad del 95% calificado como angosto (bajo algún criterio) se asocia con la idea de poca incertidumbre, mientras que un intervalo de probabilidad calificado como amplio se asocia con la idea de mayor incertidumbre.

Con respecto al problema que se aborda en este capítulo, los métodos que se utilizan actualmente para obtener mapas de habitación potencial para el establecimiento de especies no proponen alguna forma de evaluar la precisión del potencial que se obtiene. En esta sección se propone una forma de proporcionar una medida de precisión para la probabilidad de presencia dada por la expresión (1.8). Al evaluar la precisión en todos los nodos de la retícula se obtiene un mapa que se denomina *mapa de certidumbre.* Para cada nodo de la retícula se contará con la probabilidad predictiva de presencia de la especie y con una medida que indica la precisión de dicho resultado. Esto permitirá poseer mayor información con respecto a las áreas de establecimiento de las especies.

Para evaluar la certidumbre del nodo s, se registra el intervalo I_{j_s} de la partición para el que ocurre que $P\{u(s) = 1 \mid \mathbf{C}'\} \in I_{j_s}$. Con el intervalo I_{j_s} localizado, se propone evaluar la

integral

$$I_J(s) = \int_{I_{j_s}} f(\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}', J)\, d\theta_J\{\mathbf{e}_J(s)\},$$

para cada $J \in G$, donde $f(\theta_J\{\mathbf{e}_J(s)\} \mid \mathbf{C}', J)$ es la distribución marginal de $\theta_J\{\mathbf{e}_J(s)\}$. La cantidad $I_J(s)$ se interpreta como la probabilidad posterior de que el valor esperado de $\theta_J\{\mathbf{e}_J(s)\}$ se encuentre contenido en el intervalo I_{j_s}. Motivado por la expresión (1.8), la cantidad

$$I(s) = \sum_{J \in G} I_J(s)\, \pi(J \mid \mathbf{C}')$$

proporciona un nivel de certidumbre para la probabilidad de presencia $P\{u(s) = 1 \mid \mathbf{C}'\}$ obtenida para el nodo s, donde $\pi(J \mid \mathbf{C}')$ se interpreta como antes.

Para estimar la cantidad $I(s)$ puede utilizarse la distribución posterior exacta, con lo cual deberá recurrirse a un algoritmo MCMC (Sección 1.4), o bien, utilizar la aproximación Dirichlet propuesta. En caso de optar por la aproximación Dirichlet, la cantidad $I_J(s)$ se encontrará evaluando la integral de la densidad marginal $f(\theta_J(e_J(s)) \mid \mathbf{X}_J^*, J)$ sobre el intervalo I_{j_s}. Ya que la distribución que se usa como aproximación de la distribución posterior exacta corresponde a una distribución Dirichlet, la distribución marginal $f(\theta_J(e_J(s)) \mid \mathbf{X}_J^*, J)$ es una distribución beta con parámetros $(X_J^*(g) + \alpha_J(g), N + \alpha_J - X_J^*(g) - \alpha_J(g))$, donde N es la aproximación postulada al final de la sección 1.1. De esta forma, se tendrá

$$I_J(s) = \kappa_J \int_{I_{j_s}} \theta_J(g)^{X_J^*(g)+\alpha_J(g)} \{1 - \theta_J(g)\}^{N+\alpha_J - X_J^*(g) - \alpha_J(g)}\, d\theta_J(g),$$

donde $\kappa_J = \Gamma[N + \alpha_J]\,[\Gamma\{X_J^*(g) + \alpha_J(g)\}\,\Gamma\{N + \alpha_J - X_J^*(g) - \alpha_J(g)\}]^{-1}$ es la constante de normalización.

La cantidad $I(s)$ depende tanto de la distribución marginal posterior de cada J, como de la partición del intervalo $[0,1]$ utilizada para desplegar el mapa de probabilidades de presencia. Los resultados de certidumbre obtenidos para todos los nodos pueden desplegarse utilizando la misma idea que el mapa de probabilidades de presencia, es decir, utilizando alguna escala de gris o de color. En esta tesis se considera una escala de gris para desplegar los mapas que se obtengan.

1.4 MCMC

Debido a que la distribución posterior que resulta para cada pareja de covariables no corresponde a alguna distribución conocida, las cantidades de interés deberán obtenerse por medio de algún algoritmo numérico. Con estos algoritmos se obtendrán valores (muestras)

provenientes de la distribución posterior, los cuales se utilizan para obtener estimaciones de las cantidades de interés, en nuestro caso, de las cantidades $E[\theta_J\{g\} \mid \mathbf{C}', J]$ para toda $g \in F_J$ y de las cantidades $\pi(J \mid \mathbf{C}')$ para toda $J \in G$.

Para simular valores de la distribución posterior se implementó el algoritmo conocido como Metropolis-Hastings (ver Robert y Casella, 1999). Para cada pareja $J \in G$, el modelo multinomial $P(\mathbf{C}_J \mid \boldsymbol{\theta}_J, J)$ dado por la expresión (1.6) y la distribución *a priori* Dirichlet dada por la expresión (1.9) producen la distribución posterior conjunta $f(\boldsymbol{\theta}_J, J \mid \mathbf{C}_J)$ dada por la expresión (1.10).

Para obtener muestras de parejas J a partir de la distribución marginal $f(J \mid \boldsymbol{\theta}_J, \mathbf{C}_J)$, $J \in G$, se procede como sigue. Dado el conjunto de vectores de parámetros $(\boldsymbol{\theta}_J^{(t)})_{J \in G}$ y la pareja $J = J^{(t)}$ en la iteración t, en la iteración $t+1$ se selecciona al azar una pareja candidata J' del conjunto G, con distribución uniforme, por lo que cada pareja posee probabilidad $1/|G|$ de ser seleccionada. El valor de J en la iteración $t+1$ será $J^{(t+1)} = J'$ con probabilidad $\min\left\{1, \rho_1(J^{(t)}, J')\right\}$ y será $J^{(t+1)} = J^{(t)}$ con probabilidad $1 - \min\left\{1, \rho_1(J^{(t)}, J')\right\}$, donde

$$\rho_1\left(J^{(t)}, J'\right) = \frac{f(J'J \mid \boldsymbol{\theta}_{J'}, \mathbf{C}_{J'})}{f(J^{(t)} \mid \boldsymbol{\theta}_{J^{(t)}}, \mathbf{C}_{J^{(t)}})}.$$

sustituyendo y simplificando se obtiene

$$\rho_1\left(J^{(t)}, J'\right) = \frac{\left\{\prod_{g \in F_{J^{(t)}}} c_{J^{(t)}}(g)!\right\} \prod_{g \in F_{J'}} \Gamma\{\alpha_{J'}(g)\} \left\{1 - \sum_{g \in F_{J'}} \theta_{J'}(g)\nu_{J'}(g)\right\}^{N-n}}{\left\{\prod_{g \in F_{J'}} c_{J'}(g)!\right\} \prod_{g \in F_{J^{(t)}}} \Gamma\{\alpha_{J^{(t)}}(g)\} \left\{1 - \sum_{g \in F_{J^{(t)}}} \theta_{J^{(t)}}(g)\nu_{J^{(t)}}(g)\right\}^{N-n}} \times \frac{\pi(J')\Gamma(\alpha_{J'}) \prod_{g \in F_{J'}} \theta_{J'}(g)^{c_{J'}(g)+\alpha_{J'}(g)-1}\nu_{J'}(g)^{c_{J'}(g)}}{\pi(J^{(t)})\Gamma(\alpha_{J^{(t)}}) \prod_{g \in F_{J^{(t)}}} \theta_{J^{(t)}}(g)^{c_{J^{(t)}}(g)+\alpha_{J^{(t)}}(g)-1}\nu_{J^{(t)}}(g)^{c_{J^{(t)}}(g)}}.$$

En cada iteración deberá registrarse la pareja de covariables seleccionada. De esta manera, al finalizar el proceso iterativo se tendrá el número de veces que fue seleccionada cada una de las parejas. Para obtener las cantidades de interés, basta dividir dichas cantidades entre el número total de iteraciones en las que se seleccionó una muestra de J, con lo que se obtiene una estimación de la cantidad $\pi(J \mid \mathbf{C}')$, $J \in G$.

Para obtener muestras del vector $\boldsymbol{\theta}_J$ para un $J \in G$ fijo a partir de la distribución marginal $f(\boldsymbol{\theta}_J \mid \mathbf{C}_J, J)$, se procede como sigue. Dada una $J = J^{(t)}$ fija en la iteración t, en la iteración $t+1$ se selecciona un vector $\boldsymbol{\theta}'_J$ de la distribución de propuesta Dirichlet con parámetros $\mathbf{X}^*_J + \boldsymbol{\alpha}_J$. La distribución de propuesta que se utiliza es precisamente la distribución Dirichlet que se usa como aproximación de la distribución posterior exacta. Una vez que se selecciona un vector $\boldsymbol{\theta}'_J$, se define $\boldsymbol{\theta}_J^{(t+1)} = \boldsymbol{\theta}'_J$ con probabilidad $\min\left\{1, \rho_2(\boldsymbol{\theta}_J^{(t)}, \boldsymbol{\theta}'_J)\right\}$ y se define $\boldsymbol{\theta}_J^{(t+1)} =$

$\boldsymbol{\theta}_J^{(t)}$ con probabilidad $1 - \min\left\{1, \rho_2(\boldsymbol{\theta}_J^{(t)}, \boldsymbol{\theta}'_J)\right\}$, donde

$$\rho_2(\boldsymbol{\theta}_J^{(t)}, \boldsymbol{\theta}'_J) = \frac{f(\boldsymbol{\theta}'_J \mid \mathbf{C}_J, J)}{f(\boldsymbol{\theta}_J^{(t)} \mid \mathbf{C}_J, J)} \frac{f(\boldsymbol{\theta}_J^{(t)} \mid \mathbf{X}^*_J, J)}{f(\boldsymbol{\theta}'_J \mid \mathbf{X}^*_J, J)}.$$

Sustituyendo y simplificando se obtiene

$$\rho_2(\boldsymbol{\theta}_J^{(t)}, \boldsymbol{\theta}'_J) = \left\{\frac{1 - \sum_{g \in F_J} \theta'_J(g)\nu_J(g)}{1 - \sum_{g \in F_J} \theta_J^{(t)}(g)\nu_J(g)}\right\}^{N-n} \prod_{g \in F_J} \left\{\frac{\theta_J^{(t)}(g)}{\theta'_J(g)}\right\}^{X_J(g)-c_J(g)}.$$

En cada iteración se almacena el vector $\boldsymbol{\theta}_J^{(t)}$ obtenido, para cada $J \in G$. Al final del proceso iterativo bastará obtener el promedio de los elementos $\theta_J(g)$ de los vectores $\boldsymbol{\theta}_J^{(t)}$, es decir, se define

$$E[\theta_J(g) \mid \mathbf{C}', J] \approx \frac{\sum_{t=1}^{T_J} \theta_J^{(t)}(g)}{T_J},$$

donde la cantidad T_J denota el número de veces que se seleccionó una muestra del vector $\boldsymbol{\theta}_J$ en el proceso iterativo.

Los esquemas explicados para obtener muestras de parejas J y muestras de vectores $\boldsymbol{\theta}_J$, se implementan y ejecutan en un mismo programa, en el que se generó un valor J de G con probabilidad .5 y un vector $\boldsymbol{\theta}_J$ con probabilidad .5. Se utilizó el valor .5 con el objetivo de seleccionar en la misma proporción muestras de parejas de covariables de G y muestras de vectores de parámetros.

Para obtener la certidumbre de la probabilidad de presencia utilizando MCMC se procede como sigue. En cada iteración deberá registrarse el intervalo I_j de la partición del intervalo $[0,1]$ al que pertenece cada $\theta_J(g)$, $g \in F_J$, del vector $\boldsymbol{\theta}_J$ generado. El número de intervalos que se considere en esta partición será el mismo que el considerado para desplegar el mapa de probabilidades de presencia. Al término del proceso iterativo se tendrá el registro del número de veces que cada valor $\theta_J(g)$ simulado perteneció a cada intervalo de la partición considerada, para cada $J \in G$. Finalmente se procede a calcular las proporciones correspondientes, dividiendo los conteos obtenidos entre T_J. Así, si I_{j_s} denota el intervalo de la partición tal que $P\{u(s) = 1 \mid \mathbf{C}'\} \in I_{j_s}$, para el nodo s y la pareja J, la cantidad $I_J(s)$ se obtiene mediante la expresión

$$I_J(s) \approx \frac{\text{Veces que } \theta_J(\mathbf{e}_J(s)) \in I_{j_s}}{T_J}.$$

Cuando se implementa un algoritmo MCMC debe realizarse cierto monitoreo del proceso iterativo. Este monitoreo incluye (1) considerar distintos valores iniciales y (2) observar

la convergencia de la cantidad de interés, en nuestro caso, $E[\theta_J(g)]$ para cada $g \in F_J$. Al implementar el algoritmo MCMC propuesto se verificaron los puntos (1) y (2) en cada caso. Se observó que el valor inicial no es determinante en el proceso iterativo, y en cada caso, el valor $E[\theta_J(g)]$ se estabilizó después de un número suficiente de iteraciones.

1.5 Estudio de Simulación

En esta sección se propone un estudio de simulación con dos objetivos. El primero es estudiar el funcionamiento de la metodología que se propone cuando se implementa bajo diferentes condiciones. El segundo surge por el hecho de que se cuenta con diversos métodos para abordar el problema descrito. Debido a esto, el interés inmediato que se presenta es comparar los mapas que se obtienen. Cada método produce resultados que se encuentran medidos en unidades diferentes, por lo que no es posible realizar una comparación de manera directa. Una forma de solventar este problema es comparar los resultados de manera cualitativa. Sin embargo, el problema que surge en este contexto radica en que *no se conoce el mapa real de la distribución de la especie*, por lo que no se cuenta con un mapa para comparar las soluciones que se obtengan. En esta sección se propone una manera de generar un mapa de probabilidades de presencia real, mediante el cual (1) se generan los sitios de presencia utilizando la génesis identificada con que ocurren los sitios de presencia y (2) se obtiene un mapa de referencia con el cual comparar los mapas obtenidos con cada método que se utilice.

La región de estudio que se considera es la península de Yucatán, la cual se asume cubierta por una retícula regular que consta de 761 nodos, con separación de .125 grados. Cada nodo de la retícula representa un cuadrado de aproximadamente $12\,km$ por lado. Para generar el mapa de probabilidades de presencia de una especie ficticia se utiliza la siguiente idea. Ya que se asume que la especie se establece en un sitio con base en los valores de las covariables, es sensato postular que cada especie posee un vector de covariables "óptimo" $\boldsymbol{\mu} = \{\mu_1, ..., \mu_M\}$, en el que encuentra las condiciones ideales para su establecimiento. Para generar el mapa de probabilidades de establecimiento (o sea, de presencia) de una especie ficticia se postula que a medida que los valores de $\mathbf{e}(s)$ difieren de los correspondientes valores de $\boldsymbol{\mu}$, la probabilidad de presencia de la especie en el nodo s decrece de acuerdo con

$$P\{u(s) = 1\} = e^{-\frac{1}{2}\{\mu - \mathbf{e}(s)\}^t \mathbf{A}\{\mu - \mathbf{e}(s)\}}. \tag{1.13}$$

La expresión (1.13) es una forma de prescribir la manera en la que la probabilidad de presencia de la especie depende de $\mathbf{e}(s)$, a la vez que permite incorporar la noción de que dicha probabilidad decrece conforme el vector de covariables del nodo difiere de un valor idealizado. La matriz $\mathbf{A} = (t_{h,l})_{1 \leq h,l \leq M}$, permite incorporar cierta estructura con respecto a las interacciones de los componentes de $\mathbf{e}(s)$. Al especificar diversos valores de $\boldsymbol{\mu}$ y $\mathbf{A}$ se generan los mapas de probabilidad de presencia para especies ficticias. La función (1.13) puede

considerarse como la "realidad", y genera el mapa con el que se comparan los resultados que se obtengan al aplicar las metodologías existentes, incluyendo la que se propone en este capítulo.

Para generar el mapa de establecimiento potencial usando cada método, en este estudio de simulación se consideran los valores de tres covariables en cada nodo de la retícula: temperatura media, que consta de 5 categorías, precipitación media, que consta de 10 categorías, y tipo de vegetación, que consta de 11 categorías. Los valores de esas covariables corresponden a mediciones reales sobre cada nodo de la península de Yucatán.

En esta sección se ilustra el funcionamiento de la metodología propuesta en este capítulo, utilizando el vector ideal $\boldsymbol{\mu} = \{2, 4, 3.5\}$, con la matriz $\mathbf{A}$ determinada por los valores $t_{1,1} = t_{2,2} = t_{3,3} = 1$, $t_{1,2} = .6$, $t_{1,3} = .3$ y $t_{2,3} = .1$. El mapa de probabilidades de presencia que se obtuvo, es decir la realidad, se observa en la Figura 1-1(a).

Con base en las consideraciones hechas en secciones anteriores, se observó que los escenarios a considerar son determinados por los factores: *sesgo espacial*, *conocimiento a priori* y *tamaño de la muestra*, los cuales se describen en los siguientes párrafos.

Sesgo espacial. Con respecto al sesgo espacial, denotado por $\delta(s)$, éste se determina asignando una probabilidad de visitar el nodo s como inversamente proporcional a su distancia a la carretera más cercana. Se consideran las principales carreteras sobre la península de Yucatán para determinar esta cantidad. Para el factor sesgo espacial se consideran los niveles sesgo espacial alto (baja probabilidad de visita asignada a nodos lejos de carreteras) y bajo (alta probabilidad de visita asignada aún a sitios lejos de carreteras). Con el sesgo espacial determinado, el sesgo en las covariables se obtiene mediante la expresión (1.3).

Conocimiento a priori. Como ya se ha comentado, cuando un experto conocedor de la especie bajo estudio observa un mapa de potencial de presencia obtenido mediante alguna de las metodologías existentes, es capaz de señalar regiones que son de alta probabilidad de presencia de la especie y no fueron señaladas en el mapa. De la misma manera, el experto es capaz de identificar zonas donde el mapa de potencial de presencia sobrestima la presencia potencial de la especie. Para obtener esa información de manera formal, se le pide al experto que, sobre el mapa de la región de interés, delimite la zona (posiblemente fragmentada) en la que asegura que la especie bajo estudio, con alta probabilidad, es capaz de establecerse. Denótese por R_1 a esta región. Al delimitar R_1, el experto deberá proporcionar las *zonas de establecimiento potencial* de la especie. Así, si el experto sabe que en una región particular la especie no ha sido observada, y sabe que las condiciones físicas y/o climáticas de dicha zona son adecuadas para el establecimiento exitoso de la especie y que ésta no ha sido capaz de llegar a esta zona por factores externos (barreras naturales, competidores, etc.), entonces esta zona deberá ser señalada como parte de la región R_1. De la misma manera, se le pide al experto que delimite la región donde asegura que la especie, con alta probabilidad, no es capaz de establecerse. Denótese esta segunda región por R_2, la cual también puede ser fragmentada.

Para proporcionar esos mapas el experto podrá recurrir a información auxiliar, como la división política de la región, la localización de montañas, ríos, etc., pero no usar los sitios de presencia con que se cuenta como referencia para determinarlos. Sea $R_3 = R \setminus (R_1 \cup R_2)$ la zona en la que el experto declara inseguridad acerca del establecimiento de la especie, es decir, R_3 es la región no delimitada por el experto. En general se tendrá que $R_3 \neq \emptyset$ y si los mapas R_1 y R_2 son proporcionados de manera sensata, serán tales que $R_1 \cap R_2 = \emptyset$. En la práctica es posible que el experto solamente sea capaz de determinar una de las dos regiones pedidas. Como se verá en la Sección 3.1, esto no representa problema alguno para aplicar las ideas que se introducen.

Las regiones R_1 y R_2 definen el factor denominado conocimiento *a priori*, y como se verá en la Sección 3.1.2, se usarán para elicitar los parámetros de las distribuciones Dirichlet postuladas. Se consideran tres niveles (situaciones diferentes) con respecto a los mapas *a priori*. El primer nivel se estipula suponiendo que las regiones R_1 y R_2 son proporcionadas de tal manera que coinciden aproximadamente con la realidad acerca del establecimiento de la especie. Es decir, R_1 coincide con la región de alta probabilidad de presencia de la especie y R_2 coincide con la región de baja probabilidad de presencia de la especie, según los valores de probabilidad obtenidos de la expresión (1.13). El segundo nivel se estipula suponiendo que las regiones R_1 y R_2 son proporcionadas de tal manera que no coinciden con la realidad. El tercer nivel corresponde al caso en el que no se cuenta con información *a priori*, es decir, $R_1 = R_2 = \emptyset$, $R_3 = R$.

Tamaño de la muestra. Para reproducir el hecho de que los sitios de presencia ocurren típicamente a lo largo de las carreteras y cerca de asentamientos humanos, los sitios de presencia se generan simulando el agrupamiento espacial inducido por las carreteras, como se explica enseguida. Un ejemplar de la especie se establece en el nodo s con probabilidad dada por (1.13). Posteriormente el nodo s se supone visitado por observadores con probabilidad inversamente proporcional a la distancia de s a la carretera más cercana. Finalmente, el ejemplar es detectado con probabilidad d. Con respecto a la detectabilidad, en lo que sigue se postula $d = 1$. La probabilidad de visitar un nodo se modula de tal manera que el número (aleatorio) de sitios de presencia resultante, es decir n, sea de magnitud deseada. Para el tamaño de muestra se consideran 3 niveles: bajo ($0 < n \leq 29$), moderado ($30 \leq n < 59$) y alto ($n \geq 60$). Estos niveles se seleccionaron de manera arbitraria, con la única consideración de que, en general, se cuenta con pocos registros de presencia para las especies estudiadas. El mecanismo con el que se obtienen los sitios de presencia produce un sesgo espacial similar al observado en los sitios de presencia de especies reales.

Para un vector $\boldsymbol{\mu}$ y matriz $\mathbf{A}$ fijos, y para los niveles considerados para los factores, se observará un total de 18 escenarios. Para cada escenario se obtiene el mapa de probabilidades de presencia aplicando la metodología propuesta (vía la distribución posterior exacta y MCMC) y se aplican los métodos alternativos FloraMap y Domain. Los resultados se comparan de manera cualitativa con el mapa de probabilidades obtenido mediante la expresión (1.13). También se obtuvo el mapa de certidumbre (vía la distribución posterior exacta y MCMC) como se explica en la Sección 1.3. El mapa de potencial de establecimiento se obtuvo también usando las metodologías GARP y Bioclim. Sin embargo, con esos métodos se concluiría que prácticamente toda la península es de alto potencial para la presencia de la especie, por lo que no se incluyen dichos resultados. En los siguientes párrafos se discuten los resultados que se observaron en el ejercicio de simulación. Todas las gráficas que se refieran a algún escenario de la simulación se encuentran localizadas al final de este capítulo.

La metodología que se propone es robusta a sitios de presencia localizados lejos (geográficamente) de la región principal de alta probabilidad de presencia. Estos sitios producen que tanto FloraMap como Domain determinen zonas de alto potencial alrededor de dichos puntos, hecho que no se observa en los mapas obtenidos con el método propuesto. Observe por ejemplo las gráficas contenidas en cada una de las Figuras 1-4, 1-8, 1-13 y 1-14. Esta característica observada del método puede ser útil para detectar sitios de presencia anómalos, ya que aquellos sitios de presencia que se encuentran lejos de la región principal de alta probabilidad de presencia no inducen necesariamente un área de alta probabilidad alrededor de ellos. Más aún, se observa que el mapa de certidumbre asigna alto nivel de certidumbre a la probabilidad de presencia, por lo que en una aplicación a especies reales se podría concluir que dichos sitios pueden considerarse anómalos (errores de localización, especie mal identificada, etc.).

Los resultados que se obtuvieron aportan evidencia de que el método que se propone es robusto al sesgo espacial inducido por las carreteras, ya que la región de alta probabilidad se recobra razonablemente bien a pesar del agrupamiento que se observa en los sitios de presencia. Compare por ejemplo las Figuras 1-1 y 1-10, Figuras 1-2 y 1-11, o Figuras 1-3 y 1-12. Cada pareja de figuras corresponde a un nivel diferente del sesgo espacial. Observe que a pesar de considerar el sesgo espacial alto, la zona de alta probabilidad de presencia se recupera de manera adecuada.

A medida que el número de sitios de presencia aumenta, el mapa de certidumbre tiende a producir una región en la que, para casi todos los nodos, se cuenta con alta certidumbre acerca de la probabilidad de presencia obtenida. Este comportamiento puede observarse al comparar las Figuras 1-1(d), 1-2(d) y 1-3(d), o las Figuras 1-4(d), 1-5(d) y 1-6(d), por citar algunos ejemplos, que corresponden a mapas de certidumbre generados con cada uno de los niveles considerados para el tamaño de la muestra. Intuitivamente, el aumento de la certidumbre con el aumento del tamaño de la muestra se debe a que cuando se cuenta con un número grande de sitios de presencia, se espera que los vectores de covariables preferidos por la especie para cada pareja J se reflejen en los conteos $C_J(g)$, $g \in F_J$. Se observarán conteos mayores en los vectores de covariables g preferidos por la especie cuando se hizo presente sobre la región de estudio. Ya que los parámetros de la distribución marginal Beta para cada $g \in F_J$ se definen con base en los conteos, un conteo $C_J(g)$ grande (comparado con los otros conteos obtenidos), producirá que la correspondiente distribución marginal Beta $f(g \mid \mathbf{C}', J)$

sea degenerada y por lo tanto poseerá la mayor parte de su masa concentrada alrededor de su media, por lo que el correspondiente valor $I_J(s)$ será cercano a 1.

Un aspecto conocido de la teoría asintótica propia de la Estadística Bayesiana es que conforme el número de datos aumenta, el efecto que la distribución *a priori* tiene sobre la distribución posterior disminuye. En el contexto del problema que aquí se aborda, a medida que se cuenta con un mayor número de sitios de presencia el efecto de la distribución *a priori* será menor.

Con respecto a los mapas de probabilidades de presencia y de certidumbre obtenidos mediante la distribución posterior exacta (simulada utilizando MCMC) y la aproximación Dirichlet, no se observan diferencias sustanciales entre ellos que pudieran llevar a interpretaciones cualitativamente diferentes de los resultados obtenidos. Para los mapas de probabilidades, este hecho se observa en cada una de las Figuras 1-1 a 1-18, comparando las correspondientes Gráficas (b) y (c). La similitud de los mapas de certidumbre se observa en cualquiera de las Figuras 1-1 a 1-18, al comparar las correspondientes Gráficas (d) y (e). Así, la aproximación Dirichlet que se propone es una herramienta confiable para realizar la inferencia, que posee la ventaja de que las cantidades de interés se obtienen mediante expresiones cerradas.

1.6 Discusión

La forma de proceder que se ha propuesto permite establecer una definición probabilística formal para el concepto de establecimiento potencial en cada nodo s: la probabilidad predictiva dada por la expresión (1.8). Ya que el resultado obtenido es una probabilidad, la interpretación con que se cuenta para los resultados es la natural.

Aunque en el modelo (1.7) se consideran únicamente parejas de covariables, las ideas presentadas pueden extenderse de manera natural a considerar interacciones de orden mayor en el modelo (triadas, cuartetas, etc.). Sin embargo, en vista de los resultados que se han obtenido, no se espera que al considerar estas interacciones se produzca mejoría significativa en los resultados, y si se produce un aumento en el número de parámetros del modelo.

La estructura relativamente simple del modelo (1.7), en el cual se consideran sólo interacciones de parejas de covariables, es compatible con un principio que afirma que una especie, en general, considera un número reducido de covariables y criterios simples para decidir si un sitio es adecuado para su establecimiento (Peterson y Cohoon 1999). Aunque sensato, este principio requiere de pruebas experimentales para su validación. Por otro lado, la consideración de las parejas de covariables permite incluir de manera indirecta información acerca de las correlaciones entre covariables, a través de los conteos.

La consideración de sólo las parejas de covariables produce, en general, que la dimensión del parámetro de interés se reduzca. La reducción ocurre si se cumple que $\sum_{a<b} R_a R_b < |F|$, lo que en general sucede por el hecho de que el número de niveles que poseen las covariables

(es decir las cantidades R_k's) no son pequeños.

Con respecto al sesgo espacial, una manera de determinar el valor $\delta(s)$ para una especie particular surge de la interpretación de esta cantidad y de la siguiente consideración. Aunque en la práctica no se cuenta con un registro completo de lugares que se hayan visitado para evaluar la presencia de la especie bajo estudio, existen bases de datos que contienen registros de presencia de una clase completa de especies, por ejemplo los registros de presencia de la clase aves. Estas bases de datos proporcionan información de sitios visitados en los que se ha registrado la presencia de algún ejemplar de la clase correspondiente, y proporcionan información de la intensidad con que ha sido realizado el muestreo para dicha clase sobre la región de estudio. Una zona particular en la que se reporten muchos registros de presencia (aunque no sean de la especie bajo estudio) aportará evidencia de que en dicha zona se ha realizado algún esfuerzo de colecta. Este esfuerzo no se interpreta necesariamente como expediciones planeadas diseñadas especialmente para detectar a cierta especie, siendo que muchos registros de presencia se deben a avistamientos casuales.

Una aportación que se presenta en este capítulo es la estipulación de una manera de generar un mapa de probabilidades de presencia que puede considerarse como la realidad. Además de su utilidad en la tarea de simular sitios de presencia, la obtención de este mapa podría permitir realizar estudios más detallados para evaluar la forma de proceder de cada método, o bien, comparar el funcionamiento de los métodos, pues se tendrá la realidad con la cual comparar cada uno de los mapas que se obtengan.

El uso de un mapa de certidumbre, que aporta una medida de precisión para la probabilidad de presencia obtenida en cada nodo, es una herramienta que ayuda a realizar inferencias con mayor sustento estadístico. La consideración de una medida de certidumbre en mapas se ha encontrado en pocos trabajos en la literatura. Por ejemplo, Heikkinen y Högmander (1994) y Högmander y Möller (1995) proponen considerar un mapa de probabilidades de presencia para una especie de rana. La aportación de estos trabajos radica en que proporcionan una probabilidad de presencia en cada nodo, en vez de generar un mapa binario de presencia-ausencia. En esos trabajos se argumenta que la probabilidad en sí proporciona la medida de certidumbre. Por su parte, De Oliveira (2000) propone calcular una medida de variabilidad para cada nodo de una retícula en la estimación de campos aleatorios markovianos.

Los mapas de probabilidades de presencia y de certidumbre deben ser utilizados de manera conjunta, como se ilustra a continuación. Si en un nodo particular se obtiene un valor pequeño para la probabilidad de presencia junto con un valor pequeño de certidumbre, se tendrá evidencia de que la probabilidad de presencia en ese nodo podría ser en realidad mayor que la obtenida. En este caso, la probabilidad de presencia de dicho nodo deberá interpretarse con cautela. Por otro lado, si para un nodo se observa un valor pequeño de probabilidad de presencia y un valor grande (cercano a uno) de certidumbre, se tendrá evidencia de que el valor de la probabilidad de presencia en ese nodo es confiable, por lo que se podrá afirmar que el correspondiente nodo no es adecuado para el establecimiento de la

especie. Procediendo con razonamientos de este tipo, conjuntando los valores de probabilidad de presencia y certidumbre obtenidos, podrán emitirse juicios acerca de la probabilidad de presencia de la especie en regiones de interés sobre la zona de estudio, con base en mayor sustento estadístico.

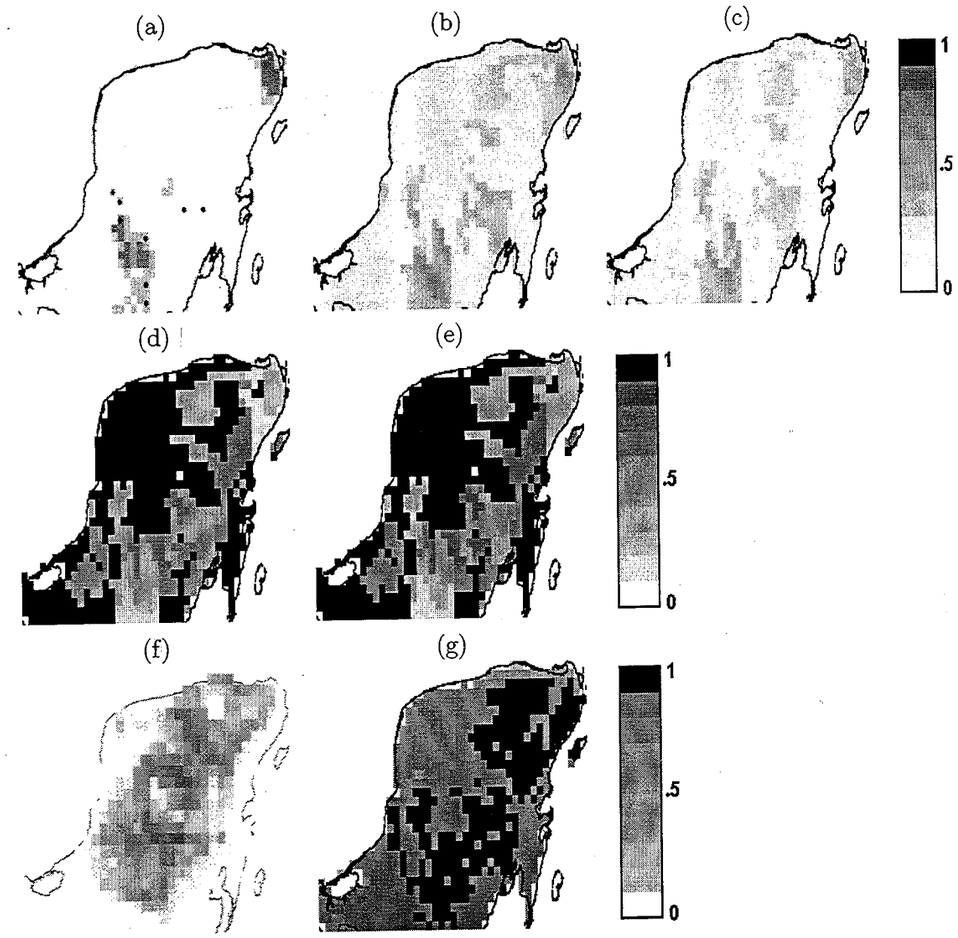

Figura 1-1: Sesgo espacial alto, *a priori* no informativa, n bajo (a) Potencial idealizado y sitios de presencia simulados ($n = 8$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

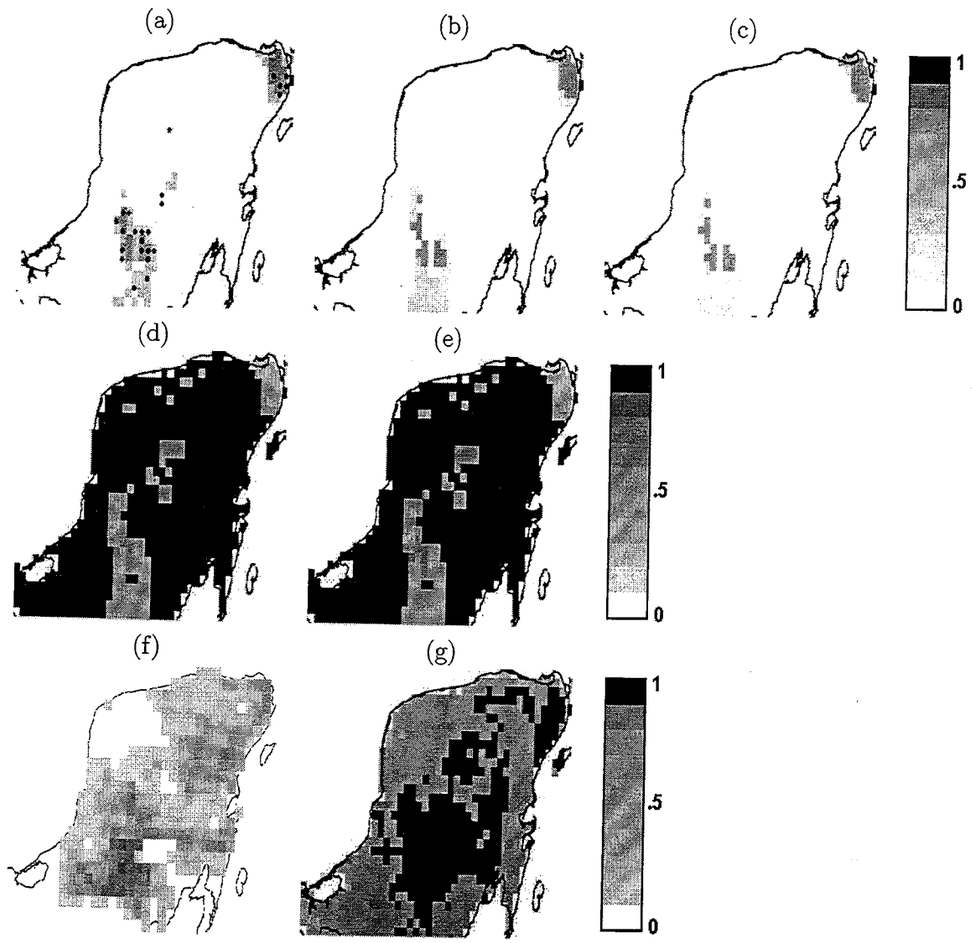

Figura 1-2: Sesgo espacial alto, *a priori* no informativa, n moderado (a) Potencial idealizado y sitios de presencia simulados ($n = 39$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

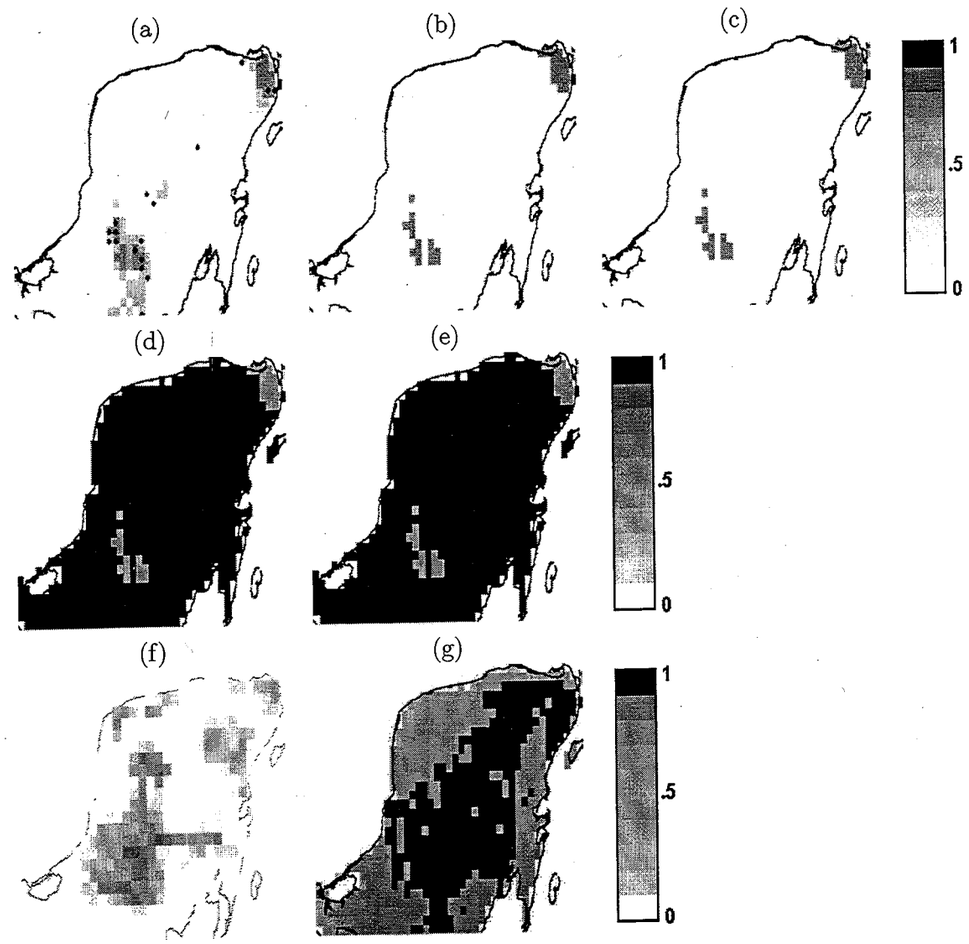

Figura 1-3: Sesgo espacial alto, *a priori* no informativa, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 65$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

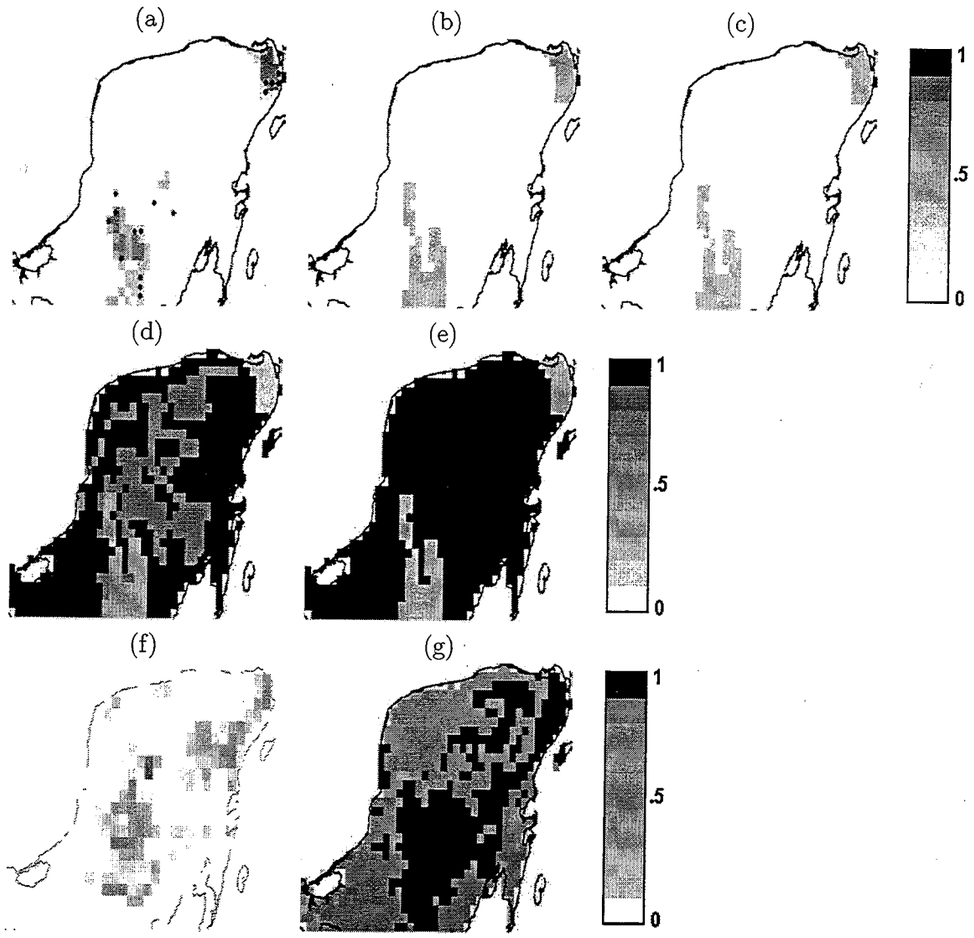

Figura 1-4: Sesgo espacial alto, *a priori* correcta, n bajo (a) Potencial idealizado y sitios de presencia simulados ($n = 17$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

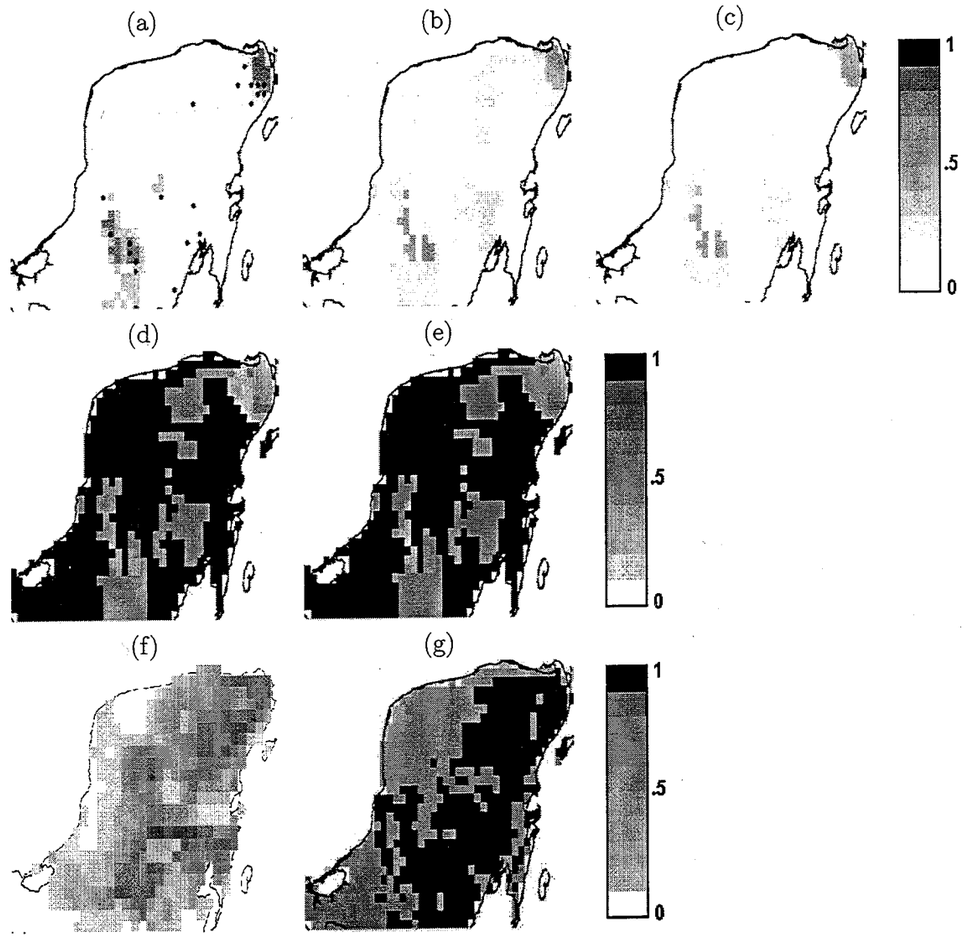

Figura 1-5: Sesgo espacial alto, *a priori* correcta, n moderado (a) Potencial idealizado y sitios de presencia simulados ($n = 39$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

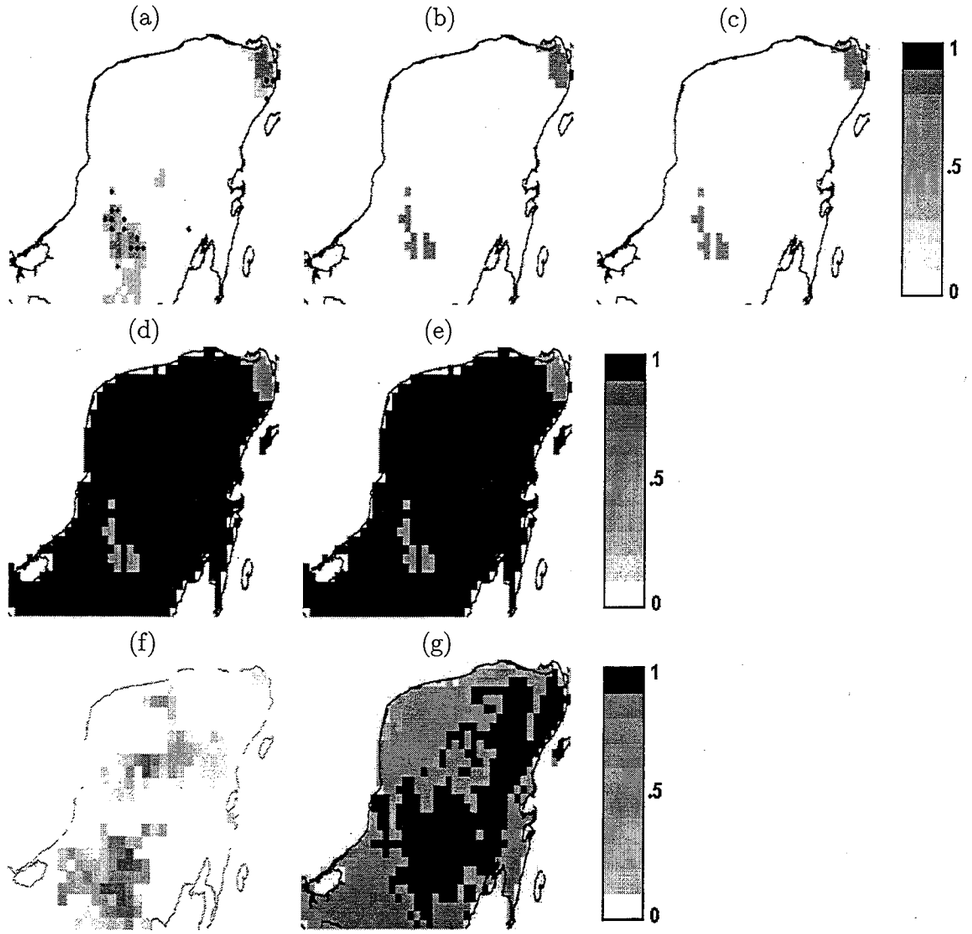

Figura 1-6: Sesgo espacial alto, *a priori* correcta, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 63$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

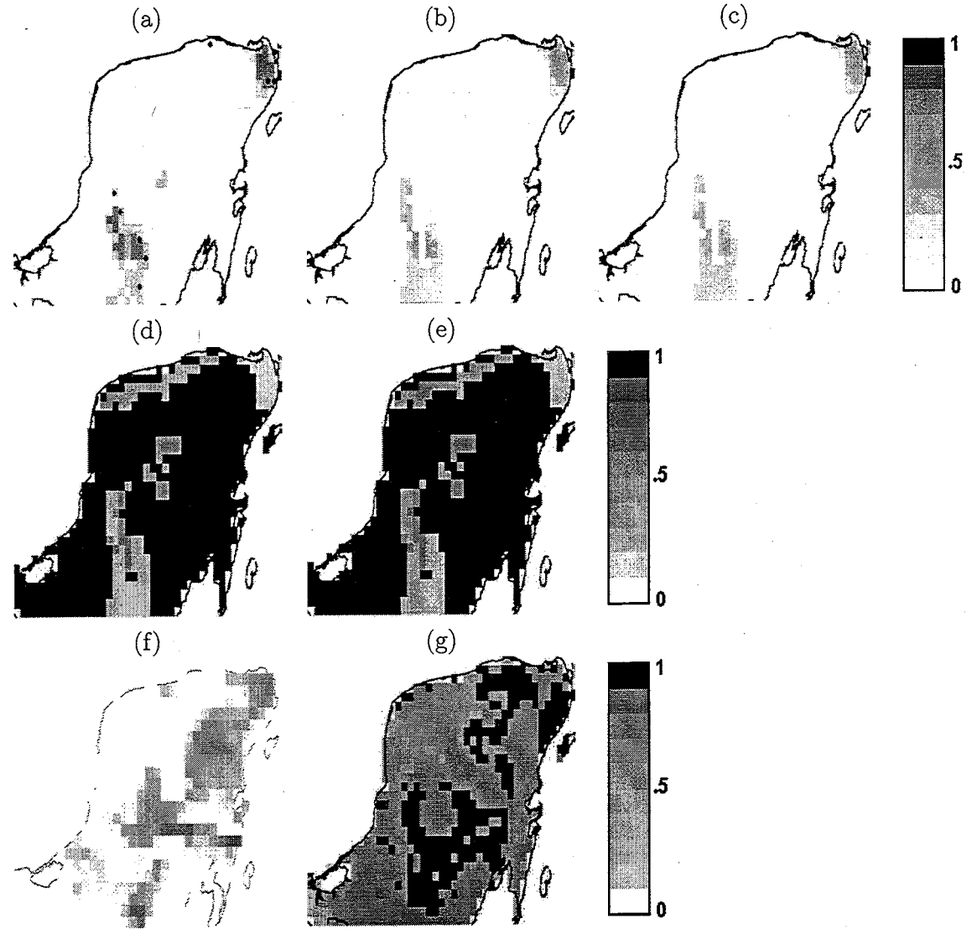

Figura 1-7: Sesgo espacial alto, *a priori* incorrecta, n bajo (a) Potencial idealizado y sitios de presencia simulados ($n = 10$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

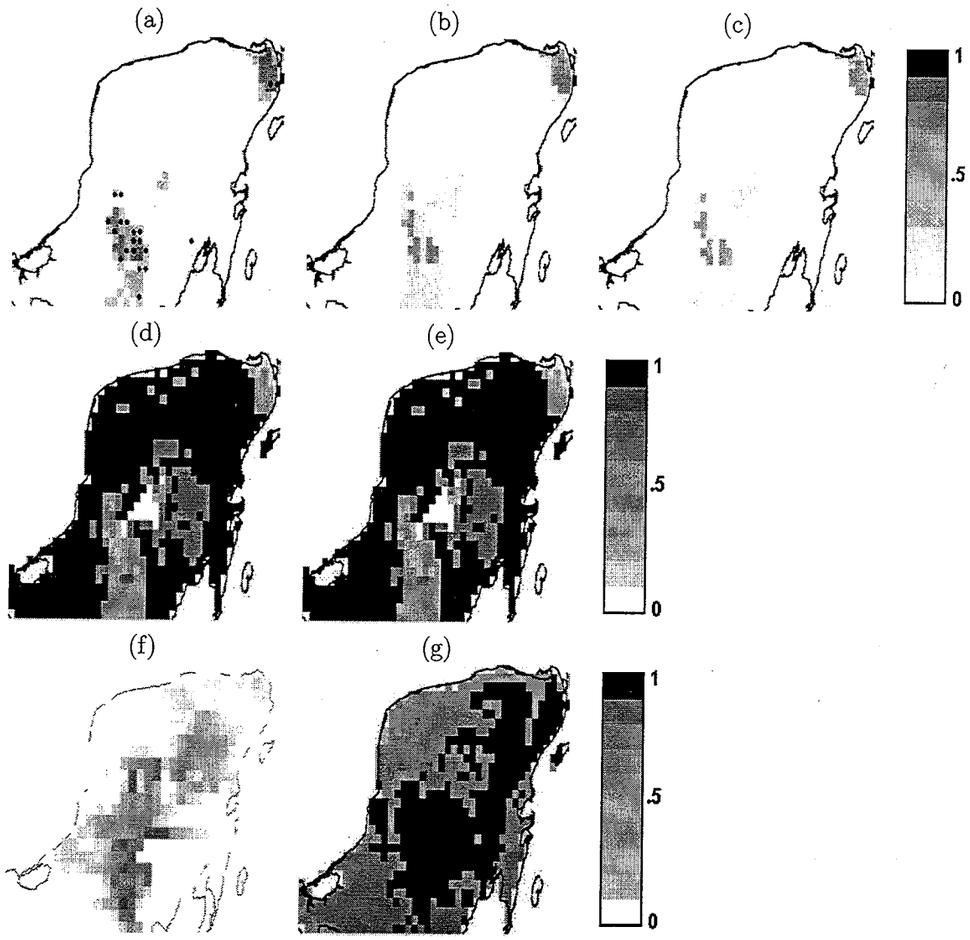

Figura 1-8: Sesgo espacial alto, *a priori* incorrecta, n moderado (a) Potencial idealizado y sitios de presencia simulados ($n = 37$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

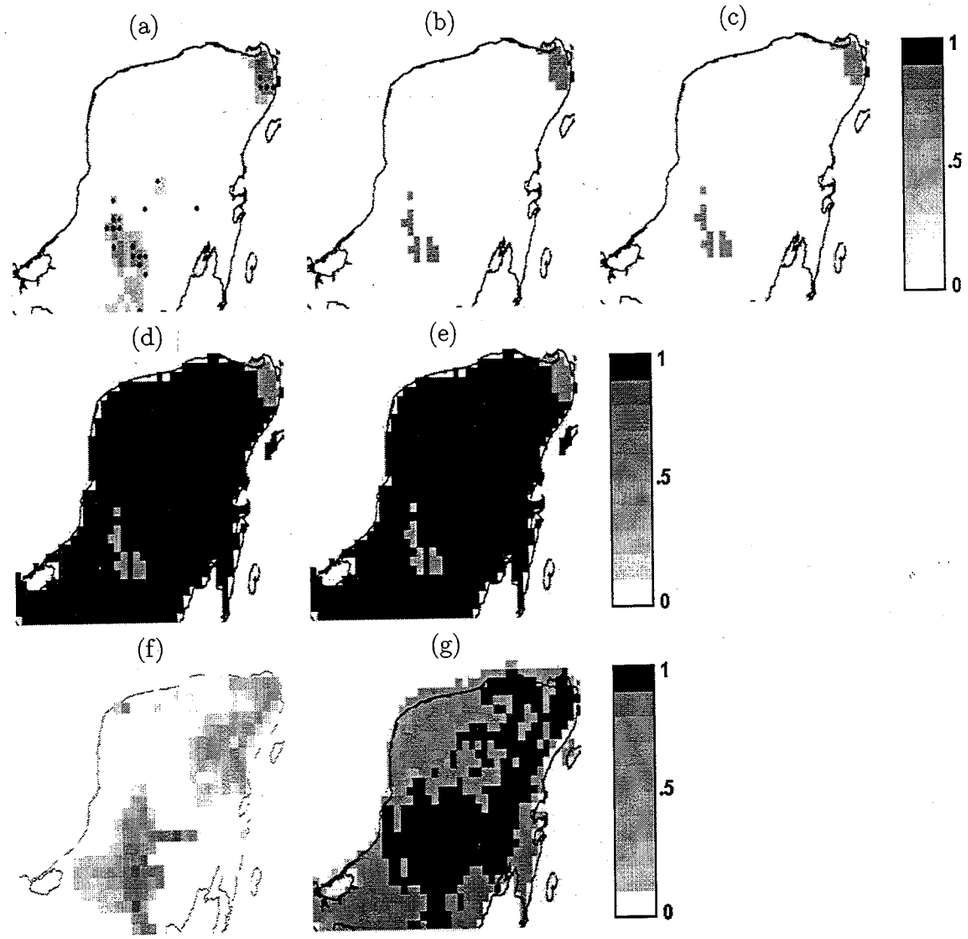

Figura 1-9: Sesgo espacial alto, *a priori* incorrecta, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 69$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

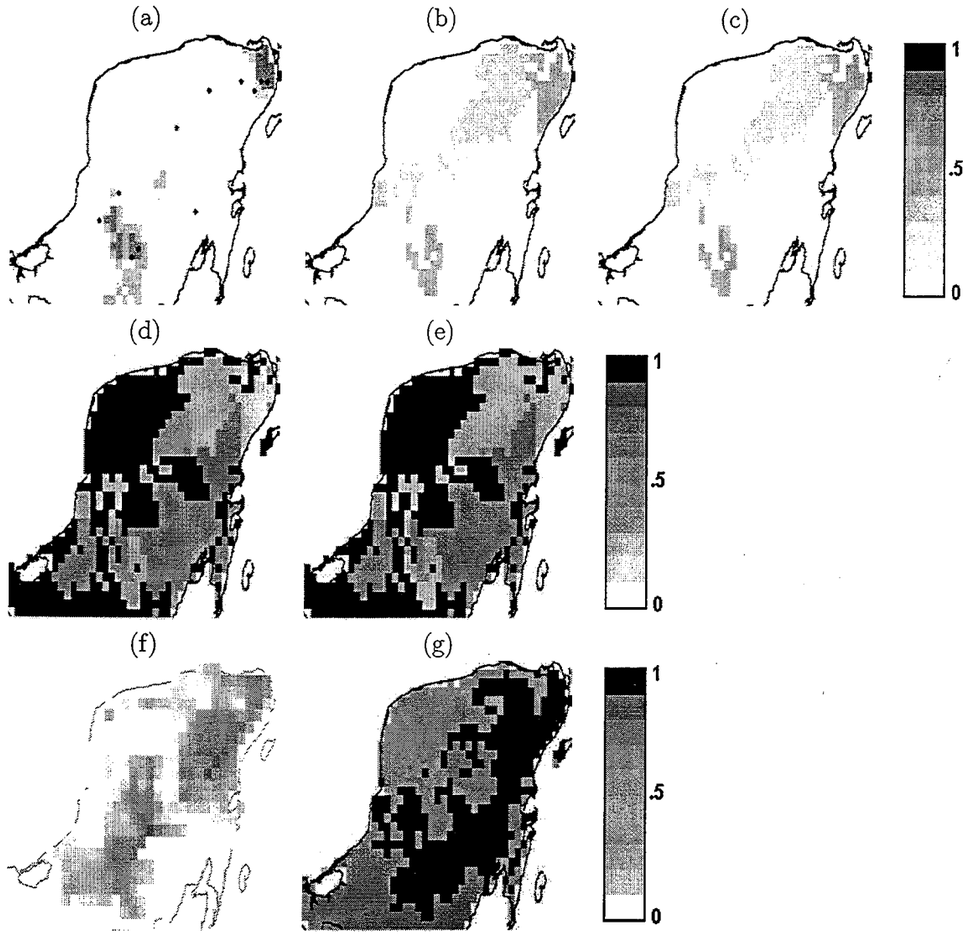

Figura 1-10: Sesgo espacial bajo, *a priori* no informativa, n bajo (a) Potencial idealizado y sitios de presencia simulados ($n = 12$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

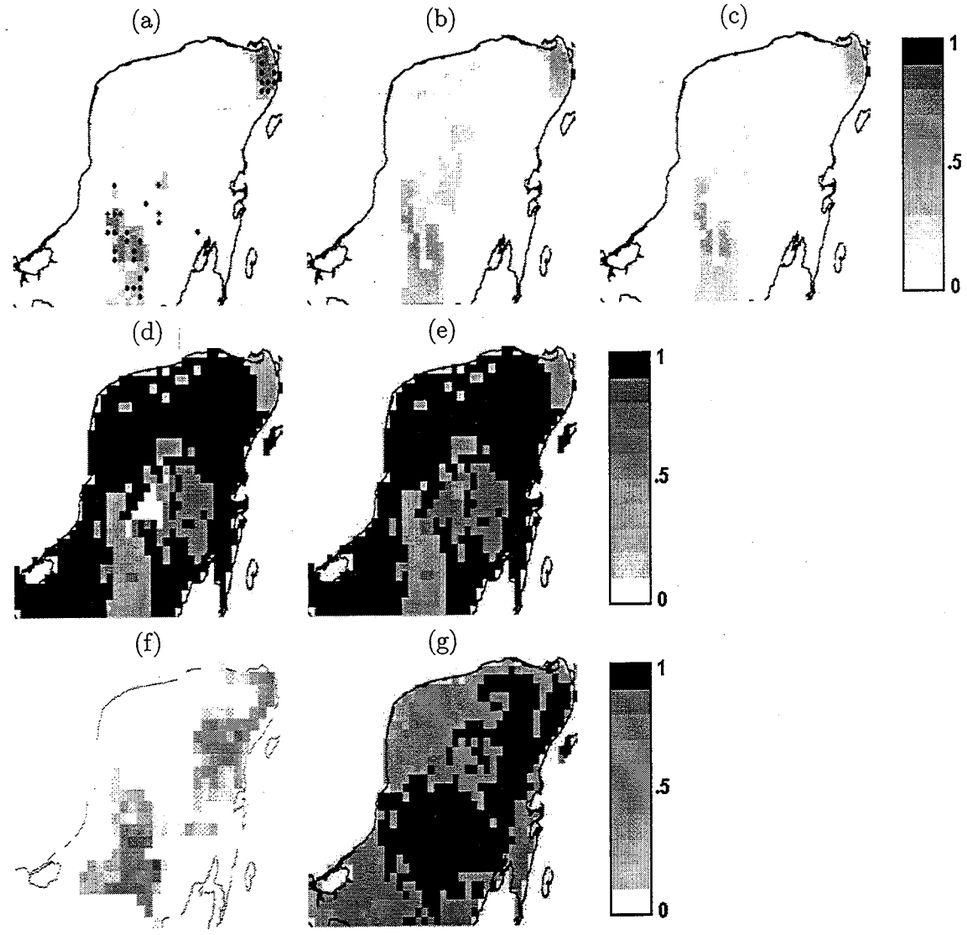

Figura 1-11: Sesgo espacial bajo, *a priori* no informativa, n moderado (a) Potencial idealizado y sitios de presencia simulados ($n = 41$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

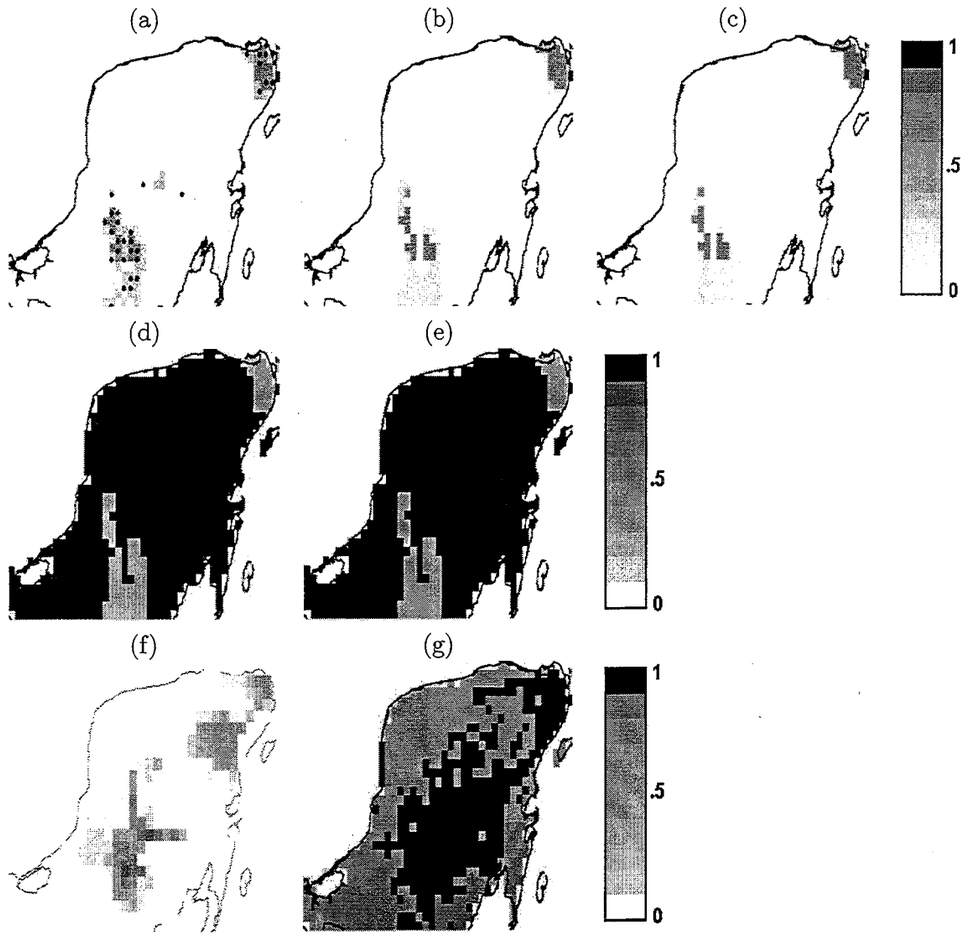

Figura 1-12: Sesgo espacial bajo, *a priori* no informativa, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 77$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

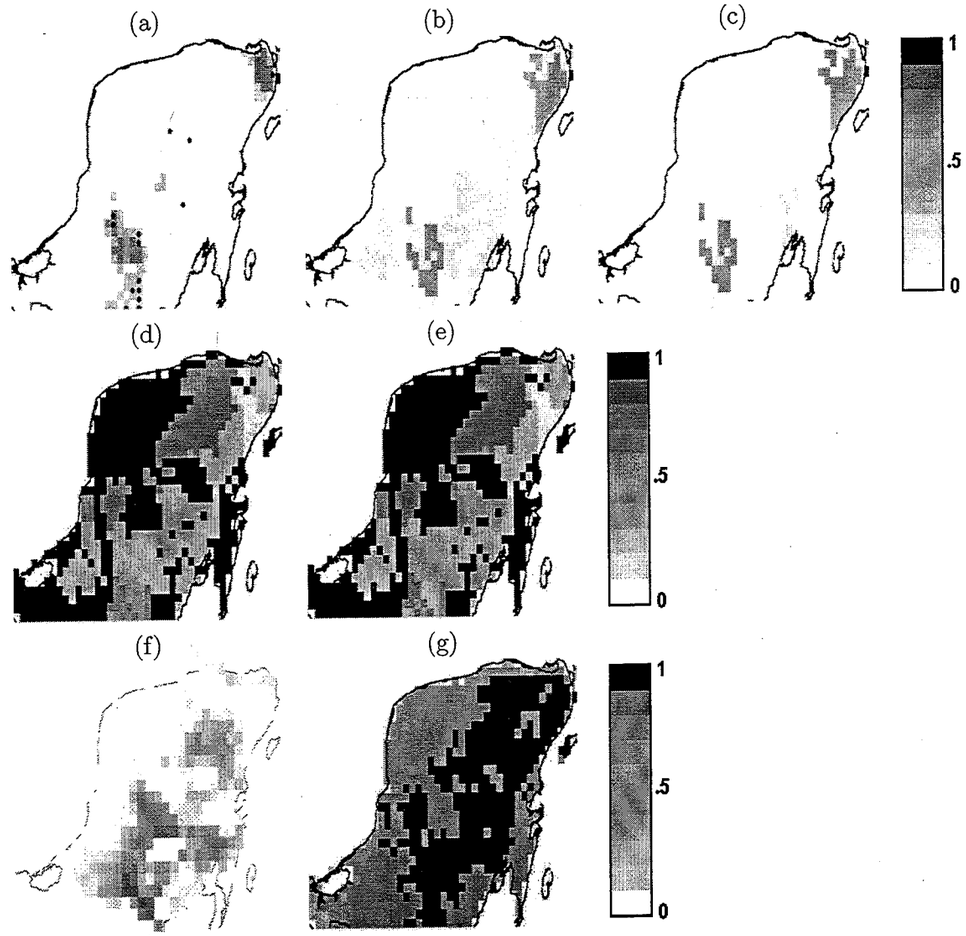

Figura 1-13: Sesgo espacial bajo, *a priori* correcta, n bajo (a) Potencial idealizado y sitios de presencia simulados ($n = 16$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

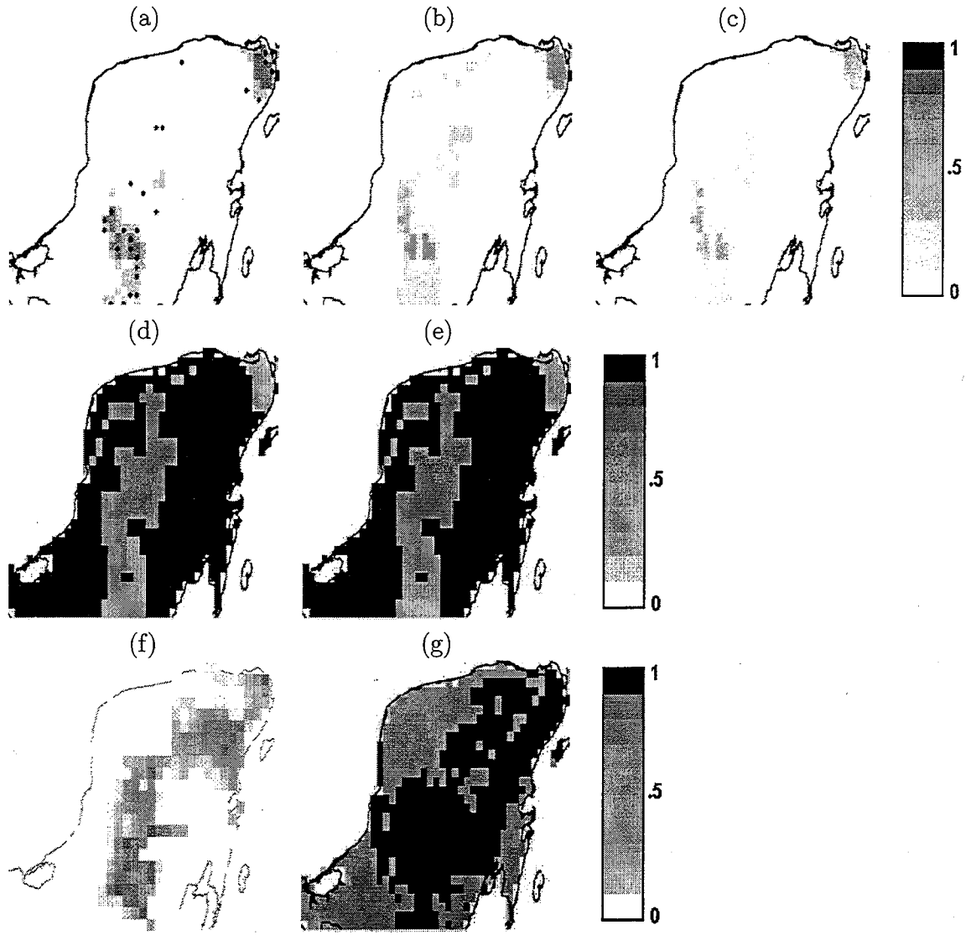

Figura 1-14: Sesgo espacial bajo, *a priori* correcta, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 36$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

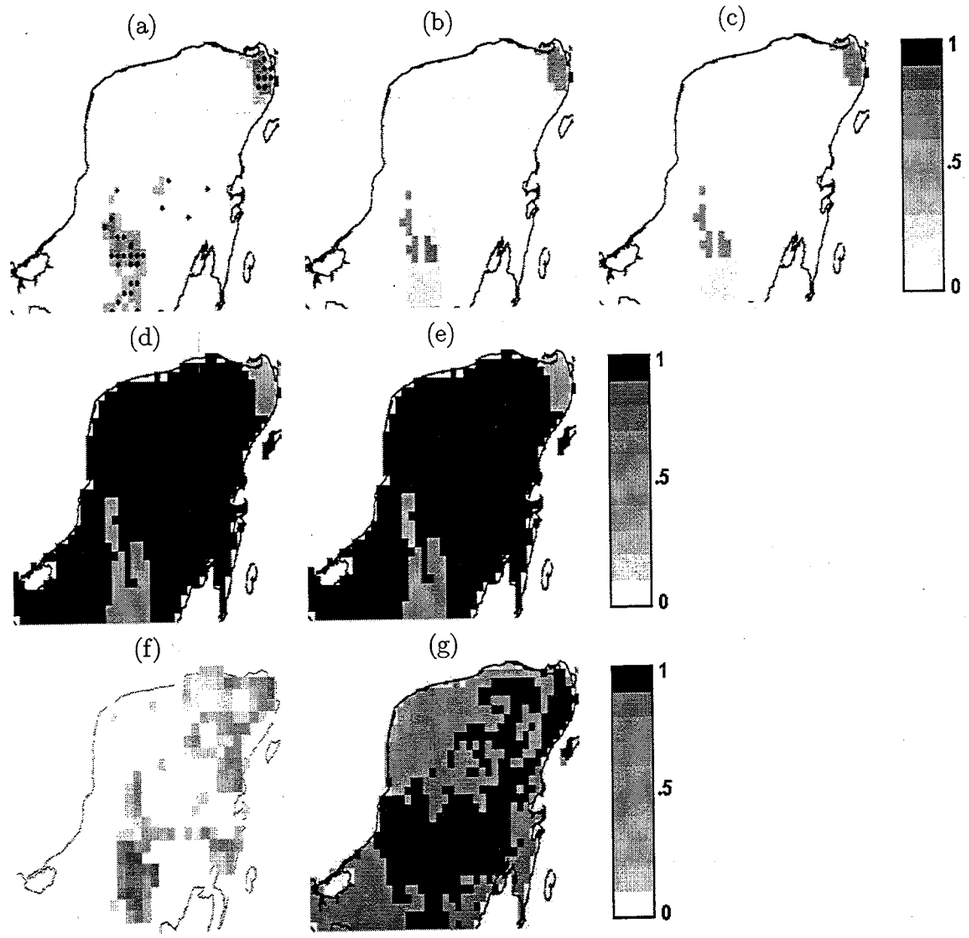

Figura 1-15: Sesgo espacial bajo, *a priori* correcta, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 83$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

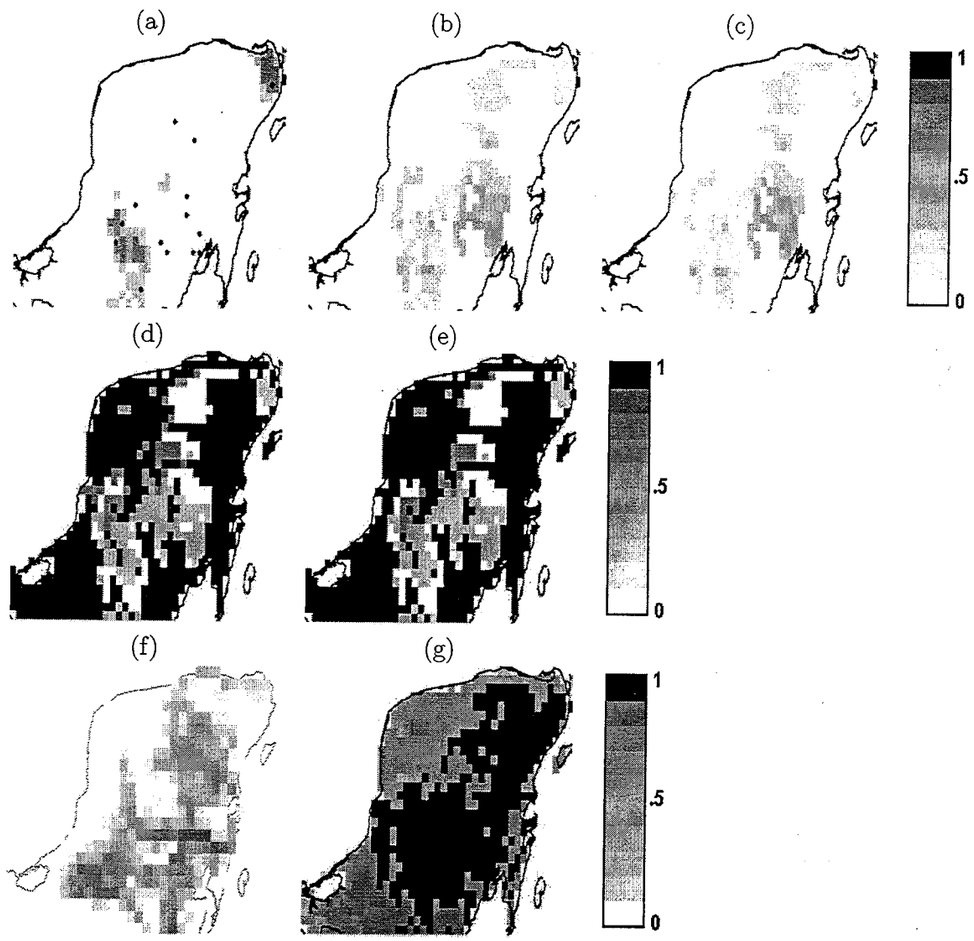

Figura 1-16: Sesgo espacial bajo, *a priori* incorrecta, n bajo (a) Potencial idealizado y sitios de presencia simulados ($n = 14$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

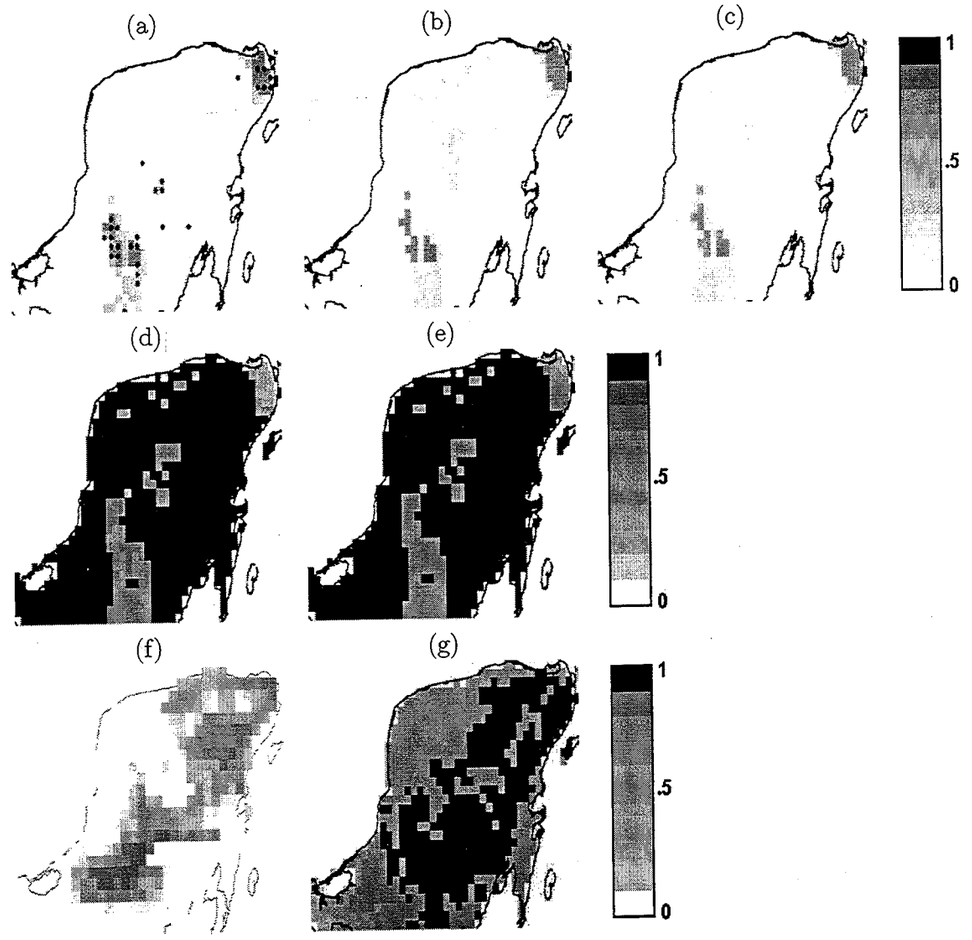

Figura 1-17: Sesgo espacial bajo, *a priori* incorrecta, n moderado (a) Potencial idealizado y sitios de presencia simulados ($n = 43$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

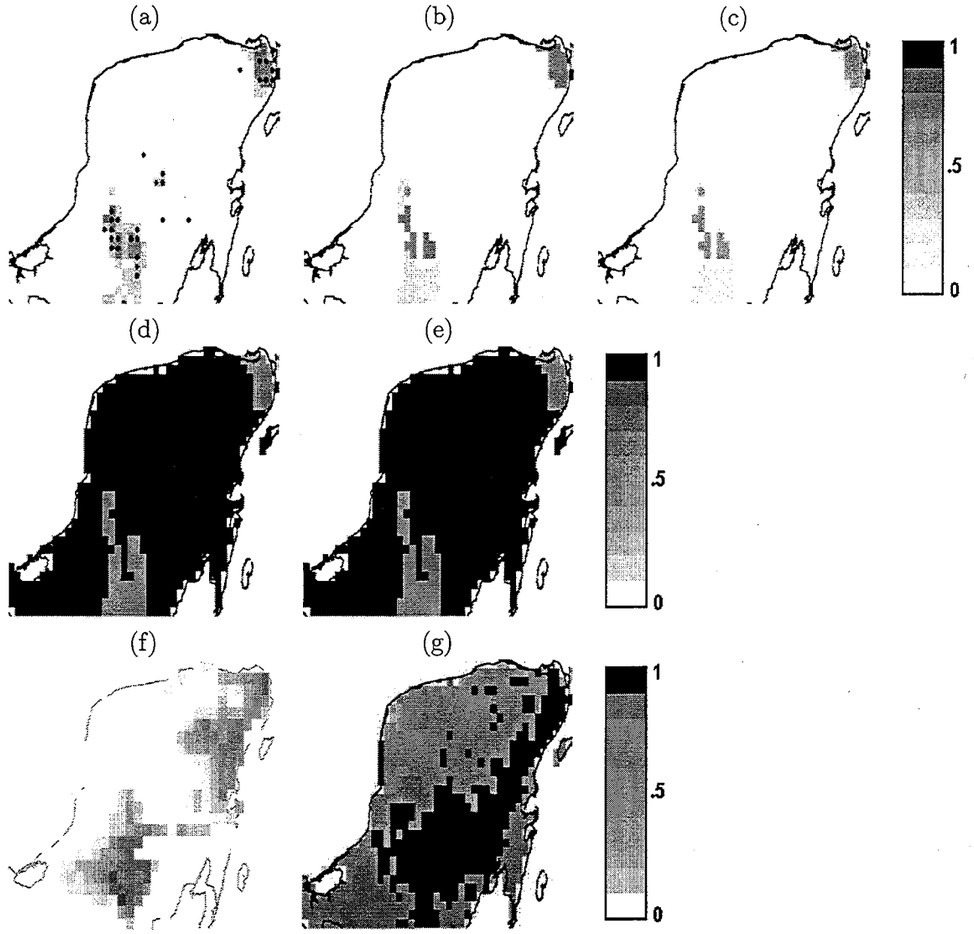

Figura 1-18: Sesgo espacial bajo, *a priori* incorrecta, n alto (a) Potencial idealizado y sitios de presencia simulados ($n = 70$). (b)-(c) Mapa de probabilidades de presencia estimado usando nuestro método con aproximación Dirichlet y MCMC, respectivamente. (d)-(e) Mapa de certidumbre estimado usando la aproximación Dirichlet y MCMC, respectivamente. (f) Potencial estimado usando FloraMap. (g) Potencial estimado usando Domain.

Capítulo 2

Búsqueda de Zonas para Proteger Especies

El problema que se aborda en este capítulo es el siguiente: encontrar un subconjuto de nodos que será propuesto para proteger especies, dado un conjunto de I especies de interés y suponiendo que se cuenta con una estimación de la probabilidad de presencia de cada especie en cada nodo de una retícula. Se supone que la región bajo estudio se encuentra cubierta por una retícula regular.

Para abordar este problema se ha considerado principalmente el enfoque de programación lineal. Bajo este enfoque, Camm *et al.* (2002), Malcolm (2001) y Polasky, Camm y Garber-Yonts (2001) proponen maximizar (o minimizar) una ecuación lineal, que se encuentra sujeta a satisfacer restricciones dadas por consideraciones prácticas. Malcolm (2001) y Costello y Polasky (2003) han abordado este problema desde un enfoque que denominan *dinámico*. Bajo este enfoque, la selección de nodos que serán declarados como zona protegida se realiza en etapas, en cada una de las cuales se selecciona un conjunto de nodos para proteger.

En los trabajos citados se suponen conocidos algunos elementos que en la práctica no son fáciles de determinar, o bien, que producen que el conjunto de soluciones posibles sea sumamente restringido. Por ejemplo, Malcolm (2001) supone que se conoce con certidumbre la presencia o ausencia de cada especie en cada nodo, es decir, la distribución de la especie. Por su parte, Camm *et al.* (2002) suponen que se conoce el número de nodos que se propondrán para proteger, por lo que la solución se restringe a poseer dicho número de nodos.

Aún en caso de que los elementos que se suponen conocidos se encontraran disponibles o haya alguna razón para considerarlos disponibles, los métodos mencionados no toman en cuenta otros elementos en su forma de proceder, principalmente relativos a las especies. Por ejemplo, no consideran que las especies bajo estudio pueden encontrarse en diferentes situaciones con respecto a la urgencia que se tenga por protegerlas. Tampoco consideran que si alguna de las especies bajo estudio es, en algún sentido, *más valiosa* que las otras, su protección debe ser priorizada. En otras palabras, no consideran que cada especie posee un

valor, que en esta tesis se denomina *valor biológico*.

Al obtener una zona para proteger, en los trabajos citados no se considera la posible preferencia por proteger zonas que no sean altamente fragmentadas. Con respecto a esta posible restricción, en el ámbito de protección de especies existe un debate, conocido como debate SLOSS (Single Large or Several Small), en el que se discute si es preferible proteger una sola región grande o varias regiones de menor tamaño, que en conjunto posean la misma área que la región grande. McDonnell *et al.* (2002) consideran el hecho de que las regiones se prefieren no fragmentadas, pero no bajo la perspectiva del debate SLOSS.

En este capítulo se aborda el problema descrito considerando los elementos citados, bajo el enfoque de la Teoría de Decisiones. Los elementos básicos para abordar un problema bajo este enfoque se describen enseguida. El primer elemento se conoce como *espacio de estados de la naturaleza*, que conjunta los posibles resultados acerca del fenómeno estudiado. El segundo elemento es el *espacio de acciones*, que contiene todas las posibles decisiones (acciones) que el usuario puede realizar. El tercer elemento es la *función de pérdida* $L(u,a) : \Theta \times \mathcal{A} \longrightarrow \mathbb{R}$, con $u \in \Theta$ y $a \in \mathcal{A}$, donde Θ denota el conjunto de estados de la naturaleza y $\mathcal{A}$ denota el conjunto de acciones. La función de pérdida asigna una pérdida $L(u,a)$ a cada decisión posible $a \in \mathcal{A}$, dependiendo del estado de la naturaleza $u \in \Theta$ que ocurra. Una vez definidos estos elementos, la decisión que se tome será aquella que minimice $E\left[L(u,a)\right]$, es decir, aquella decisión que produzca la pérdida esperada mínima. Para un estudio detallado de la Teoría de Decisiones puede consultarse, DeGroot (1970) o Berger (1985), entre otros.

Usando los elementos descritos, en esta tesis se aborda el problema de acuerdo con el siguiente esquema. Primeramente se aborda el caso más simple, en el que se debe decidir si se protege un nodo particular y se estudia una sola especie (Sección 2.1). Se define la variable aleatoria $u(s)$, que asume el valor 1 si la especie se encuentra presente en s, y asume el valor 0 de otra manera, lo que produce el espacio de estados de la naturaleza $\Theta_s = \{0,1\}$. Si $a(s)$ define la decisión que se tome en el nodo s, el espacio de acciones es $\mathcal{A}_s = \{0,1\}$, con $a(s) = 0$ si se decide no proteger el nodo s y $a(s) = 1$ si se decide protegerlo. Con estos elementos se propone una función de pérdida $L(u(s), a(s)) : \Theta_s \times \mathcal{A}_s \longrightarrow \mathbb{R}$, en la que la decisión de proteger un nodo s se tomará si se cumple que $E[L(u(s),1)] \leq E[L(u(s),0)]$.

Con base en la función de pérdida del caso simple, se aborda el problema de decidir si se protege un nodo, considerando ahora I especies en el estudio. El espacio de estados de la naturaleza será $\Theta = \Theta_s^1 \times \cdots \times \Theta_s^I$, donde Θ_s^i representa el espacio de estados de la naturaleza para la especie i en el nodo s, mientras que el espacio de acciones permanece igual. La función de pérdida que se propone para este caso se define como la suma ponderada de funciones de pérdida definidas para cada especie en el caso simple. Esta función de pérdida involucra cantidades que representan la importancia que se asigna a cada especie para ser protegida.

Con los elementos definidos para el caso de decidir si se protege un nodo se aborda el problema de proteger una región, es decir, se aborda el problema de seleccionar un conjunto

de nodos para proteger (Sección 2.2). El espacio de estados de la naturaleza es ahora $\Theta = \prod_{s \in R} \left[\Theta_s^1 \times \cdots \times \Theta_s^I\right]$. Ya que cualquier subconjunto de nodos es una solución posible, el espacio de acciones es el conjunto potencia de R, es decir, $\mathcal{A} = \mathcal{P}(R)$. La función de pérdida $L(\mathbf{U}, A) : \Theta \times \mathcal{A} \rightarrow \mathbb{R}$ que se propone, con $\mathbf{U} \in \Theta$ y $A \in \mathcal{A}$, resulta de generalizar la función de pérdida del caso anterior (I especies, decisión en un nodo). La región que se propondrá para proteger será aquel subconjunto de nodos $A \in \mathcal{A}$ que produzca la mínima pérdida esperada.

En la práctica, al abordar este problema pueden considerarse diversas restricciones sobre el conjunto de nodos que se propondrá para proteger. En esta tesis se consideran dos de ellas, que son comunes en la práctica. La primera se denomina *restricción por presupuesto*, la cual surge por el hecho de que, en general, se cuenta con un presupuesto fijo destinado a la protección de las especies. La restricción por presupuesto se involucra considerando como espacio de soluciones a los conjuntos de nodos que implican una inversión menor o igual que el presupuesto con que se cuenta. Así, se restringe el espacio de decisiones.

La segunda restricción que se considera surge cuando por razones económicas y/o biológicas se prefiere que la región que se proponga para proteger no sea altamente fragmentada. Esta restricción se denomina *restricción por conexidad* y se involucra considerando un término adicional en la función de pérdida, mediante el cual se manipula el nivel de influencia que se desea imponer a la preferencia por zonas no fragmentadas. Este término permitirá observar soluciones desde la perspectiva del debate SLOSS. El problema de encontrar una solución a la que se imponen restricciones se aborda en la Sección 2.3.

Aunque debido a la restricción por presupuesto se considera el espacio de acciones restringido, éste posee en general un número tal de elementos que no es posible buscar exhaustivamente un elemento que posea pérdida esperada mínima. Se recurre por lo tanto a algoritmos de búsqueda no exhaustivos (Sección 2.4). Existen dos algoritmos ampliamente utilizados para realizar búsquedas sobre espacios que contienen un gran número de elementos: el algoritmo *greedy* y el algoritmo *simulated annealing*.

El algoritmo greedy realiza una búsqueda no aleatoria de un conjunto de nodos que minimiza la pérdida esperada. Por su forma de proceder es posible que la solución que se obtenga corresponda a un mínimo local de la función a minimizar. Por su parte, el algoritmo simulated annealing realiza una búsqueda aleatoria de la solución, mediante un proceso iterativo. Dada una solución en una iteración, la idea de este algoritmo consiste en generar una solución candidata, agregando o suprimiendo un elemento a la solución actual. La solución candidata se acepta como nueva solución con cierta probabilidad, la cual depende de la magnitud de la diferencia de la pérdida esperada de la solución actual y la solución candidata. En la Sección 2.4 se procede a utilizar ambos algoritmos a la vez, proponiendo como solución inicial para el algoritmo simulated annealing la solución que se obtenga con el algoritmo greedy. Esta forma de proceder permite aprovechar las ventajas operativas y teóricas que ofrecen ambos algoritmos.

Una vez que se ha definido la función de pérdida que se utilizará y se ha determinado la manera de realizar la búsqueda de la solución, se procede a implementar el método. Al realizar algunas consideraciones biológicas con respecto a las especies, es posible determinar cantidades particulares para los valores de los que depende la función de pérdida. Estas cantidades se relacionan con los factores que los enfoques usados para abordar este problema no consideran. Así, la función de pérdida que se utiliza considera (1) una cantidad que refleja la importancia de proteger cada especie (cantidad que se denota por w_i), (2) una cantidad que se interpreta como el costo biológico de la especie (cantidad denotada por z_i) y (3) un parámetro que se relaciona con el debate SLOSS (denotado por β), el cual permite observar soluciones con diferente grado de fragmentación.

Para investigar el funcionamiento de la metodología que se propone se realizó un estudio de simulación en el que se consideran 3 especies (Sección 2.6). Para cada una de estas especies se obtuvo un mapa de probabilidades de presencia de establecimiento, utilizando las ideas del ejercicio de simulación del Capítulo 1, es decir, se postuló un vector de covariables que define el vector óptimo para el establecimiento de cada especie. Estos vectores se utilizan para generar un mapa de presencia que se considera como el mapa real. Al proponer diversos valores para las cantidades β, w_i y z_i, $1 \leq i \leq I$, se generan escenarios que permiten investigar el efecto de esos valores al obtener la región que se propone para proteger.

De esta simulación se observa que el parámetro β es útil desde el punto de vista del debate SLOSS. A medida que el parámetro β se incrementa, se obtiene como solución una región menos fragmentada. Si se postulan varios valores para este parámetro, será posible comparar las soluciones que se obtengan por medio de las pérdidas esperadas, como se explica al final de la Sección 2.6.

Para observar el efecto de las cantidades w_i's sobre la región a proteger se procedió a postular $z_i = z$ para toda i. Se observó que la zona que se propone para proteger posee mayor cantidad de nodos pertenecientes a la región de alta probabilidad de presencia de la especie a la que se asignó el mayor valor w_i.

Para observar el efecto de las cantidades z_i's se procedió a postular $w_i = 1/I$, es decir, se postula que cada especie posee la misma urgencia por ser protegida. Se observó que la zona que se obtiene para proteger posee más nodos pertenecientes a la zona de alta probabilidad de presencia de la especie que posee mayor valor z_i.

La forma de proceder que se propone en este capítulo se basa en la postulación de una función de pérdida que considera cantidades relevantes acerca de las especies. La consideración de que las especies pueden estar en diferente situación de amenaza, por medio de asignar un peso w_i a cada una de ellas es una aportación relevante de este capítulo. Otra aportación es la consideración de que cada especie posee un valor biológico, el cual se introduce en la función de pérdida a través de la cantidad z_i. La consideración de las cantidades z_i y w_i permite obtener regiones para proteger de acuerdo con las necesidades de las especies bajo estudio, las cuales se reflejan a través de las magnitudes de esas cantidades.

2.1 Toma de Decisión en un Sitio

Supóngase por el momento que el objetivo que se persigue es decidir si el sitio s debe ser protegido y se estudia una sola especie. En particular, se supone que los sitios de interés son nodos determinados por una retícula que cubre la región bajo estudio. Así, se supone que $s \in R$, donde R es el conjunto de tales nodos. El espacio de decisiones puede especificarse por medio de la variable binaria $a(s)$, que toma el valor 0 si se decide no proteger el nodo s, y toma el valor 1 si se decide protegerlo. Así, el conjunto de decisiones se postula como $\mathcal{A}_s = \{0, 1\}$, y $a(s) \in \mathcal{A}_s$ indica la acción a realizar en el nodo s.

El espacio de estados de la naturaleza para la especie, de aquí en adelante denominado simplemente *espacio de estados*, puede especificarse mediante la variable aleatoria binaria $u(s)$, que toma el valor 0 si la especie no se encuentra presente en s, mientras que toma el valor 1 si se encuentra presente. Así, $\Theta_s = \{0, 1\}$ representa el espacio de estados para la especie considerada. Note que la variable aleatoria $u(s)$ es la misma que se define en el capítulo anterior, por lo que $p(s) = P(u(s) = 1)$ podrá ser la estimada por medio de la metodología propuesta en el Capítulo 1. En general, siempre que la cantidad $p(s)$ denote la probabilidad de presencia de la especie en el nodo s, y si se cuenta con esta cantidad para los nodos de interés, las cantidades $p(s)$, $s \in R$, podrán utilizarse para aplicar la metodología que aquí se propone.

La función de pérdida que se propone para tomar la decisión en el nodo s, $L_s(u(s), a(s)) : \Theta_s \times \mathcal{A}_s \longrightarrow \mathbb{R}$, con $u(s) \in \Theta_s$ y $a(s) \in \mathcal{A}_s$, se resume en la Tabla 1.

Tabla 1: Función de pérdida para la especie en $s \in R$.

$\Theta_s \backslash \mathcal{A}_s$	$a(s) = 0$	$a(s) = 1$
$u(s) = 0$	$x(s)$	$y(s)$
$u(s) = 1$	$z(s)$	$t(s)$

De acuerdo con la Tabla 1, la pérdida asociada a tomar la decisión de no proteger un nodo $(a(s) = 0)$ cuando la especie no se encuentra presente en él $(u(s) = 0)$, es $L_s(0, 0) = x(s)$. Al definir la función de pérdida por medio de la Tabla 1, se asume implícitamente que las cantidades $x(s)$, $y(s)$, $z(s)$ y $t(s)$ son comparables, es decir, que se encuentran medidas en la misma unidad (pesos mexicanos, por ejemplo). En la Sección 2.5 se presentan algunas consideraciones que permiten asignar valores particulares a las cantidades de la Tabla 1.

Siguiendo el procedimiento estándar de la Teoría de Decisiones, la decisión que se tomará en el nodo s será aquella que produzca la menor pérdida esperada. Así, si se define $L_s(a(s)) = E[L_s(u(s), a(s))]$, la pérdida esperada correspondiente a la acción $a(s) \in \mathcal{A}_s$. La decisión que se tomará en el nodo s es $a^*(s) = \arg\min_{a(s) \in \mathcal{A}_s} \{L_s(a(s)\}$. De acuerdo con la Tabla 1, la pérdida esperada correspondiente a la acción de no proteger $(a(s) = 0)$ y proteger $(a(s) = 1)$

el nodo s es

$$\begin{aligned} L_s(0) &= x(s) - p(s)x(s) + p(s)z(s) \\ L_s(1) &= y(s) - p(s)y(s) + p(s)t(s), \end{aligned}$$

respectivamente, por lo que se tomará la decisión de proteger el nodo s si se cumple que $L_s(1) \leq L_s(0)$, es decir, si $y(s) - x(s) \leq p(s)\{z(s) - x(s) + y(s) - t(s)\}$. Si se procede a calcular la pérdida esperada para cada acción en cada nodo de la región de interés, el mapa a proteger estará determinado por todos aquellos nodos que satisfacen

$$p(s) \geq \frac{y(s) - x(s)}{z(s) - x(s) + y(s) - t(s)}.$$

Cuando se consideran I especies, para cada una de ellas se tendrá una función de pérdida similar a la resumida en la Tabla 1. Cada tabla estará definida por los valores correspondientes $x_i(s)$, $y_i(s)$, $z_i(s)$ y $t_i(s)$, $1 \leq i \leq I$. Así, para cada especie se propone considerar la función de pérdida resumida en la Tabla 2.

Tabla 2: Función de pérdida para la i-ésima especie en $s \in R$.

$\Theta_s \backslash \mathcal{A}_s$	$a(s) = 0$	$a(s) = 1$
$u_i(s) = 0$	$x_i(s)$	$y_i(s)$
$u_i(s) = 1$	$z_i(s)$	$t_i(s)$

La variable aleatoria $u_i(s)$ tomará el valor 0 si la i-ésima especie no se encuentra presente en el nodo s y tomará el valor 1 si se encuentra presente. La probabilidad de presencia de la i-ésima especie se denota ahora por $p_i(s)$.

Ya que se cuenta con I especies, para postular la presencia $(u_i(s) = 1)$ o ausencia $(u_i(s) = 1)$ de cada una en el nodo s, se define el vector aleatorio $\mathbf{U}_s = (u_1(s), ..., u_I(s))$. El espacio de decisiones es nuevamente el conjunto $\mathcal{A}_s$, pues aún se desea tomar la decisión en un solo nodo. Si ahora se denota por Θ_s^i al espacio de estados para la especie i, entonces $\Theta_s = \Theta_s^1 \times \cdots \times \Theta_s^I$ define el espacio de estados para las especies consideradas. Este espacio contiene todas las posibles combinaciones de presencia-ausencia de las I especies en el nodo s. Si por ejemplo se tiene $I = 2$, el espacio Θ_s será $\{(0,0),(1,0),(0,1),(1,1)\}$, que representa las situaciones: ausencia de las dos especies, presencia sólo de la primera especie, presencia sólo de la segunda especie y presencia de las dos especies, respectivamente. La función de pérdida, $L_s(\mathbf{U}_s, a(s)) : \Theta_s \times \mathcal{A}_s \longrightarrow \mathbb{R}$, con $\mathbf{U}_s \in \Theta_s$ y $a(s) \in \mathcal{A}_s$, que se propone para tomar la decisión de proteger o no proteger un nodo cuando se consideran I especies, se estipula como

$$L_s(\mathbf{U}_s, a(s)) = \sum_{i=1}^{I} w_i L_s^i(u_i(s), a(s)), \tag{2.1}$$

con $w_i \in [0,1]$, $\sum_{i=1}^{I} w_i = 1$, donde $L_s^i(u_i(s), a(s))$ es la función de pérdida resumida en la Tabla 2. La cantidad w_i se interpreta como el grado de importancia que se asigna a la i-ésima especie para ser protegida. Si se postula que $w_i \approx 0$, entonces la i-ésima especie no será considerada muy importante para proteger. Si por el contrario se postula $w_i \approx 1$, entonces la i-ésima especie será considerada muy importante para proteger. Si no existe preferencia por proteger alguna especie, se podrá postular $w_i = 1/I$ para toda i. En la Sección 3.2.2 se discute cómo asignar valores sensatos a las w_i's. La posibilidad de considerar una cantidad que pondere la importancia que cada especie posee para ser protegida fue sugerida por Polasky *et al.* (2001), aunque en su trabajo supone que todas las especies son igualmente importantes para proteger.

Utilizando para cada especie la función de pérdida definida en la Tabla 2, se obtiene que la pérdida esperada para la acción *no proteger* $(a(s) = 0)$ y *proteger* $(a(s) = 1)$ el nodo s es

$$\begin{aligned} L_s(0) &= \sum_{i=1}^{I} w_i \{x_i(s) - p_i(s)x_i(s) + p_i(s)z_i(s)\} \quad \text{y} \\ L_s(1) &= \sum_{i=1}^{I} w_i \{y_i(s) - p_i(s)y_i(s) + p_i(s)t_i(s)\}, \end{aligned} \tag{2.2}$$

respectivamente. Para obtener las expresiones (2.2) se supone implícitamente que

$$P\{u_1(s) = q_1, ..., u_I(s) = q_I \mid \mathbf{e}(s), s \in R\} = \prod_{i=1}^{I} p_i(s)^{q_i} [1 - p_i(s)]^{1-q_i},$$

donde q_i puede asumir los valores cero (ausencia de la i-ésima especie) o uno (presencia de la i-ésima especie). Es decir, se supone que las especies se establecen en la región de estudio de manera independiente, condicionadas a los valores de las covariables. En otras palabras, dados los vectores de covariables en cada nodo de la región, los lugares de establecimiento se suponen seleccionados independientemente por las especies. Este supuesto se adopta en las metodologías existentes, con el objetivo de facilitar los cálculos requeridos para obtener una región a proteger. Como antes, la decisión de proteger el nodo s se tomará si se cumple la desigualdad $L_s(1) \leq L_s(0)$, es decir, si

$$\sum_{i=1}^{I} w_i \{y_i(s) - x_i(s)\} \leq \sum_{i=1}^{I} w_i p_i(s) \{z_i(s) - x_i(s) + y_i(s) - t_i(s)\}.$$

Suponiendo que la decisión se toma independientemente en cada nodo, la zona que será propuesta para ser protegida es simplemente el conjunto de nodos en los que la decisión fue

proteger, es decir

$$A = \left\{ s \in R : \sum_{i=1}^{I} w_i \{y_i(s) - x_i(s)\} \leq \sum_{i=1}^{I} w_i p_i(s) \{z_i(s) - x_i(s) + y_i(s) - t_i(s)\} \right\}. \quad (2.3)$$

El conjunto (2.3) proporciona una solución a una forma particular del problema: la decisión se toma en cada nodo y no se impone restricción alguna a la solución. En la práctica, el problema de interés consiste en determinar una región que será propuesta como protegida y no en decidir si un solo nodo debe o no ser protegido. En las siguientes secciones se utilizan las ideas presentadas en esta sección para proponer una forma de proceder cuando el objetivo es encontrar una *región* (conjunto de nodos) para proteger.

2.2 Buscando una Región: sin Restricciones

Cuando se aborda el problema con el objetivo de proteger una región, cualquier subconjunto de nodos de R puede ser una solución, por lo que el espacio de decisiones es ahora el conjunto potencia de R, es decir, $\mathcal{A} = \mathcal{P}(R)$. La función de pérdida $L(\mathbf{U}, A) : \Theta \times \mathcal{A} \longrightarrow \mathbb{R}$, con $\Theta = \prod_{s \in R} \Theta_s$, $\mathbf{U} = \prod_{s \in R} \mathbf{U}_s$ y $A \in \mathcal{A}$, que se propone es

$$L(\mathbf{U}, A) = \sum_{s \in R} L_s(\mathbf{U}_s, a(A, s)), \quad (2.4)$$

donde $L_s(\mathbf{U}_s, a(A, s))$ es la función de pérdida definida por la expresión (2.1) y

$$a(A, s) = \begin{cases} 0 & \text{si } s \notin A \\ 1 & \text{si } s \in A. \end{cases}$$

La notación $a(A, s)$ generaliza lo que antes se denotaba por $a(s)$, que se utilizó cuando la decisión se tomaba en un solo nodo, y es de hecho la función indicadora $I_A(s)$. Aplicando el valor esperado a la expresión (2.4) y utilizando los valores de pérdida esperada dados en (2.2), se obtiene que la pérdida esperada correspondiente a un conjunto $A \in \mathcal{A}$ está dada por

$$\begin{aligned} L(A) \;=\; & \sum_{s \notin A} \sum_{i=1}^{I} w_i \{x_i(s) - p_i(s)x_i(s) + p_i(s)z_i(s)\} + \\ & \sum_{s \in A} \sum_{i=1}^{I} w_i \{y_i(s) - p_i(s)y_i(s) - p_i(s)t_i(s)\}. \end{aligned} \quad (2.5)$$

La región A^* que será propuesta para ser protegida será el conjunto de nodos $A \in \mathcal{A}$ que produzca la menor pérdida esperada, es decir

$$A^* = \arg\min_{A \in \mathcal{A}} \{L(A)\}.$$

En caso de no imponer restricciones a la solución y suponiendo que la decisión se toma de manera independiente en cada nodo, el conjunto A^* coincide con el conjunto dado por (2.3). Este hecho se prueba en la siguiente proposición.

Proposición 3 *Sea $L(\mathbf{U}, A) : \Theta \times \mathcal{A} \longrightarrow \mathbb{R}$ la función de pérdida definida por la expresión (2.4). Si no se impone restricción alguna a la región que se desea proteger, el conjunto A^* que minimiza $L(A) = E[L(\mathbf{U}, A)]$ es el conjunto de nodos para los que la decisión óptima es proteger, utilizando las pérdidas esperadas dadas en (2.2).*

Demostración. Sean $s_1, ..., s_N$ los nodos contenidos en la región bajo estudio. Para cada nodo, sean $L_s(0)$ y $L_s(1)$ las pérdidas esperadas dadas en la expresión (2.2). Sea $L_s^*(a^*(s))$ la pérdida esperada correspondiente a la decisión óptima $a^*(s)$ en el nodo s, es decir

$$L_s^*(a^*(s)) = \min\{L_s(0), L_s(1)\}.$$

Ya que cada $L_s^*(a^*)$ corresponde a la decisión a^* que posee la mínima pérdida esperada en el nodo s, la cantidad

$$D(s_1, ..., s_N) = \sum\nolimits_{s \in R} L_s^*(a^*(s))$$

es la mínima pérdida esperada posible para el conjunto de nodos R, pues la decisión se toma independientemente en cada nodo. Esta suma puede escribirse de la siguiente manera

$$D(s_1, ..., s_N) = \sum_{\{s \in R : L_s(0) < L_s(1)\}} L_s^*(0) + \sum_{\{s \in R : L_s(1) \leq L_s(0)\}} L_s^*(1).$$

La zona para proteger será entonces $A = \{s \in R : L_s(1) \leq L_s(0)\}$, y la cantidad $D(s_1, ..., s_N)$ puede escribirse como

$$D(s_1, ..., s_N) = \sum_{s \notin A} L_s^*(0) + \sum_{s \in A} L_s^*(1).$$

Ya que $L_s^*(0) = L_s(0)$ para $s \notin A$ y $L_s^*(1) = L_s(1)$ para $s \in A$, se obtiene

$$D(s_1, ..., s_N) = \sum_{s \notin A} L_s(0) + \sum_{s \in A} L_s(1).$$

Sustituyendo las expresiones dadas en (2.2) se obtiene que

$$D(s_1,...,s_N) = \sum_{s\notin A}\sum_{i=1}^{I} w_i \left\{x_i(s) - p_i(s)x_i(s) + p_i(s)z_i(s)\right\} + \sum_{s\in A}\sum_{i=1}^{I} w_i \left\{y_i(s) - p_i(s)y_i(s) + p_i(s)t_i(s)\right\},$$

que es precisamente $E[L(\mathbf{U}, A)]$. Ya que la cantidad $D(s_1,...,s_N)$ representa la mínima pérdida posible al considerar la decisión en cada uno de los nodos, se tiene que $D(s_1,...,s_N) = \arg\min_{A\in\mathcal{A}}\{L(A)\}$, por lo que A^* es el conjunto de nodos en donde la decisión es proteger. ■

2.3 Buscando una Región: Imponiendo Restricciones

En la práctica, es natural suponer que la región que se propone para proteger se encuentre sujeta a satisfacer algunas condiciones, las cuales surgen por razones prácticas y/o biológicas. En esta sección se consideran dos restricciones comunes en el contexto del problema de proteger especies. La primera de ellas se denomina *Restricción por Presupuesto*, que como se explica en la sección 2.3.1, surge debido a que en la práctica se cuenta con un presupuesto limitado destinado a proteger especies. La segunda restricción que se considera se denomina *Restricción por Conexidad*, la cual surge por razones principalmente biológicas y da lugar a un debate con respecto a la forma que debe tener una región que se propone para proteger (Sección 2.3.2). En las siguientes secciones se presenta la forma de involucrar las restricciones en el proceso de tomar la decisión, es decir, en la determinación de la región que se propondrá para proteger.

2.3.1 Restricción por Presupuesto

En proyectos de protección y/o manejo de la biodiversidad es común que se cuente con un presupuesto determinado, el cual limita el número de nodos que pueden adquirirse para la protección de especies. Sea B el presupuesto con que se cuenta, que en esta tesis se supone dado en pesos mexicanos. Cada nodo se supone asociado a algún indicador de su valor, que puede ser monetario, o bien, que es determinado utilizando alguna información adicional con que se cuente. Sea $c(s)$ el valor asignado al nodo $s \in R$. Así, $c(s)$ representa la cantidad que se deberá invertir si el nodo s es seleccionado para ser protegido. En general no se cuenta con una definición precisa de lo que representa la cantidad $c(s)$. Este costo puede ser, por ejemplo, el valor promedio del terreno representado por dicho nodo, o bien, cualquier indicador pertinente de su valor. Más aún, $c(s)$ puede representar el costo involucrado en comprar, cercar y administrar la porción de tierra representada por s. La cantidad $c(s)$

puede utilizarse para definir alguna de las cantidades $x(s)$, $y(s)$, $z(s)$ y $t(s)$ de las Tablas 1 y 2, como se hace en la Sección 2.5, pero en general estas cantidades pueden definirse sin considerar la cantidad $c(s)$. En lo sucesivo, la cantidad $c(s)$ se supone dada en pesos mexicanos.

Si se postula que se propondrá la región $A \in \mathcal{A}$ para proteger, deberá invertirse un total de $\sum_{s\in A} c(s)$ pesos, cantidad que debe ser menor o igual que B. Así, una región $A \in \mathcal{A}$ que se proponga como candidata a ser protegida deberá satisfacer $\sum_{s\in A} c(s) \leq B$. Esta restricción puede involucrarse considerando el espacio de acciones (decisiones) definido por $\mathcal{A}' = \{A \in \mathcal{A} : \sum_{s\in A} c(s) \leq B\}$. Al utilizar $\mathcal{A}'$ como conjunto de búsqueda, una región cuyo costo exceda el presupuesto B no será considerada como candidata para ser protegida.

Aún cuando el espacio de acciones $\mathcal{A}'$ posee en general menos elementos que $\mathcal{A}$, no existe garantía de que la búsqueda pueda realizarse de manera exhaustiva, por lo que es necesario considerar algún algoritmo de búsqueda (Sección 2.4) para obtener una solución. En lo sucesivo se considera $\mathcal{A}'$ como el espacio de búsqueda.

2.3.2 Restricción por Conexidad

En el ámbito de protección de especies existe un debate con respecto a la fragmentación de la región que debe protegerse. Este debate se expresa como sigue: dada un área fija Y, ¿es preferible proteger una sóla región con área Y, o proteger varias regiones de menor tamaño que en conjunto posean área Y? Este debate se conoce como el debate SLOSS (*Single Large Or Several Small*) y para ambas opciones existen argumentos a favor y en contra tanto biológicos (Baz y García-Boyero, 1996) como económicos (Drechsler y Wätzold, 2001).

En esta sección se propone una forma de proceder que permite obtener regiones a proteger a la luz del debate SLOSS y observar la diferencia en pérdida esperada de una región fragmentada y una que no lo sea. Para esto debe ser posible medir el grado de fragmentación que posea una región A. Una manera de hacerlo es mediante el perímetro de A, que en el contexto que nos ocupa, se define como la longitud del contorno de la superficie determinada por los nodos que conforman la región A que se considere. Es claro que una región altamente fragmentada posee mayor perímetro que una región que no lo sea. Si se supone que cada nodo $s \in R$ representa un cuadro de una unidad de longitud por lado, un conjunto A conformado por 9 nodos dispersos sobre R producirá un perímetro de 36 unidades, en tanto que si los 9 nodos conforman un cuadrado de 3×3, el perímetro de A será de 12 unidades. Denótese por $H(A)$ al perímetro de la región A.

La estrategia que aquí se adopta para involucrar el perímetro de tal manera que tenega el efecto deseado en la región a proteger que se obtendrá es considerar a $H(A)$ como un término adicional en la función de pérdida que se utilice. La influencia que este término tendrá sobre la solución será modulada por medio de un parámetro β asociado a $H(A)$. De esta manera,

la función de pérdida que se propone es

$$L_\beta(\mathbf{U}, A) = L(\mathbf{U}, A) + \beta H(A), \tag{2.6}$$

donde $\beta \in [0, \infty)$. La región que se propondrá para proteger se define como

$$A^*_\beta = \arg\min_{A \in \mathcal{A}'} \{L_\beta(A)\},$$

donde $L_\beta(A) = E\left[L_\beta(\mathbf{U}, A)\right]$ es la pérdida esperada correspondiente a tomar la decisión de proteger la región $A \in \mathcal{A}'$. Aplicando el valor esperado a la función (2.6) y utilizando las cantidades $x(s)$, $y(s)$, $z(s)$ y $t(s)$ de la Tabla 2, se obtiene

$$\begin{aligned} L_\beta(A) \;=\; & \sum_{s \notin A} \sum_{i=1}^{I} w_i \left\{x_i(s) - p_i(s)x_i(s) + p_i(s)z_i(s)\right\} + \\ & \sum_{s \in A} \sum_{i=1}^{I} w_i \left\{y_i(s) - p_i(s)y_i(s) - p_i(s)t_i(s)\right\} + \beta H(A). \end{aligned} \tag{2.7}$$

El parámetro β podrá ser manipulado por el usuario con el fin de observar diferentes configuraciones (en el campo geográfico) de la región que se propondrá para proteger. Esta forma de proceder permite explorar diferentes posibilidades en cuanto a la conexidad de la región que se obtenga, considerando el debate SLOSS. Si se postula $\beta = 0$, no se considera el perímetro de la región, por lo que se podrá obtener como solución una región altamente fragmentada. A medida que el valor β se incrementa se impone mayor importancia a la preferencia por zonas no fragmentadas. Por otro lado, si se utilizan sucesivamente los valores $\beta = \beta_1$ y $\beta = \beta_2$ en la expresión (2.6) será posible comparar las correspondientes soluciones $A^*_{\beta_1}$ y $A^*_{\beta_2}$ mediante la pérdida esperada $L_{\beta_i}(A^*_{\beta_i})$, $i = 1, 2$, de las mismas, como se propone en la Sección 2.6.

Ya que el parámetro β puede asumir cualquier valor no negativo, deberá experimentarse con éste a fin de detectar los valores que producen cambios en la región que se propondrá para proteger. La magnitud de los cambios en β que produzcan cambios en la región que se obtenga, depende de las unidades de $L(\mathbf{U}, A)$. La idea de introducir el parámetro β es establecer de manera cualitativa la conveniencia de considerar regiones más conexas, en comparación con el aumento relativo de su pérdida esperada. Esto se espera logre de manera dinámica modular el parámetro β en un rango adecuado.

La idea de introducir un parámetro manipulable en una función a minimizar fue utilizada por McDonnell *et al.* (2002) en el problema de diseñar reservas ecológicas espacialmente conexas. Utilizando el enfoque de programación lineal definen la expresión a minimizar como función del área y el perímetro de la región. En su trabajo se introduce una cantidad similar a la β de la expresión (2.6) para ponderar la magnitud del perímetro de una región

en la función a minimizar, con respecto a la magnitud de su área. Sin embargo, esta forma de proceder no se motiva a partir del debate SLOSS.

2.4 Búsqueda de la Solución

El problema de encontrar un elemento que minimice una función cuando el espacio de búsqueda posee un número grande de elementos ha sido abordado mediante algoritmos de búsqueda no exhaustivos, que son fáciles de implementar y que, en general, producen soluciones adecuadas. Existen dos algoritmos iterativos ampliamente utilizados para este fin, los cuales se describen a continuación. Sea $L_\beta(A)$ la función que se desea minimizar, en nuestro caso, la función de pérdida esperada definida para conjuntos de nodos en $\mathcal{A}'$ (o en $\mathcal{A}$). Ya que en esta sección se propone la manera de buscar una región para proteger dado un valor β fijo, la funcion a minimizar se denota simplemente por $L(A)$, suprimiendo el subíndice β.

El primer algoritmo se conoce como *algoritmo greedy* (McDonnell *et al.* 2002). Sea $A^{(t)} \in \mathcal{A}'$ la solución con que se cuenta en la iteración t. En la iteración $t+1$ se propone encontrar una solución vecina A' de $A^{(t)}$ que posea menor pérdida esperada. Una solución vecina se define como el conjunto de nodos (la solución) $A^{(t)}$ al que se añade un solo nodo como parte de la solución. De esta manera, la región A' poseerá un nodo más que $A^{(t)}$. En la iteración $t+1$ se consideran todas las soluciones vecinas posibles de $A^{(t)}$ y aquella que produzca la mayor disminución en $L(A)$ con respecto a $L(A^{(t)})$ se considera la nueva solución, es decir, $A^{(t+1)}$. En cada iteración se repite este proceso hasta que ninguna solución vecina produzca disminución en $L(A)$. Con este procedimiento, en la primera iteración se evaluarán N posibles soluciones, cada una de las cuales constará de un solo nodo. En la segunda iteración se evaluarán $N-1$ posibles soluciones. Cada una de ellas constará de dos nodos, uno de los cuales será el nodo que produjo la menor pérdida esperada en la primera iteración, y así sucesivamente. En caso de que dos o más soluciones vecinas produzcan la misma disminución en la función objetivo, la nueva solución se selecciona al azar entre dichas soluciones.

Es claro que en este algoritmo un nodo que en alguna iteración ha sido seleccionado como parte de la solución formará parte de la solución final que se obtenga. En otras palabras, una decisión tomada no se puede revocar. Este hecho produce que, en algunas situaciones, la solución que se obtenga corresponda a un mínimo local y no a un mínimo global (McDonnell *et al.* 2002). Para solucionar esto es necesario que el algoritmo sea capaz de "ir hacia atrás" en el proceso de búsqueda, con el fin de escapar de mínimos locales. En el contexto del problema que se aborda, esto significa que debe ser posible suprimir nodos de una solución que se tenga, lo que es posible por medio del algoritmo *Simulated Annealing* (Bertsimas y Tsitsiklis, 1993).

De manera similar que el algoritmo greedy, la idea del algoritmo Simulated Annealing consiste en proponer, en cada iteración, una solución vecina A' de $A^{(t)}$. En este algoritmo

una solución vecina en la iteración $t+1$ se obtiene seleccionando al azar un nodo $s \in R$. Si resulta que $s \in A^{(t)}$, se postula que $s \notin A'$, en cuyo caso la solución vecina poseerá un nodo menos que $A^{(t)}$. Si por el contrario resulta que $s \notin A^{(t)}$, entonces se postula que $s \in A'$. Así, la solución vecina poseerá un nodo más que $A^{(t)}$. La solución vecina A' será aceptada como nueva solución, es decir se postulará $A^{(t+1)} = A'$, con probabilidad

$$\rho = \min\left\{1,\, e^{-\frac{1}{T(t)}\left\{L(A')-L(A^{(t)})\right\}}\right\},$$

donde $T(t)$ es una función no creciente conocida como esquema de "enfriamiento". Para t fijo, la cantidad $T(t)$ se conoce como *temperatura* al tiempo t. En cada iteración el parámetro $T(t)$ decrece de acuerdo con un esquema previamente establecido. De acuerdo con este algoritmo, si la solución vecina A' produce un aumento en el valor de $L(A)$, ésta se acepta como nueva solución, es decir $A^{(t+1)} = A'$, con probabilidad $\exp\{-[L(A') - L(A^{(t)})]/T(t)\}$. Si por el contrario A' produce una disminución en $L(A)$, ésta se acepta como solución actual con probabilidad 1. Con esta forma de proceder una solución vecina puede estar conformada por un número menor de nodos que la solución actual, lo que no ocurre en el algoritmo greedy. Después de un número determinado de iteraciones, o bien, cuando $T(t)$ alcance un valor previamente establecido, la región A^* que se propondrá será la región con que se cuente en la última iteración.

Existen diversas propuestas para definir el esquema de enfriamiento (Cohn y Fielding, 1998). En este capítulo se utiliza el *esquema de enfriamiento logarítmico*, en el que se define $T(t) = d/\ln(1+t)$, donde d es una constante positiva que se relaciona con la cantidad de "energía" necesaria para escapar de un mínimo local. Este esquema de enfriamiento garantiza que el algoritmo Simulated Annealing converge a un mínimo global (Geman y Geman, 1984).

En McDonnell *et al.* (2002) se comparan los resultados obtenidos al utilizar los dos algoritmos descritos para el problema de diseñar reservas naturales espacialmente conexas. Después de un extenso estudio concluyen que, usualmente, el algoritmo Simulated Annealing produce mejores resultados (en términos de la función de pérdida) que el algoritmo greedy, aunque éste último es más rápido en obtener una solución.

Para explotar las ventajas que ofrecen ambos algoritmos, en esta tesis se propone utilizarlos de manera conjunta, como se explica enseguida. Ya que el algoritmo Simulated Annealing requiere de una región A_0 inicial, la idea que se explota consiste en utilizar como A_0 al resultado que se obtenga con el algoritmo greedy. Aunque es posible implementar el algoritmo greedy utilizando la restricción considerada para el presupuesto, se propone encontrar la solución greedy sin considerar esta restricción. Esto puede producir que el costo de la región A_0 exceda el presupuesto con que se cuenta. Para solventar esto basta seleccionar al azar, de la solución greedy obtenida, tantos nodos como sea posible adquirir con el presupuesto B. Ya que la región A_0 se utiliza como solución inicial para el algoritmo Simulated Annealing, el procedimiento de aleatorización que se utilice para seleccionar los nodos de A_0 no es relevante. La consideración de que la solución inicial se obtiene utilizando el algoritmo greedy produce, en general, que la convergencia del algoritmo Simulated Annealing se alcance en un número menor de iteraciones, debido a que la solución greedy proporciona una solución que, en el peor de los casos, será una solución subóptima, ya sea por corresponder a un mínimo local, o por haberse suprimido nodos de una solución óptima del algoritmo greedy para no exceder el presupuesto con que se cuenta.

Al implementar un algoritmo como el descrito en los párrafos anteriores se requiere cierto monitoreo para verificar que éste proporciona soluciones adecuadas para el problema que se aborda. El monitoreo que se realiza incluye ejecutar el algoritmo utilizando diferentes soluciones iniciales (diferentes conjuntos de nodos como solución inicial). Esto permite observar si el algoritmo produce soluciones similares en cada ejecución, o bien, soluciones cuya variación en la función de pérdida sea pequeña después de un número determinado de iteraciones. Si esto no ocurre, será evidencia de que el algoritmo no esta realizando una busqueda adecuada, o bien, que existe más de un conjunto de nodos que pueden considerarse óptimos para proteger. En los ejercicios de simulación (Sección 2.6) y en las aplicaciones (Capítulo 4), se encontró una solución óptima después de un número suficiente de iteraciones.

2.5 Caso Particular

Al abordar el problema de encontrar una región para proteger especies reales, existen consideraciones que permiten proponer valores particulares para las cantidades involucradas en las Tablas 1 y 2, introducidas en la Sección 2.1. A su vez, estos valores definen expresiones particulares para las correspondientes funciones de pérdida esperada. En esta sección se realizan consideraciones particulares para definir los valores $x(s)$, $y(s)$, $z(s)$ y $t(s)$ para una especie.

Sea s el nodo en el que se desea tomar la decisión de proteger o no proteger. Cuando se toma la decisión correcta de no proteger el nodo (la especie no se encuentra presente en s), no se incurre en pérdida alguna, por lo que es sensato postular $L_s(0,0) = 0$, es decir $x(s) = 0$. Si por el contrario se decide no proteger el nodo pero la especie se encuentra presente, se postula $L_s(1,0) = z(s)$. De acuerdo con esta consideración, la cantidad $z(s)$ se interpreta como el costo (pérdida) que representa perder a la especie en el nodo s.

Si se decide proteger el nodo s cuando la especie no se encuentra presente, por el hecho de protegerlo se invertirá la cantidad $c(s)$. En este caso esta cantidad se considera una pérdida, pues se estará utilizando de manera inadecuada. Así, se postula $L_s(0,1) = c(s)$, es decir, $y(s) = c(s)$. Finalmente, cuando se toma la decisión correcta de proteger el nodo (la especie se encuentra presente en él), la cantidad $c(s)$ invertida no se considera una pérdida, pues se estará utilizado de manera correcta. Más aún, tomar la decisión correcta representa una ganancia (pérdida negativa) debida a la protección del nodo en favor de proteger a la especie. Por esta consideración es sensato postular que la ganancia que se obtiene al tomar

esta decisión es precisamente la cantidad que se perdería en caso de decidir erróneamente no proteger el nodo, es decir, se postula $L_s(1,1) = -z(s)$.

El valor $z(s)$ depende de la especie en cuestión y no de la posición geográfica del nodo considerado, por lo que se postula $z(s) = z$, suprimiendo la dependencia de s. En otras palabras, el costo debido a perder a la especie (o sea el valor de la especie) se supone constante sobre la región de estudio. De esta forma, z es el costo biológico de la especie.

La función de pérdida que se obtiene para cada especie después de realizar las consideraciones indicadas se resume en la Tabla 3, en la que se asume que las cantidades z y $c(s)$ son comparables, es decir, se encuentran medidas en la misma unidad (pesos, por ejemplo). En el Capítulo 3 se propone una forma de postular dichos valores utilizando información del experto e información adicional.

Tabla 3: Función de pérdida para la especie en $s \in R$ (caso particular).

$\Theta_s \backslash \mathcal{A}_s$	$a(s)=0$	$a(s)=1$
$u(s)=0$	0	$c(s)$
$u(s)=1$	z	$-z$

Para el caso de tomar la decisión en un nodo considerando I especies, denotando por z_i al valor biológico correspondiente a la i-ésima especie, se tendrá por sustitución que la pérdida esperada para cada acción posible en $\mathcal{A}_s$ es

$$\begin{aligned} L_s(0) &= \sum_{i=1}^{I} w_i p_i(s) z_i, \quad \text{y} \\ L_s(1) &= c(s) - c(s)\sum_{i=1}^{I} w_i p_i(s) - \sum_{i=1}^{I} w_i p_i(s) z_i, \end{aligned} \tag{2.8}$$

respectivamente. La decisión de proteger el nodo s se tomará si $c(s)\left\{1 - \sum_{i=1}^{I} w_i p_i(s)\right\} \leq 2\sum_{i=1}^{I} w_i p_i(s) z_i$. Si en particular $I = 1$ (una especie), se decidirá proteger el nodo s si $c(s)\{1 - p(s)\} \leq 2p(s)z$, es decir, si $p(s) \geq c(s)\left[c(s) + 2z\right]^{-1}$. En este caso la zona para proteger estará conformada por los nodos en los que la probabilidad de presencia de la especie sea mayor que el umbral dado por la cantidad $c(s)\left[c(s) + 2z\right]^{-1}$.

Suponiendo que la decisión se toma de manera independiente en cada nodo, la zona que se propone para ser protegida es

$$A^* = \left\{ s \in R : c(s)\left(1 - \sum_{i=1}^{I} w_i p_i(s)\right) \leq 2\sum_{i=1}^{I} w_i p_i(s) z_i(s) \right\}. \tag{2.9}$$

Cuando se considera el problema desde el punto de vista de proteger una región sin

considerar restricciones sobre la solución, la pérdida esperada que se obtiene es

$$L(A) = \sum_{s \notin A}\sum_{i=1}^{I} w_i p_i(s) z_i + \sum_{s \in A}\left\{ c(s) - c(s)\sum_{i=1}^{I} w_i p_i(s) - \sum_{i=1}^{I} w_i p_i(s) z_i \right\}. \tag{2.10}$$

Finalmente, cuando se considera la restricción por conexidad en la función de pérdida, la pérdida esperada es

$$L_\beta(A) = \sum_{s \notin A}\sum_{i=1}^{I} w_i p_i(s) z_i + \sum_{s \in A}\left\{ c(s) - c(s)\sum_{i=1}^{I} w_i p_i(s) - \sum_{i=1}^{I} w_i p_i(s) z_i \right\} + \beta Z(A). \tag{2.11}$$

mediante la cual la zona que se propondrá para proteger será $A^*_\beta = \arg\min_{A \in \mathcal{A}'}\{L_\beta(A)\}$.

2.6 Estudio de Simulación

En esta sección se presenta un experimento de simulación que permite explorar la metodología que se propone para encontrar regiones para proteger. La función de pérdida que se considera es la propuesta en la expresión (2.6), en la que se usan los valores particulares para las cantidades involucradas de acuerdo con las ideas de la Sección 2.5. La pérdida esperada que se debe minimizar está dada por la expresión (2.10). La región de estudio que se considera es la península de Yucatán, con la retícula R descrita en la Sección 1.5 (761 nodos). Para fines de este ejercicio de simulación se procedió a definir el costo de cada nodo mediante la función

$$c(s) = 0.2 * \sqrt{|lat(s) - 20.5|^3 + |long(s) - 20.5|^3}, \tag{2.12}$$

donde $(lat(s), long(s))$ representan las coordenadas geográficas de los nodos. Esta función se seleccionó de manera arbitraria, observando que los costos que se generan para los nodos posean cierta estructura espacial. En la Figura 2-6(a) se observa el mapa de costos que resultó de la expresión (2.12), y que es la que se utiliza en este estudio.

Se consideran tres especies sobre la región de interés. La probabilidad de presencia de cada una de ellas en cada nodo se obtuvo mediante la expresión (1.13), para la cual se postularon los elementos $\boldsymbol{\mu}_1 = \{2, 4, 3.5\}$, $\boldsymbol{\mu}_2 = \{4, 3, 3\}$, $\boldsymbol{\mu}_3 = \{4, 4, .5\}$ y la matriz simétrica $\mathbf{A}$ determinada por los valores $a_{1,1} = a_{2,2} = a_{3,3} = 1$, $a_{1,2} = .6$, $a_{1,3} = .3$ y $a_{2,3} = .1$ para las tres especies. Los mapas de probabilidades de presencia que se obtienen con estos elementos se presentan en las Figuras 2-1(a), (b) y (c), respectivamente. En esos mapas se observan las probabilidades $p_i(s)$, $i = 1, 2, 3$, que se utilizan para encontrar la región a proteger. Para la primera especie se observa que la zona de mayor probabilidad de presencia es conformada principalmente por dos regiones disjuntas (Figura 2-1(a)). Se observa que una de las regiones

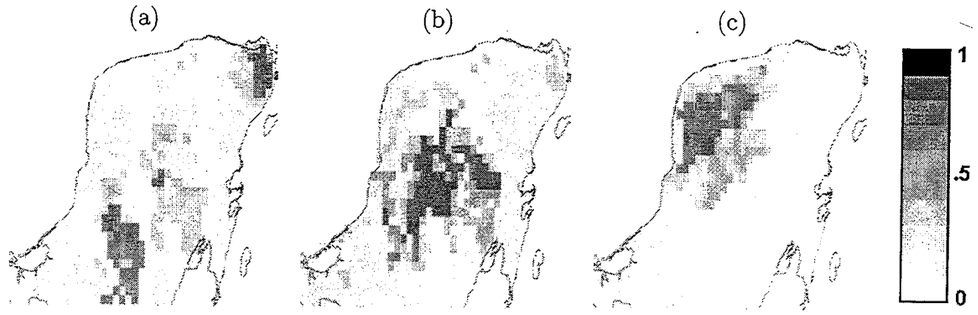

Figura 2-1: Mapas de probabilidades de presencia generados mediante (a) $\mu_1 = \{2, 4, 3.5\}$, (b) $\mu_2 = \{4, 3, 3\}$ y (c) $\mu_3 = \{4, 4, .5\}$.

que posee alta probabilidad de presencia se localiza en una zona que puede considerarse no costosa, la cual se localiza hacia el centro de la Península (Figura 2-6(a)). La otra región de alta probabilidad de presencia se localiza en una zona que puede considerarse costosa (al noreste de la Península). Para la segunda especie (Figura 2-1(b)) se observa que la zona de alta probabilidad de presencia se localiza en la zona considerada no costosa, mientras que la zona de alta probabilidad de presencia para la tercera especie (Figura 2-1(c)) se localiza en una zona que posee costo intermedio.

Con base en las cantidades de las que depende la función de pérdida esperada dada en la expresión (2.11) se detectó la existencia de factores que determinan diferentes escenarios que pueden explorarse. Para cada escenario se observa el funcionamiento de la metodología bajo ciertas características que pueden presentarse en la realidad. Los factores detectados se describen en los siguientes párrafos.

Dada la cantidad $c(s)$ para cada nodo de la retícula, sea $C_T = \sum_{s \in R} c(s)$, que representa el costo de toda la región. El primer factor que se considera es el presupuesto con que se cuenta. La cantidad C_T se considera como referencia para determinar los presupuestos que se postularán en el estudio. Para este ejercicio se obtuvo $C_T = 6,283.8$ pesos, que se encontró al sumar las cantidades $c(s)$ generadas mediante la expresión (2.12). Se propone considerar como presupuesto las cantidades $.0025C_T$ y $.01C_T$, es decir, $B = 15.71$ pesos y $B = 62.82$ pesos. Los valores .0025 y .01 se seleccionaron de tal manera que se refleje un hecho que comúnmente se presenta en la práctica: se cuenta con un presupuesto restringido para realizar acciones para proteger. La cantidad $B = 15.71$ se denomina *presupuesto bajo*, mientras que la cantidad $B = 62.82$ se denomina *presupuesto alto*.

El segundo factor que se considera se encuentra determinado por el vector $\{z_1, z_2, z_3\}$. Recordando, la cantidad z_i se interpreta como el costo que representa perder a la especie en la región de estudio, es decir, el valor biológico asignado a la especie. Se consideran cuatro conjuntos de valores, los cuales determinan cuatro niveles para el vector $\{z_1, z_2, z_3\}$. Para el primer nivel se postula que el costo de perder a cada especie es el mismo. En este caso se

determinó $z_1 = z_2 = z_3 = 10$. Los otros tres niveles se obtienen al considerar, por turnos, que una de las especies posee una pérdida menor asociada a perderla que las otras dos especies. Se consideró $z_i = 5$ y $z_j = 15$ para $j \neq i$, $i = 1, 2, 3$.

El tercer factor a considerar es la postulación del vector $\{w_1, w_2, w_3\}$. Para éste se consideran cuatro niveles. En el primero se postula que las tres especies poseen la misma importancia para ser protegidas, es decir, $w_i = 1/3$, $i = 1, 2, 3$. Para los otros tres niveles se considera, por turnos, una de las especies como más importante para proteger, mientras que las otras dos se consideran igual de importantes. Así, se postula $w_i = .6$, $i = 1, 2, 3$ y $w_j = .2$, $j \neq i$. Los valores .6 y .2 se seleccionaron arbitrariamente.

El último factor a considerar es el parámetro β, involucrado en la función de pérdida esperada dada por la expresión (2.11), que es la que se utiliza en esta sección. Para este parámetro se consideran los valores 0 y 2, que fueron determinados después de inspeccionar diversos valores de β, como se sugiere en la Sección 2.3. Desde luego, $\beta = 0$ significa que la restricción por conexidad no es tomada en cuenta. En este caso, $\beta = 2$ representa un valor alto en relación a las unidades de $L(\mathbf{U}, A)$ y permite observar regiones no fragmentadas para proteger.

Los factores considerados y sus correspondientes niveles se resumen en la Tabla 4. Estos niveles permiten inspeccionar un total de 64 escenarios. Los resultados se resumen en los siguientes párrafos.

Tabla 4 : Factores considerados para la simulación y los correspondientes niveles.

Factor		Niveles		
B	15.71 pesos	62.82 pesos		
$\{z_1, z_2, z_3\}$	$\{10, 10, 10\}$	$\{5, 15, 15\}$	$\{15, 5, 15\}$	$\{15, 15, 5\}$
$\{w_1, w_2, w_3\}$	$\{.333, .333, .333\}$	$\{.6, .2, .2\}$	$\{.2, .6, .2\}$	$\{.2, .2, .6\}$
β	0	2		

Con respecto al presupuesto, es claro que cuando se cuenta con una cantidad restringida se limita el número de nodos que pueden seleccionarse para proteger. Se observa que cuando se cuenta con un presupuesto restringido ($B = 15.71$ pesos), las zonas que se proponen para proteger se encuentran conformadas en su mayoría por nodos localizados en la zona donde los nodos son menos costosos. Cuando se cuenta con un presupuesto mayor ($B = 62.82$ pesos), se incrementa el número de nodos localizados en la zona costosa que se consideran parte de la solución. La diferencia en el número de nodos que posee la zona que se propone para proteger puede observarse al comparar las Figuras 2-2(a) y 2-4(a), 2-2(c) y 2-4(c), y en general, al comparar cualquier pareja de gráficas en las Figuras 2-2 y 2-4 ó Figuras 2-3 y 2-5, considerando el mismo inciso para cada una de ellas.

Con el fin de estudiar el efecto de las cantidades $\{z_1, z_2, z_3\}$ en la zona a proteger, se procedió a observar las zonas resultantes para los cuatro niveles de este factor que se presen-

tan en la Tabla 4, con el presupuesto B y el parámetro β fijos, y postulando las cantidades $w_i = 1/3$, para $i = 1, 2, 3$. Si se considera el presupuesto bajo ($B = 15.71$) y $\beta = 0$, al comparar las Figuras 2-2(a), 2-2(i), 2-3(e) y 2-3(m) se observa que la especie i para la que se postula $z_i = .6$, tiende a poseer mayor cantidad de nodos pertenecientes a la correspondiente zona de alta probabilidad de presencia. En el caso de postular el presupuesto alto, es decir $B = 62.82$, este hecho se observa con mayor claridad, como se ilustra en las Figuras 2-4(b), 2-4(j), 2-5(f) y 2-5(n).

Para estudiar el efecto que las cantidades $\{w_1, w_2, w_3\}$ producen en la región que se propondrá para proteger, se procede a fijar el valor del presupuesto B y del parámetro β, y utilizar $z_i = 10$, $i = 1, 2, 3$. Las cantidades $\{w_1, w_2, w_3\}$ se fijan de acuerdo con los niveles descritos en la Tabla 4. Se observa que la especie i a la cual se asignó mayor importancia, postulando $w_i = .6$, posee mayor influencia en la región a proteger que se obtiene. Para el caso en el que se considera el prespuesto $B = 15.71$ y $\beta = 0$, compare las Figuras 2-2(b), 2-2(d), 2-2(f) y 2-2(h). En la Figura 2-2(d) se observa que la zona propuesta para proteger posee más nodos correspondientes a la zona de alta probabilidad de presencia de la primera especie. De la misma manera, cuando se postula para la segunda especie el valor $w_2 = .6$, la zona para proteger posee más nodos correspondientes a la zona de alta probabilidad de presencia para esta especie (Figura 2-2(f)). En la Figura 2-2(h) se observa el mismo efecto, al postular la tercera especie como más importante para proteger ($w_3 = .6$). En el caso de considerar el presupuesto $B = 62.82$, este mismo hecho se observa al comparar las Figuras 2-4(b), 2-4(d), 2-4(f) y 2-4(h). Por ejemplo, si la tercera especie se considera más importante para proteger, es decir, se postula $w_3 = .6$, la zona que resulta para proteger posee más nodos de la zona real de alta probabilidad de presencia correspondiente a esta especie (Figura 2-4(h)).

Con respecto al parámetro β, se observa que los valores postulados permiten obtener diferentes niveles de fragmentación de la zona propuesta. El caso $\beta = 0$ corresponde a no considerar importante que la región que se proponga para proteger sea conexa. En este caso se observa que, en general, las regiones que se obtienen contienen mayor cantidad de nodos dispersos sobre la región de estudio. Observe por ejemplo las Figuras 2-2(a), (c), (i) o (k). El aspecto fragmentado de la región que se obtiene como solución disminuye con el aumento del valor de β. Así, en el caso de $\beta = 2$ se obtienen regiones más conexas como zonas para proteger. Observe las Figuras 2-2(b), (d), (j), (l) y compárelas con las Figuras 2-2(a), (c), (i), (k), respectivamente. Este hecho se observa también al comparar las Gráficas (a) y (b), (c) y (d), (e) y (f), (g) y (h), (i) y (j), (k) y (l), (m) y (n) y (o) y (p), en cualquiera de las Figuras 2-2, 2-3, 2-4 y 2-5.

La forma de proceder que se propone para abordar el problema de encontrar una zona para proteger permite comparar de manera cuantitativa las regiones que se obtuvieron, por medio de la pérdida esperada. Para valores fijos de $\{z_1, z_2, z_3\}$, $\{w_1, w_2, w_3\}$ y B, y considerando por turnos H valores $\beta_1, ..., \beta_H$, con $\beta_1 = 0$, se propone construir una tabla en la que se observe

el porcentaje de cambio en la pérdida esperada que la región obtenida con el valor β_h, $h > 2$, genera comparada con la pérdida generada con β_1. La comparación puede realizarse por medio de la Tabla 5, en donde se utiliza la notación $L_{\beta_1\beta_h} = L_{\beta_h}(A^*_{\beta_h})/L_{\beta_1}(A^*_{\beta_1})$, donde A^*_β denota la zona que se obtuvo para proteger utilizando el valor β fijo.

Tabla 5: Comparación de Pérdida Esperada

β	% Pérdida
β_1	0
β_2	$\left(L_{\beta_1\beta_2} - 1\right) 100\%$
$\vdots$	$\vdots$
β_H	$\left(L_{\beta_1\beta_H} - 1\right) 100\%$

Por ejemplo, para las regiones que se observan en las Figuras 2-2(a) y (b), las cuales corresponden a los valores $\beta = 0$ y $\beta = 2$ respectivamente, se obtiene la tabla

β	% Pérdida
0	0
2	2.97%

en tanto que para las regiones que se observan en las Figuras 2-2(c) y (d) se obtiene la tabla

β	% Pérdida
0	0
2	−1.72%

De la primera tabla se concluye que la región obtenida con el valor $\beta = 2$ representa una pérdida esperada que es 2.97% mayor que la región obtenida con $\beta = 0$. En este caso, se propondrá para proteger la región A^*_0. En la segunda tabla se observa que la región obtenida utilizando $\beta = 2$ representa una pérdida que es 1.72% menor (pérdida negativa) que la pérdida obtenida con $\beta = 0$. Por lo tanto, en este caso la región que se propondrá para proteger es A^*_2.

2.7 Discusión

La metodología introducida en este capítulo para encontrar una región que se propondrá para proteger no garantiza que la región que se obtenga poseerá nodos de alta probabilidad de presencia de cada una de las especies consideradas. Sin embargo, ya que para encontrar la región para proteger se considera (1) la probabilidad de presencia de cada especie en cada nodo, (2) un valor biológico para cada especie (dado por la cantidad z_i) y (3) un nivel de importancia que cada especie posee para ser protegida (dado por la cantidad w_i),

la zona que se propondrá para proteger estará conformada, principalmente, por nodos de alta probabilidad de presencia de aquellas especies cuya pérdida sea más costosa y/o que se consideren más importantes para ser protegidas.

Una aportación relevante de este capítulo es la posibilidad que se da a un usuario de observar diferentes soluciones, dependiendo de la importancia que se asigna a proteger regiones conexas. Esto se realizó considerando el parámetro β en la función de pérdida. Para determinar los valores de este parámetro que permiten observar diferentes niveles de conexidad en las regiones que se obtengan, deberá experimentarse con diferentes valores de β para determinar el rango de valores que producen cambios significativos en las regiones que se propongan para proteger.

Aunque en esta tesis se consideran solamente 2 restricciones para la región a proteger, la manera en la que éstas se consideran ejemplifica la forma en la que otras restricciones pueden involucrarse en el proceso de encontrar una región para proteger. Así, se podrá utilizar un espacio de acciones aún más restringido, o bien, considerar un término adicional en la función de pérdida. Por ejemplo, si además de considerar las dos restricciones mencionadas, se determina el número de nodos que se propondrá para proteger, denotado aquí por k, se procederá a restringir el espacio de soluciones y considerar ahora $\mathcal{A}'' = \{A \in \mathcal{A} : \sum c(s) \leq B$ y $|A| \leq k\}$.

La función de pérdida resumida en las Tablas 1, 2 y 3, requiere que las cantidades que las definen se encuentren medidas en las mismas unidades. Sin embargo, estas funciones pueden también utilizarse si las cantidades $x(s)$, $y(s)$, $z(s)$ y $t(s)$ se determinan de tal manera que reflejen alguna cantidad subjetiva. Por ejemplo, si un usuario no posee las herramientas necesarias para fijar las cantidades de la Tabla 3 siguiendo las ideas que se presentan en la Sección 3.2, el usuario podrá proceder a utilizar una función de pérdida subjetiva definida por la tabla

$\Theta \backslash \mathcal{A}$	$a(s) = 0$	$a(s) = 1$
$u(s) = 0$	a	b
$u(s) = 1$	c	d

En este caso, el experto deberá proporcionar las cantidades a, b, c y d, y en lugar de requerir que las cantidades sean comparables, se deberá cumplir que dichas cantidades guarden la debida relación en magnitud, de acuerdo con la interpretación que el experto asigne a cada una de ellas. Por ejemplo, si el experto considera que para una especie es más grave no proteger un nodo donde la especie se encuentra presente ($L(1,0)$), que proteger un nodo donde la especie no se encuenta presente ($L(0,1)$), entonces deberá postular las cantidades b y c de tal manera que $c > b$. Siguiendo razonamientos de este tipo, se procede a encontrar la región a proteger con base en esta función de pérdida subjetiva.

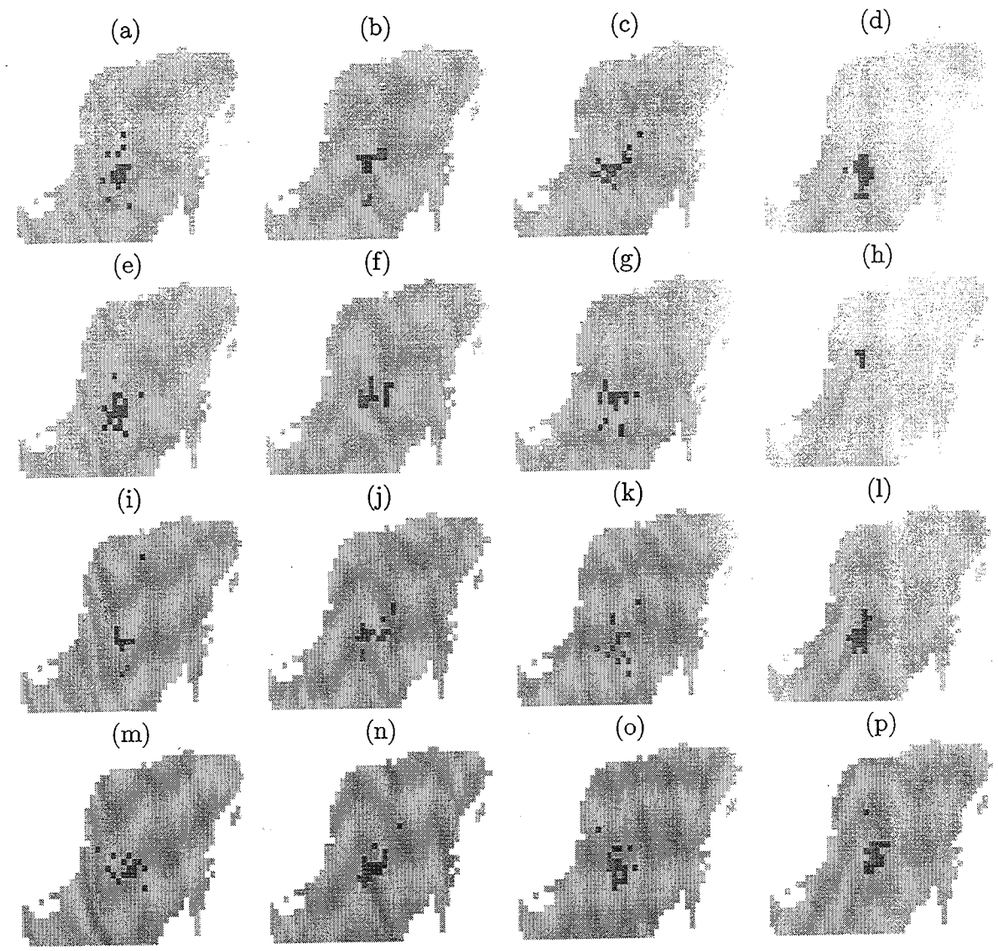

Figura 2-2: Zonas obtenidas para proteger. El primer dígito del nombre de cada figura indica el nivel del presupuesto considerado. El segundo dígito indica el nivel considerado para las cantidades $\{z_1, z_2, z_3\}$. El tercer dígito indica el nivel considerado para las cantidades $\{w_1, w_2, w_3\}$. El último dígito indica el nivel considerado para parámetro β. Los niveles se presentan en la Tabla 4. (a) e1111 (b) e1112 (c) e1121 (d) e1122 (e) e1131 (f) e1132 (g) e1141 (h) e1142 (i) e1211 (j) e1212 (k) e1221 (l) e1222 (m) e1231 (n) e1232 (o) e1241 (p) e1242.

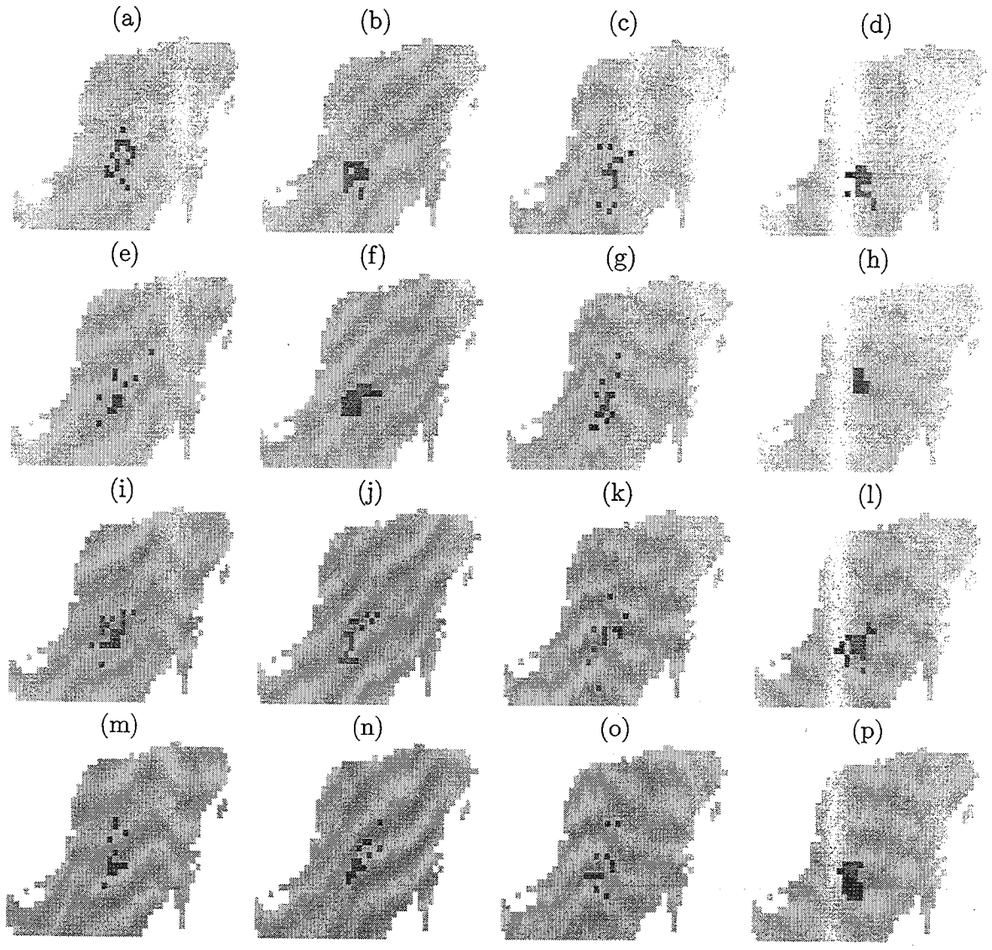

Figura 2-3: Zonas obtenidas para proteger. El primer dígito del nombre de cada figura indica el nivel del presupuesto considerado. El segundo dígito indica el nivel considerado para las cantidades $\{z_1, z_2, z_3\}$. El tercer dígito indica el nivel considerado para las cantidades $\{w_1, w_2, w_3\}$. El último dígito indica el nivel considerado para parámetro β. Los niveles se presentan en la Tabla 4. (a) e1431 (b) e1432 (c) e1441 (d) e1442 (e) e1311 (f) e1312 (g) e1321 (h) e1322 (i) e1331 (j) e1332 (k) e1341 (l) e1342 (m) e1411 (n) e1412 (o) e1421 (p) e1422.

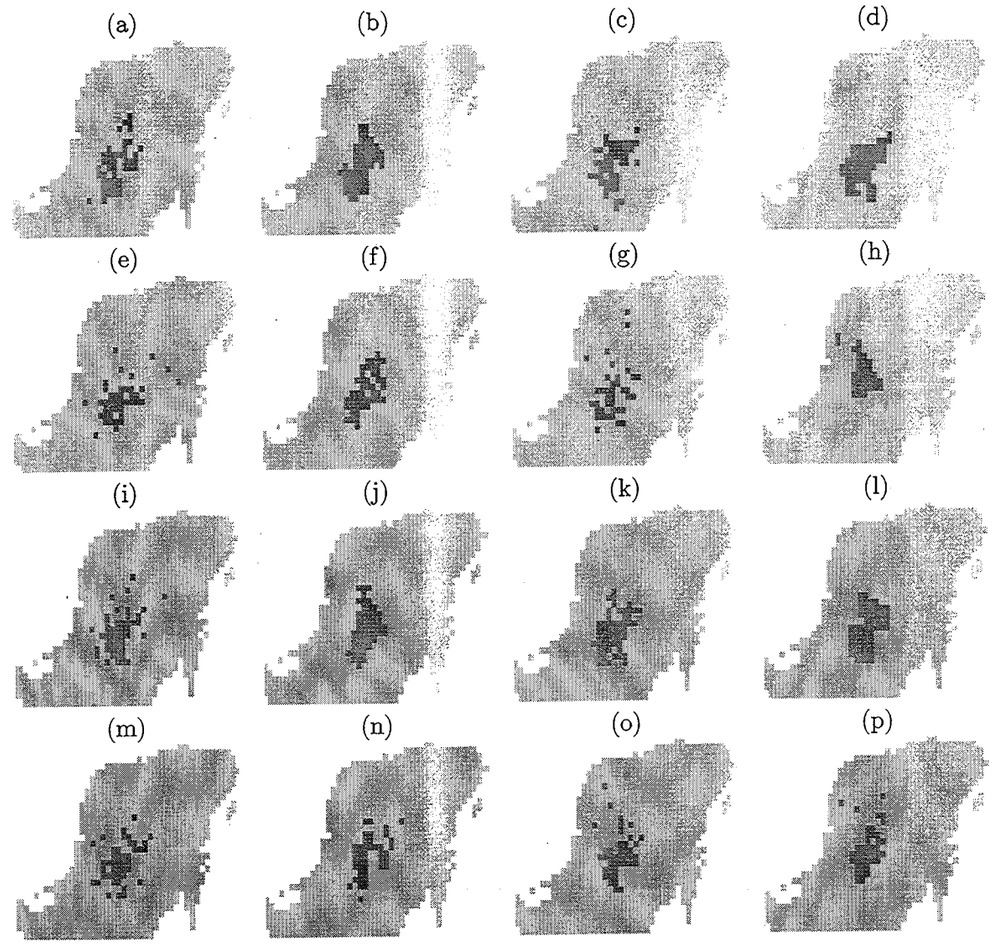

Figura 2-4: Zonas obtenidas para proteger. El primer dígito del nombre de cada figura indica el nivel del presupuesto considerado. El segundo dígito indica el nivel considerado para las cantidades $\{z_1, z_2, z_3\}$. El tercer dígito indica el nivel considerado para las cantidades $\{w_1, w_2, w_3\}$. El último dígito indica el nivel considerado para parámetro β. Los niveles se presentan en la Tabla 4. (a) e2111 (b) e2112 (c) e2121 (d) e2122 (e) e2131 (f) e2132 (g) e2141 (h) e2142 (i) e2211 (j) e2212 (k) e2221 (l) e2222 (m) e2231 (n) e2232 (o) e2241 (p) e2242.

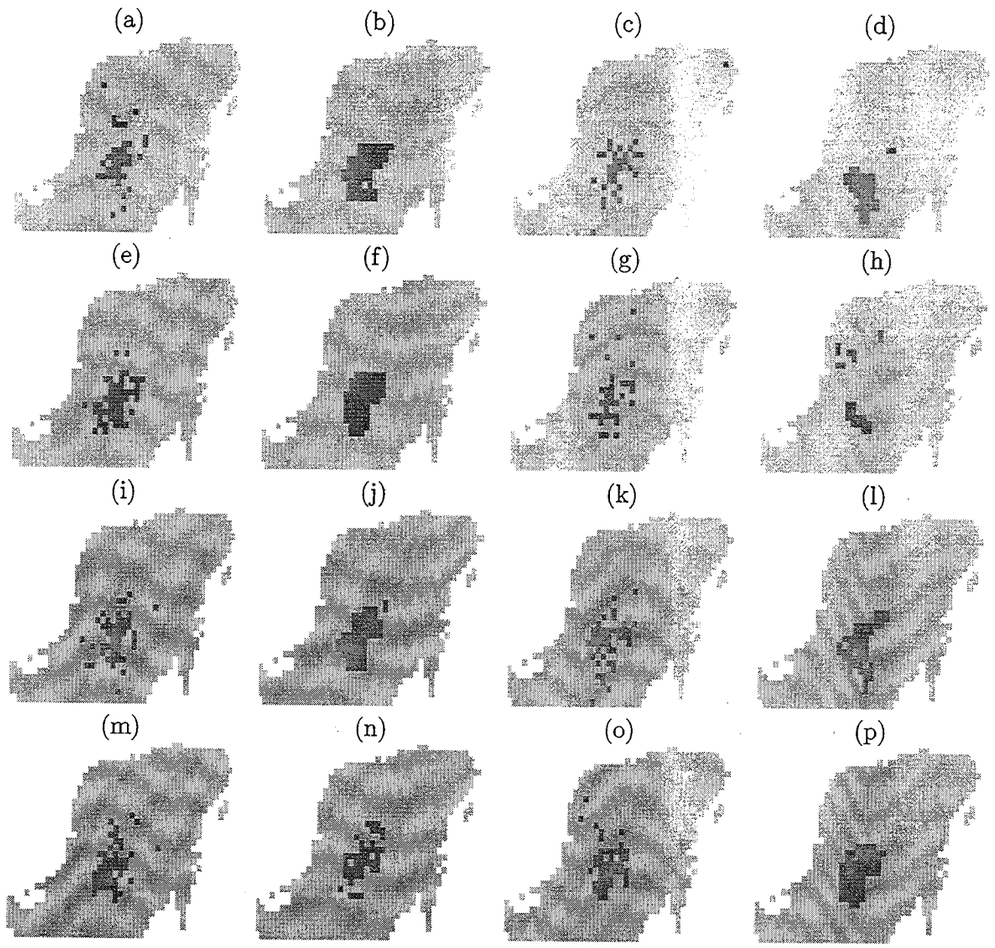

Figura 2-5: Zonas obtenidas para proteger. El primer dígito del nombre de cada figura indica el nivel del presupuesto considerado. El segundo dígito indica el nivel considerado para las cantidades $\{z_1, z_2, z_3\}$. El tercer dígito indica el nivel considerado para las cantidades $\{w_1, w_2, w_3\}$. El último dígito indica el nivel considerado para parámetro β. Los niveles se presentan en la Tabla 4. (a) e2431 (b) e2432 (c) e2441 (d) e2442 (e) e2311 (f) e2312 (g) e2321 (h) e2322 (i) e2331 (j) e2332 (k) e2341 (l) e2342 (m) e2411 (n) e2412 (o) e2421 (p) e2422.

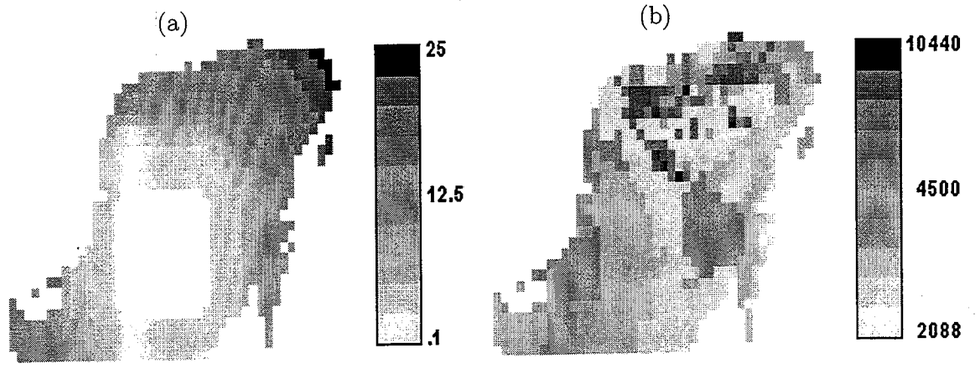

Figura 2-6: (a) Mapa de costos utilizado en el ejercicio de simulación, obtenido mediante la expresión (2.12). (b) Mapa de costos utilizado en los casos de estudio, generado de acuerdo con las ideas de la Sección 3.2.1.

Capítulo 3

Elicitación y Postulación de Valores

En el Capítulo 1 se abordó el problema de inferir la probabilidad de presencia de una especie en nodos determinados por una retícula con base en los valores de M covariables. La inferencia se realizó considerando las $C_{M,2}$ parejas de covariables. Para cada pareja los datos con que se cuenta se resumen por medio de un vector de conteos, $\mathbf{c}_J = (c_J(g))_{g \in F_J}$, donde $c_J(g)$ representa el número de sitios de presencia para los que el vector de covariables correspondiente a la pareja J es g. Se consideró un modelo mezcla (ver expresión (1.7)), en el que cada componente corresponde a una distribución multinomial para cada pareja de covariables. El parámetro de interés para $J \in G$, es $\boldsymbol{\theta}_J = (\theta_J(g))_{g \in F_J}$, donde $\theta_J(g)$ representa la probabilidad de que un ejemplar de la especie bajo estudio se establezca en un sitio en el que los valores de la pareja de covariables J sean g. Para hacer la inferencia se consideró una distribución *a priori* Dirichlet para cada pareja, dada por la expresión (1.9), la cual depende del vector de parámetros $\boldsymbol{\alpha}_J = (\alpha_J(g))_{g \in F_J}$.

El segundo problema que se aborda en esta tesis (Capítulo 2), consiste en seleccionar un conjunto de nodos para proteger especies de interés. La función de pérdida que se propone (Sección 2.5) depende de las cantidades $c(s)$ (costo asignado al nodo s), w_i (importancia de la especie i para ser protegida) y z_i (valor biológico de la especie i), con $i = 1, ..., I$, y $s \in R$.

Este capítulo tiene tres objetivos: (1) elicitar los parámetros de la distribución Dirichlet propuesta en el Capítulo 1 para cada pareja de covariables, (2) postular valores para las probabilidades *a priori* $\pi(J)$ correspondientes a cada pareja de covariables presentes en el modelo (1.7) y (3) postular valores para las cantidades de las que depende la función de pérdida que se usará para proponer regiones para proteger.

Para abordar (1), (2) y (3) se supone que para cada nodo de la región de interés se cuenta con los valores de M covariables, medidas en escala discreta, y en general, se supone que se cuenta con cierta información que un experto versado en la especie puede aportar con respecto a las zonas de presencia y/o ausencia de la especie.

La palabra *elicitar* (o *elicitación*), aunque ha sido utilizada para traducir del inglés la expresión *to elicit*, no es una palabra reconocida de manera oficial en el idioma español.

Este anglicismo se utilizará en lo sucesivo para denotar que ciertos valores serán obtenidos a través de información dada por un experto.

La información con que se cuenta, que será aportada por un experto, consiste en las regiones de presencia y/o ausencia de la especie bajo estudio. En este capítulo se propone una manera en la que el usuario puede aportar la información que posee acerca de las zonas de establecimiento de la especie, y se propone una manera en la que dicha información puede utilizarse para elicitar los parámetros de la distribución Dirichlet.

Para obtener la información *a priori* se propone pedir al experto que, sobre la región de interés, delimite la subregión (o subregiones) en la que asegura que la especie bajo estudio es capaz de establecerse con alta probabilidad. Denótese esta zona por R_1. De la misma manera, se pide al experto que delimite la región (o regiones) en la que asegura que la especie, con alta probabilidad, no es capaz de establecerse, la cual se denota por R_2. Las regiones R_1 y R_2 se denominan *mapas a priori* y su determinación produce, por complemento, la región $R_3 = R \setminus (R_1 \cup R_2)$ en la que el experto no posee conocimiento suficiente para catalogar a dicha región como de presencia o ausencia de la especie.

Como antes, la región de interés se supone cubierta por una retícula regular, la cual define un conjunto de nodos que se denota por R. Dada la pareja J, para cada nodo se considera que se cuenta con los valores de las covariables indicadas por J, es decir, se cuenta con el vector $\mathbf{e}_J(s) = (e_a(s), e_b(s))$ para cada $s \in R$.

Para elicitar el parámetro α_J se utiliza el hecho de que este parámetro se interpreta como la cantidad de información contenida en la distribución *a priori*. Bajo esta interpretación la idea que se explota consiste en medir la cantidad de información que el experto aportó en forma de los mapas *a priori*. Para esto se asume que cada nodo de la retícula puede aportar una unidad de información acerca del establecimiento de la especie. Para medir la cantidad de información se introduce la noción de 3 y 2-contradicción, que se basa en la idea de que un vector de covariables se observe en tres o en dos regiones proporcionadas por el experto. Ya que cada región posee una interpretación propia con respecto al conocimiento del experto acerca de la presencia o ausencia de la especie bajo estudio, este hecho se considera una contradicción. Con base en este concepto, se propone obtener el número de nodos no contradictorios contenidos en los mapas proporcionados por el experto, y el parámetro α_J se define como función del número de nodos no contradictorios.

Para elicitar los parámetros $\alpha_J(g)$, $g \in F_J$, se utiliza la igualdad $\alpha_J(g) = \alpha_J E[\theta(g)]$, la cual es conocida para los parámetros de la distribución Dirichlet, donde $\alpha_J = \sum_{g \in F_J} \alpha_J(g)$ y cada $E[\theta(g)]$ se interpreta como la probabilidad *a priori* de presencia de la especie en un nodo en el que el vector de covariables sea $g \in F_J$. Así, se procede a elicitar la cantidad $E[\theta(g)]$ para toda $g \in F_J$, midiendo la coherencia que el experto posee en las regiones que delimite para este vector de covariables. Para esto, a cada región R_1, R_2 y R_3 se asigna una probabilidad *a priori* de presencia de la especie, denotadas por p_1, p_2 y p_3. Para cada $g \in F_J$ se cuenta el número de nodos observados en cada una de las regiones que poseen el vector

de covariables g. Los conteos obtenidos se ponderan por medio de las probabilidades p_1, p_2 y p_3 y se suman, lo que proporciona un peso $w(g)$ para g. Las cantidades $w(g)$ obtenidas se normalizan, con lo que se elicita $E[\theta(g)]$.

Ya que la cantidad α_J se interpreta como la cantidad de información contenida en la distribución *a priori* para la pareja J, para obtener un valor para las cantidades $\pi(J)$, $J \in G$, se procede a normalizar las cantidades α_J. El esquema de elicitación propuesto puede validarse utilizando la idea de modelo mezcla, calculando la probabilidad de presencia *a priori*, utilizando las cantidades $E[\theta(g)]$ y $\pi(J)$ obtenidas a partir del conocimiento del experto (Sección 3.1.2). Las probabilidades de presencia *a priori* se despliegan en un mapa y se comparan con las regiones *a priori* dadas por el experto. Si el experto aportó información relevante acerca de las zonas de establecimiento y/o no establecimiento de la especie, se observará que las regiones de mayor probabilidad *a priori* de presencia coincidirán con la región de presencia *a priori*, y las regiones de menor probabilidad *a priori* de presencia coincidirán con la región de ausencia *a priori*.

Para postular el costo $c(s)$ de cada nodo de la región de interés (Sección 3.2.1), la cual en esta tesis es la península de Yucatán, se utiliza información contenida en el mapa de Regionalización Económica (García y Alonso, 1999) y el Mapa de Regiones de Especialización Productiva (Aké, Jiménez y Ruenes, 1999). El Mapa de Regionalización Económica proporciona el valor económico por km^2 de la producción agropecuaria y forestal sobre el estado de Yucatán. Por su parte, el Mapa de Regiones de Especialización Productiva proporciona información de las principales actividades productivas que se realizan en la Península. La información de estos mapas se conjunta para producir un valor estimado de la cantidad $c(s)$.

Por su parte, la cantidad w_i se interpreta como la importancia que se asigna a una especie para ser protegida. Una forma de determinar la importancia de una especie es considerar el denominado *Índice de Rareza* (Sección 3.2.2). Existen diversas maneras de definir este índice, una de las cuales es sencillamente el tamaño del área de distribución de la especie (Arita *et al.*, 1997). Es un hecho que el área de distribución de una especie no se conoce con certidumbre. Sin embargo, es posible obtener información acerca del área de distribución de cada especie, por medio de un experto versado en la especie. Siguiendo la notación del Capítulo 1, la región R_1^i puede considerarse como una aproximación del área de distribución de la i-ésima especie. La cantidad w_i se postula como el inverso del tamaño del área determinada por R_1^i, es decir, el inverso del número de nodos contenidos en R_1^i. Así, una especie con distribución restringida será considerada más importante para proteger que una especie con amplia distribución.

Con respecto a la cantidad z_i, ésta se interpreta como el valor biológico que posee la i-ésima especie. En la literatura se propone un mecanismo que permite obtener un estimador de este valor, por medio de aplicar una encuesta a un gran número de personas, lo que en el contexto de esta tesis no es posible. En este capítulo se propone una manera alterna para obtener valores para estas cantidades, utilizando nuevamente las regiones en las que el

usuario afirma que la especie bajo estudio es capaz de establecerse (Sección 3.2.3). Dada la región R_1^i para la especie i y asumiendo que se cuenta con las cantidades $c(s)$ para todo $s \in R$, se propone definir la cantidad z_i como la suma de los costos de los nodos contenidos en R_1^i.

3.1 Elicitación

En la literatura existen diversos trabajos que proponen mecanismos para determinar distribuciones *a priori* con base en el conocimiento de un usuario. Por ejemplo, Winkler (1967a) propone, por medio de un cuestionario, obtener algunos fractiles utilizando el conocimiento del usuario. Los fractiles se utilizan para determinar una función de densidad de probabilidad, dibujando una curva aproximada a través de esos puntos. Winkler (1967b, 1969) y Savage (1971) proponen considerar reglas de tasación (*scoring rules*), cuya función es motivar al usuario a ser honesto durante el proceso de elicitación. Bajo este esquema, se motiva al experto a realizar su mejor esfuerzo en el proceso de elicitación, mediante la asignación de algún pago de acuerdo con la regla de tasación estipulada.

Un hecho generalmente aceptado es que no existe una distribución *a priori* que pueda considerarse como la mejor. De hecho, ya que cada experto posee diferente información *a priori* acerca del fenómeno estudiado, es natural que las correspondientes distribuciones *a priori* sean también diferentes. Citando a Winkler (1967a):

> "todas las valuaciones que sean consistentes o coherentes son admisibles en tanto que el usuario sienta que éstas corresponden a sus juicios".

Por lo tanto, la tarea de elicitación consiste en encontrar valores que reflejen el conocimiento que el experto posee. En esta sección se considera una forma particular de obtener información *a priori*, la cual ha sido utilizada en la práctica en el área de Ecología, aunque de manera informal. Nos referimos al caso en el que la información *a priori* con que se cuenta puede obtenerse en forma de subregiones, que en lo sucesivo se denominarán mapas *a priori*, sobre la región de interés. El objetivo de esta sección es proponer una forma de utilizar dichos mapas para elicitar los parámetros de la distribución *a priori* Dirichlet introducida en el problema descrito en el Capítulo 1.

En general, los mapas *a priori* surgen cuando se estudia un fenómeno que puede asumir un número finito de valores o estados sobre la región y un experto determina subregiones donde, de acuerdo con su conocimiento, cada valor del fenómeno puede ocurrir. Aunque en este capítulo se aborda el caso cuando el fenómeno de interés puede asumir uno de dos estados, las ideas que se presentan pueden generalizarse fácilmente al caso en el que el fenómeno de interés puede asumir uno de Υ estados.

3.1.1 Mapas *a priori* de Presencia-Ausencia

Aunque la tarea particular que nos ocupa es utilizar información de un experto con respecto a las zonas de establecimiento potencial de especies de interés, en esta sección se presentan las ideas en un contexto general. Se supone que sobre una región de interés se estudia un fenómeno que asume un valor de dos posibles en cada sitio. Como antes, sea R el conjunto de nodos determinados por la retícula regular que cubre la región de interés. Sea $u(s)$ variable aleatoria que indica el valor del fenómeno en s. Así, $u(s)$ puede asumir el valor 0 ó 1. Por ejemplo, en el Capítulo 1 se utilizó esta variable para denotar la presencia $(u(s) = 1)$ o ausencia $(u(s) = 0)$ de una especie en el nodo s. Como antes, sea $\mathbf{e}(s)$ el vector de covariables medidos/observados en s. Todas las covariables se suponen medidas en escala discreta. Una idea central que se utiliza en esta sección es la postulación de que la información relevante acerca del valor que asume $u(s)$ en el nodo s se encuentra contenida en las covariables, es decir, en $\mathbf{e}(s)$, y no en la posición geográfica del nodo. Este supuesto también es la base del modelo que se propone en el Capítulo 1 (ver expresión 1.1). En esta tesis se conceptualiza que cada nodo puede aportar una unidad de información con respecto al fenómeno de interés.

Formalizando, los mapas *a priori* se obtienen cuando un experto (posiblemente varios, en cuyo caso deberán proporcionar un mapa que represente el consenso de ellos) delimita zonas sobre la región de estudio, caracterizadas por el hecho de que, de acuerdo con su experiencia y/o conocimiento, sólo un valor del fenómeno estudiado se encuentra en cada una de ellas, con alta probabilidad. De esta manera el experto proporcionará dos regiones sobre la región de estudio, denotadas por R_1 y R_2, cada una de las cuales puede ser disconexa. Es natural esperar que las regiones R_1 y R_2 proporcionadas serán tales que $R_1 \bigcap R_2 = \emptyset$, pues en un nodo no puede registrarse el fenómeno y a la vez no registrarse. Aunque el experto delimitará R_1 y R_2 como regiones, para efecto de este trabajo R_i denotará el conjunto de nodos que se encuentran contenidos en dicha zona. Así, R_1 esta conformada por los nodos en los que el experto afirma que solamente se registra el valor 1 del fenómeno. Por su parte, R_2 estará conformada por los nodos donde el experto afirma que solamente se registra el valor 0 del fenómeno. Note que $R_1 \cup R_2 \subseteq R$, y si la contención es estricta (lo que en general ocurrirá) el experto habrá proporcionado, de manera indirecta, una región (posiblemente disconexa) en la que su conocimiento no es suficiente para determinar cuál valor del fenómeno se presenta en ella. Sea $R_3 = R \setminus (R_1 \cup R_2)$, la cual se encuentra conformada por los nodos en los que el experto no posee información suficiente para asegurar la ocurrencia de uno de los dos valores del fenómeno.

Es claro que en aplicaciones reales el usuario no conoce el verdadero valor del fenómeno bajo estudio en cada s, por lo que los mapas que proporcione contendrán ciertas imprecisiones, las cuales se denominan *contradicciones*. El concepto de contradicción se motiva a partir de la siguiente consideración. En el caso hipotético en el que un experto conociera a la perfección los vectores de covariables que deben estar presentes en cada R_i con respecto

al fenómeno de interés (situación denominada *experto ideal*), cada vector $f \in F$ registrado en algún nodo de R_i no será registrado en nodos de R_j, $j \neq i$. Si lo anterior no se satisface, el experto habrá proporcionado *información contradictoria* en el campo de las covariables, concepto que se formaliza en la Definición 4. La información dada por el experto se medirá en términos de los nodos *no contradictorios* contenidos en los mapas *a priori*.

Definición 4 *Sean $s_1 \in R_1$, $s_2 \in R_2$ y $s_3 \in R_3$. Los nodos s_1, s_2, s_3 definen una 3-contradicción para $f \in F$ si los correspondientes $\mathbf{e}(s_1)$, $\mathbf{e}(s_2)$ y $\mathbf{e}(s_3)$ satisfacen $\mathbf{e}(s_1) = \mathbf{e}(s_2) = \mathbf{e}(s_3) = f$.*

Note que el conjunto R_3 se incluye en la Definición 4 aunque este conjunto es conformado por aquellos nodos donde el usuario no posee suficiente información para clasificarlos en alguna de las regiones R_1 o R_2. La inclusión de R_3 se sustenta en la consideración de que esta región es conformada por nodos donde el usuario *está seguro* de no saber dónde clasificarlos, y esto no constituye una contradicción por sí misma. Así, las tres regiones pueden considerarse como portadoras de información útil acerca del conocimiento del experto con respecto al fenómeno de interés. Las ideas que aquí se presentan pueden utilizarse si por alguna razón se decide no considerar R_3, definiendo $R' = R_1 \cup R_2$ como el conjunto de nodos de interés. En este caso no será posible encontrar 3-contradicciones, pero sí otro tipo de contradicción, como se establece en la Definición 5. Esto mismo ocurrirá si el experto sólo delimita una región *a priori*.

Asumiendo que se consideran las tres regiones R_1, R_2 y R_3, y ya que el interés radica en utilizar la información no contradictoria, se propone encontrar y eliminar las 3-contradicciones de estos conjuntos. De manera intuitiva la idea para hacerlo es la siguiente: se fija un nodo en alguna de las regiones proporcionadas por el usuario. Sea $s_1 \in R_1$ el nodo seleccionado. Dado el nodo s_1, la idea es examinar las regiones R_2 y R_3 y determinar si existen nodos $s_2 \in R_2$ y $s_3 \in R_3$ que posean el mismo vector de covariables que s_1. En caso de que existan dichos nodos, se procede a excluir los tres nodos del estudio. Para cada f, este procedimiento se repite, cada vez con los nodos que permanecen después de eliminar los nodos que se encuentran involucrados en una 3-contradicción, hasta que no se encuentren 3-contradicciones en la región.

El concepto de 3-contradicción no involucra un orden específico en el que los nodos de cada R_i serán examinados para localizar este tipo de contradicción, y de acuerdo con la Definición 4, un nodo puede estar involucrado en varias 3-contradicciones. Por ejemplo, si para algún $s_1 \in R_1$ existen nodos $s_2 \in R_2$ y s_3, $s_3' \in R_3$ tales que $\mathbf{e}(s_1) = \mathbf{e}(s_2) = \mathbf{e}(s_3)$ y $\mathbf{e}(s_1) = \mathbf{e}(s_2) = \mathbf{e}(s_3')$, el nodo s_1 estará involucrado en por lo menos 2 diferentes 3-contradicciones. Sin embargo, lo relevante en la formulación es el concepto genérico de lo que es una 3-contradicción y no los nodos físicamente involucrado en la contradicción. En el ejemplo descrito se procederá a excluir los nodos s_1 y s_2, junto con alguno de los nodos s_3, s_3'. Para excluir los nodos 3-contradictorios según las ideas presentadas, enseguida se propone

una forma sencilla de proceder que permite saber cuántos nodos deben ser eliminados. De hecho, el mecanismo que se propone permite obtener la cantidad de nodos no contradictorios sin necesidad de excluir físicamente a los nodos contradictorios como parte del mecanismo.

Motivados por el postulado de que la información dada por el experto con respecto al fenómeno de interés se encuentra contenida en las covariables, se define el conjunto

$$R_i(f) = \{s \in R_i : \mathbf{e}(s) = f\},$$

para $i = 1, 2, 3$, $f \in F$, siendo F el conjunto de todas las configuraciones posibles de covariables sobre R.

Para cada $f \in F$, sea $r_i^3(f) = |R_i(f)|$ y sea $d^3(f) = \min\{r_1^3(f),\ r_2^3(f),\ r_3^3(f)\}$. Ya que $r_i^3(f)$ representa el número de nodos contenidos en R_i que poseen el vector de covariables f, la cantidad $d^3(f)$ representa una medida del número de 3-contradicciones para el vector de covariables f, contenidas en R_1, R_2 y R_3. La principal propiedad que posee $d^3(f)$ consiste en que es independiente de los nodos físicamente involucrados en las 3-contradicciones. Si resulta $d^3(f) = 0$ se concluirá que no existen 3-contradicciones para f. El número total de 3-contradicciones contenidos en los mapas R_1, R_2 y R_3 es $d^3 = \sum_{f \in F} d^3(f)$.

De acuerdo con estas ideas, para excluir todas las 3-contradicciones basta remover, de cada conjunto $R_i(f)$, un total de $d^3(f)$ elementos para cada $f \in F$. Así, se excluirá un total de $3d^3(f)$ nodos para cada $f \in F$, lo que producirá que un total de $3d^3$ nodos sean removidos de R.

Sea D_1 ($\subseteq R$) el conjunto de nodos que permanecen después de la búsqueda y eliminación de todas las 3-contradicciones. Este conjunto no es único en cuanto a los nodos que lo conforman. Sin embargo, el hecho relevante radica en que el número de nodos contenidos en este conjunto es el mismo, independientemente de los nodos físicamente seleccionados para eliminar las 3-contradicciones.

Por construcción, el conjunto D_1 no contiene nodos involucrados en 3-contradicciones. Es posible, sin embargo, encontrar otro tipo de contradicciones en este conjunto. Esto se establece en la siguiente definición.

Definición 5 *Sean $s_{i_1} \in R_{i_1} \cap D_1$ y $s_{i_2} \in R_{i_2} \cap D_1$, $1 \leq i_1 < i_2 \leq 3$. Los nodos s_{i_1}, s_{i_2} definen una 2-contradicción para f si los correspondientes $\mathbf{e}(s_{i_1})$ y $\mathbf{e}(s_{i_2})$ satisfacen $\mathbf{e}(s_{i_1}) = \mathbf{e}(s_{i_2}) = f$.*

De esta definición se deduce que en D_1 hay tres posibles tipos de 2-contradicción, una por cada par de conjuntos $R_{i_1} \cap D_1$, $R_{i_2} \cap D_1$, $1 \leq i_1 < i_2 \leq 3$. Sin embargo, una vez que las 3-contradicciones han sido removidas, sólo un tipo de 2-contradicción puede encontrarse en D_1, para cada $f \in F$. Aunque este es un hecho intuitivamente claro, se presenta su demostración en la siguiente Proposición.

Proposición 6 *En el conjunto D_1, y para cada $f \in F$, puede encontrarse a lo más un tipo de 2-contradicción.*

Demostración. Suponga que en D_1 hay dos tipos de 2-contradicciones para una $f \in F$ fija. Sin pérdida de generalidad, suponga que dichas contradicciones se encuentran en $R_1 \cap D_1$, $R_2 \cap D_1$ y $R_2 \cap D_1$, $R_3 \cap D_1$. Por definición de 2-contradicción, existen nodos $s_i \in R_i \cap D_1$, $i = 1, 2$, tales que $\mathbf{e}(s_1) = \mathbf{e}(s_2) = f$ y nodos $z_j \in R_j \cap D_1$, $j = 2, 3$, tales que $\mathbf{e}(z_2) = \mathbf{e}(z_3) = f$. Combinando esos hechos, se tienen nodos $s_1 \in R_1$, $s_2 \in R_2$ y $z_3 \in R_3$ tales que $\mathbf{e}(s_1) = \mathbf{e}(s_2) = \mathbf{e}(z_3) = f$, lo que constituye una 3-contradicción en D_1. Esto contradice el hecho de que D_1, por construcción, es conformado por nodos que no están involucrados en 3-contradicciones. ■

Para eliminar las 2-contradicciones se sigue la misma idea que la utilizada para eliminar las 3-contradicciones. Sea $r_i^2(f) = |R_i(f) \cap D_1|$. Defina el conjunto $L_2 = \{(i_1, i_2) : 1 \leq i_1 < i_2 \leq 3\}$, que especifica todos los posibles tipos de 2-contradicciones. Sea $d_l^2(f) = \min\left\{r_{i_1}^2(f), r_{i_2}^2(f)\right\}$, $l \in L_2$, el cual proporciona el número de nodos 2-contradictorios que se encuentran en los conjuntos $R_{i_1}(f) \cap D_1$ y $R_{i_2}(f) \cap D_1$ para $f \in F$. Si $d_l^2(f) = 0$ se concluye que en los conjuntos $R_{i_1}(f) \cap D_1$ y $R_{i_2}(f) \cap D_1$ no existen nodos 2-contradictorios. Si por el contrario resulta que $d_l^2(f) > 0$, se procede a excluir $d_l^2(f)$ elementos de cada uno de esos conjuntos, con lo que se eliminan las 2-contradicciones para esta f. Bajo este esquema, el número total de 2-contradicciones es $d^2 = \sum_{f \in F} \sum_{l \in L_2} d_l^2(f)$.

Tal como en el caso de las 3-contradicciones, es posible que un nodo se encuentre involucrado en varias 2-contradicciones. Sin embargo, la idea relevante es el concepto genérico de 2-contradicción y no el nodo físicamente involucrado en la misma. Sea D_2 ($\subseteq D_1$) el conjunto de nodos que permanecen después de que todos los nodos contradictorios han sido removidos de R.

El conjunto D_2 es conformado por todos los nodos no contradictorios contenidos en los conjuntos R_1, R_2 y R_3. Con este conjunto a la mano se procede a introducir medidas útiles para determinar la cantidad de información contenida en los mapas *a priori*. La idea inmediata que surge es considerar $|D_2|$, el número de nodos no contradictorios que resultan después del proceso de eliminación de las contradicciones, como la cantidad de información. Sin embargo, la existencia de diferentes tipos de contradicciones motiva a considerar un factor de penalización (o ponderación) para cada uno de ellos. El concepto de penalización se introduce en la siguiente definición, donde también se proponen tres cantidades que representan la cantidad de información dada por el experto a través de los mapas R_1, R_2 y R_3.

Definición 7 *La cantidad de información dada por el experto contenida en los conjuntos*

R_1, R_2 y R_3 está dada por:

$$Q = |R| - 3\gamma_{1,1}d^3 - \sum_{f\in F}\sum_{l\in L_2} 2\gamma_{2,l}d_l^2(f), \tag{3.1}$$

donde $\gamma_{h,l} \in [0,1]$, $h = 1,2$, $l \in L_2$ y $d_l^2(f)$ es el número de 2-contradicciones para $f \in F$ correspondiente a la pareja de índices $l \in L_2$. La Proporción de Información dada por el experto es $P = Q/|R|$. Una medida de la Cantidad de Información Relativa se define por $S = Q/(|R| - Q)$.

La cantidad $\gamma_{h,l}$, $h = 1,2$, $l = (i_1, i_2)$, correspondiente a la 2-contradicción presente en los conjuntos $R_{i_1} \cap D_1$ y $R_{i_2} \cap D_1$, penaliza la gravedad de esta contradicción. La cantidad $\gamma_{h,l}$ puede interpretarse como aquella que determina el porcentaje de información que se considera útil de nodos contradictorios al calcular las cantidades de la Definición 7. Por ejemplo, si se postula $\gamma_{2,l} = .5$, cada par de nodos contradictorios aporta un 50% de la información total que poseen a la cantidad de información dada por el experto. Si se postula $\gamma_{h,l} = 1$ para toda h y l, cada triada de nodos involucrados en una 3-contradicción y cada par de nodos involucrados en una 2-contradicción serán removidos por aportar información totalmente contradictoria, por lo que la información que podrían aportar dichos nodos no se considera. En este caso la cantidad Q será simplemente el número de nodos no contradictorios proporcionados por el usuario, es decir, $Q = |R| - 3d^3 - 2d^2$. Por otro lado, si $\gamma_{h,l} = 0$ para alguna h y l, la contradicción correspondiente no es considerada, y en ese caso la información contenida en los nodos involucrados se considera parte de la información dada por el experto.

Para cada cantidad propuesta en la Definición 7, el mínimo valor posible corresponde al caso de información completamente contradictoria, mientras que el máximo valor posible corresponde al caso del denominado experto ideal. Es claro que $Q \in [0, |R|]$, $P \in [0,1]$ y $S \in [0, \infty]$. El uso de las cantidades Q, P o S depende de la información que se desee obtener del usuario. Por ejemplo, si el interés radica simplemente en cuantificar la información dada por el experto, la cantidad Q puede ser útil. Por otra parte, si el interés radica en determinar el porcentaje de información contenida en los mapas *a priori*, la cantidad P puede ser útil. En la Secciones 3.1.2 y 3.1.3 se presentan ejemplos donde la cantidad S se utiliza en un problema práctico que surge en Ecología.

En el siguiente teorema se demuestra que las cantidades que se proponen para medir la información dada por el experto son únicas, bajo el entendido de que se procede eliminando primero las 3-contradicciones.

Teorema 8 *Para regiones R_1, R_2 y R_3 dadas por el experto, las cantidades Q, P y S son únicas, independientemente de los nodos que se eliminen por estar involucrados en* 3 *ó* 2*-contradicciones.*

Demostración. Sea $r_i^3(f) = |R_i(f)|$ y $d^3(f) = \min\{r_1^3(f), r_2^3(f), r_3^3(f)\}$, para un $f \in F$. La cantidad $r_i^3(f)$ representa el número de nodos contenidos en R_i que poseen el vector

de covariables f. Por lo tanto, la cantidad $d^3(f)$ es el número de 3-contradicciones para f contenidas en R_1, R_2 y R_3. Las cantidades $d^3(f)$ y d^3 no especifican cuáles nodos están involucrados en las 3-contradicciones, sino el número de nodos involucrados en estas. Al excluir un total de $d^3(f)$ elementos de cada $R_i(f)$ se eliminan las 3-contradicciones, y el número de nodos que permanece después de la eliminación es $|R| - 3d^3$, independientemente de cuáles nodos fueron eliminados.
Esta misma idea se sigue con las 2-contradicciones, por lo que resulta que la cantidad Q no depende de los nodos físicamente eliminados. Así, la cantidad Q es única, y por lo tanto, lo son también las cantidades P y S por ser funciones de Q y $|R|$. ■

3.1.2 Elicitación de Parámetros de la Distribución Dirichlet

El problema en el que se desea utilizar el conocimiento del experto es inferir la probabilidad de presencia de una especie en nodos de interés. Así, el fenómeno que se estudia es la presencia-ausencia de una especie. Este problema se aborda en el Capítulo 1, donde se propuso la distribución Dirichlet como *a priori* para el parámetro de interés $\boldsymbol{\theta}_J = (\theta_J(g))_{g\in F_J}$, $J = (a,b) \in G$, donde $F_J = \{1,...,R_a\} \times \{1,...,R_b\}$ y $\theta_J(g) = P(\mathrm{e}_J(s) = g)$ denota la probabilidad de que la especie bajo estudio se establezca en un nodo de la región con vector de covariables g para la pareja de covariables J. Las ideas introducidas en la sección anterior se utilizan para elicitar los parámetros de la distribución Dirichlet para cada pareja de covariables, la cual está dada por la expresión (1.9).

Sea $u(s)$ la variable aleatoria que denota la presencia o ausencia de la especie en s. Como se postula en la sección anterior, el experto divide la región R en las regiones R_1, que según su conocimiento corresponde a $u(s) = 1$, región R_2, que según su conocimiento corresponde a $u(s) = 0$, y el complemento $R_3 = R \setminus (R_1 \cup R_2)$. Sea D_1 el conjunto obtenido después de que los $3d^3$ nodos involucrados en las 3-contradicciones han sido removidos. Se procede ahora a eliminar las 2-contradicciones.

En el contexto de este problema, el peor tipo de 2-contradicción se presenta en los conjuntos $R_1 \cap D_1$ y $R_2 \cap D_1$. Esto se debe a que, por el significado que poseen los nodos contenidos en R_1 y R_2, esta contradicción es equivalente a afirmar "la especie puede establecerse y a la vez no puede establecerse en un nodo con vector de covariables g". Debido a esta consideración es sensato postular que $\gamma_{2,l} = 1$, $l = (1,2)$. Así, la información contenida en los nodos involucrados en este tipo de contradicción no será considerada. Las otras posibles 2-contradicciones son menos graves. Por ejemplo, una 2-contradicción que involucre nodos contenidos en los conjuntos $R_1 \cap D_1$ y $R_3 \cap D_1$ equivale a afirmar "la especie puede establecerse en un nodo con cierto vector de covariables g y a la vez no estoy seguro de ello", lo que es menos contradictorio que la primera contradicción citada. Así, para estas contradicciones puede postularse $\gamma_{2,l} = .5$, con $l = (1,3)$.

Una vez que se han determinado los valores $\gamma_{h,l}$, se procede a utilizar la Definición 7 para

elicitar la cantidad α_J de la distribución Dirichlet. El parámetro α_J de esta distribución puede interpretarse como la cantidad de información contenida en la distribución *a priori* (Gelman *et al.* 1995, p. 76). De esta manera, el valor $\alpha_J = 0$ se interpreta como ausencia de información *a priori*, mientras que $\alpha_J = +\infty$ se interpreta como la máxima cantidad de información *a priori* posible (que corresponde al denominado experto ideal). La interpretación que posee α_J permite, en principio, proponer la cantidad Q de la Definición 7 como un valor sensato para ella. Sin embargo, el rango de α_J es $[0, \infty)$, mientras que el rango de Q es $[0, |R|]$. Por esta razón se propone utilizar la cantidad S en lugar de Q como el valor de α_J, es decir, se define $\alpha_J = S$. Con este parámetro elicitado, se procede a proponer valores para los parámetros $\alpha_J(g)$, para cada $g \in F_J$.

Para la distribución Dirichlet es un hecho conocido que $\alpha_J(g) = \alpha_J E[\theta_J(g)]$, donde $E[\theta_J(g)]$ es el valor esperado *a priori* de $\theta_J(g)$. Así, se procede a elicitar las cantidades $E[\theta_J(g)]$. Los mapas *a priori* delimitados por el experto representan su conocimiento acerca de las zonas de presencia y/o ausencia de la especie. Se postula que la probabilidad *a priori* de presencia para los nodos contenidos en R_1 es p_1, para nodos contenidos en R_2 es p_2 y para nodos contenidos en R_3 es p_3. En particular, para las aplicaciones y simulaciones de esta tesis se postula $p_1 = .99$, $p_2 = .01$ y $p_3 = .5$. En general, los valores p_i pueden ser determinados por el experto de acuerdo con el significado que posea cada región R_i y considerando la seguridad que posee acerca de las regiones que proporcione. Usando esas ideas se define

$$w(g) = \sum_{i=1}^{3} p_i \frac{|R_i(g)|}{|\{s \in R : \mathbf{e}_J(s) = g\}|}, \tag{3.2}$$

que representa una ponderación (o peso) que se asigna a cada $g \in F_J$ de acuerdo con la manera en la que el vector de covariables g se distribuye en las regiones R_1, R_2 y R_3. Intuitivamente, la expresión (3.2) mide la coherencia del experto, en el campo de covariables, para cada $g \in F_J$, cuando delimita las regiones R_i. Normalizando las cantidades (3.2), se propone definir

$$E[\theta_J(g)] = \frac{w(g)}{\sum\limits_{g' \in F_J} w(g')}.$$

Utilizando la igualdad que relaciona los parámetros de la distribución Dirichlet se obtiene $\alpha_J(g) = \alpha_J \left[w(g) / \sum_{g' \in F_J} w(g') \right]$.

Con respecto a las cantidades $\pi(J)$, ya que el parámetro α_J de la distribución *a priori* correspondiente a la pareja J se interpreta como la cantidad de información contenida en la

distribución *a priori*, se propone definir

$$\pi(J) = \frac{\alpha_J}{\sum\limits_{J' \in G} \alpha_{J'}}.$$

Una vez que se han elicitado los parámetros de cada una de las distribuciones Dirichlet y las cantidades $\pi(J)$, es posible inspeccionar si las regiones proporcionadas por el experto aportan información relevante acerca del fenómeno bajo estudio, en nuestro caso, las zonas de presencia y ausencia de la especie. Para esto se propone calcular

$$P\{u(s) = 1\} = \sum_{J \in G} \pi(J) E\left[\theta_J\{e_J(s)\}\right].$$

La cantidad $P\{u(s) = 1\}$, con $s \in R$, se interpreta como la probabilidad *a priori* de presencia de la especie en el nodo s, y como se observa, se calcula utilizando solamente las cantidades elicitadas, es decir, el conocimiento del experto. Si se calcula esta expresión para todos los nodos de la retícula con que se cuenta, es posible desplegar un mapa, que se denomina *mapa de probabilidades a priori de presencia.* Para desplegar este mapa, se considera una partición del intervalo $[0, 1]$ y una escala de color o de gris, siguiendo las ideas mediante las cuales se despliegan los mapas de probabilidades de presencia y certidumbre obtenidos en el Capítulo 1. Si el experto aportó información *a priori* relevante se espera observar que los contornos del mapa de probabilidades de presencia *a priori* coincidan con los mapas *a priori* proporcionados por el experto. En caso que esto no ocurra, se tendrá evidencia de que el experto no conoce realmente las regiones de establecimiento y/o de no establecimiento de la especie, o bien, que el conocimiento del experto acerca del establecimiento de la especie se basa en covariables que no se están considerando al momento de realizar la aplicación.

3.1.3 Ejemplo: Mapas *a priori* sobre una Región Ficticia.

Con el fin de observar el funcionamieno del proceso de elicitación propuesto, en esta sección se considera una región ficticia, sobre la cual se dibujan mapas *a priori* según las ideas de la sección anterior. Sea R la retícula regular conformada por 30 filas y 30 columnas, lo que genera un total de 900 nodos ($|R| = 900$). Se considera una sola covariable que consta de 5 niveles, cuya distribución sobre la región puede observarse en la Figura 3-1(b). Así, se tiene que $F = \{1, 2, 3, 4, 5\}$. Se presentan dos escenarios en los que se obtienen las cantidades Q, P y S bajo diferentes situaciones con respecto a los mapas proporcionados por el experto y se elicitan los parámetros de la distribución Dirichlet. En el primer escenario se considera el caso en el que los mapas *a priori* son grandes con respecto a la región bajo estudio, los cuales se presentan en la Figura 3-1(c). El segundo escenario presenta el caso en el que los mapas *a priori* son pequeños con respeto a la región bajo estudio, los cuales se presentan en

la Figura 3-1(d).

Aplicando las ideas de la sección anterior, si se postula $\gamma_{h,l} = 1$ para toda h y toda l, en el primer escenario se obtienen las cantidades $Q = 130$, $P = .14$ y $S = .169$, en tanto que en el segundo se obtienen las cantidades $Q = 759$, $P = .843$ y $S = 5.383$. Note que la cantidad de información es mayor para el escenario 2, en el que los mapas *a priori* poseen menor área que los mapas *a priori* del escenario 1.

Si ahora se asume que el peor tipo de 2-contradicción se encuentra en los conjuntos dados por $l_1 \in L_2$, por lo que se portula $\gamma_{2,l_1} = 1$, $\gamma_{1,1} = \gamma_{2,l_2} = \gamma_{2,l_3} = .5$, denotando que las 3-contradicciones y las otras posibles 2-contradicciones se penalizan considerando el 50% de la información que contienen, se obtiene $Q = 439$, $P = .49$ y $S = .95$ para el primer escenario y $Q = 831$, $P = .92$ y $S = 12.04$ para el segundo.

Se observa que al calcular las cantidades utilizando las diferentes ponderaciones para las contradicciones, se obtiene mayor magnitud en las cantidades Q, P y S que si se postula que todas las contradicciones son igualmente penalizadas. Por ejemplo, para el primer escenario se obtuvo la cantidad $Q = 130$ con $\gamma_{h,l} = 1$ para toda h y l, y $Q = 493$ con las ponderaciones postuladas. También se observa que las cantidades Q, P y S obtenidas para el escenario 2 son mayores que las cantidades obtenidas para el escenario 1.

Este sencillo ejercicio ilustra el efecto de las cantidades $\gamma_{h,l}$ sobre la cantidad de información aportada por el experto. También ilustra que la cantidad de información no depende del tamaño de los mapas *a priori* proporcionados por el experto, sino de la coherencia con que los valores de las covariables se encuentren dispersos en dichos mapas.

Siguiendo las ideas introducidas en la sección anterior, se procedió a realizar el ejercicio de elicitación de los parámetros de la Dirichlet, que en este caso, es el vector $\boldsymbol{\alpha} = (\alpha(1), \alpha(2), \alpha(3), \alpha(4), \alpha(5))$, que corresponden a cada uno de los valores de la covariable considerada. Se procede a elicitar $E(\boldsymbol{\theta})$ por medio de la igualdad conocida que satisfacen los parámetros de la distribución Dirichlet.

Para el caso de considerar $\gamma_{h,l} = 1$ para toda h y toda l se obtiene que los vectores de esperanzas *a priori* son $E(\boldsymbol{\theta}) = (.22, .14, .21, .22, .22)$ y $E(\boldsymbol{\theta}) = (.42, .24, .24, .08, .02)$, respectivamente. Estos valores producen los correspondientes vectores de parámetros $\boldsymbol{\alpha} = (.039, .024, .031, .038, .038)$ y $\boldsymbol{\alpha} = (2.31, 1.32, 1.30, .46, .13)$ para los escenarios 1 y 2, respectivamente. Para el caso de considerar los diferentes tipos de 2-contradicción, sea $\gamma_{1,1} = \gamma_{2,l} = 1$ y sea .5 el peso asignado a las otras 2-contradicciones. Para el primer escenario el vector de valores esperados *a priori* es $E(\boldsymbol{\theta}) = (.229, .140, .182, .227, .222)$, en tanto que para el segundo escenario es $E(\boldsymbol{\theta}) = (.42, .24, .24, .08, .024)$. Estos vectores producen que los parámetros de las correspondientes distribuciones Dirichlet sean $\boldsymbol{\alpha} = (.261, .160, .208, .259, .254)$ para el primer escenario y $\boldsymbol{\alpha} = (3.440, 1.970, .942, .683, .197)$ para el segundo.

En el primer escenario, los parámetros elicitados no son muy diferentes entre sí. Esto ocurre debido a que todos los valores de las covariables se encuentran presentes en las tres regiones *a priori*, aproximadamente en la misma proporción, por lo que se encuentran muchas

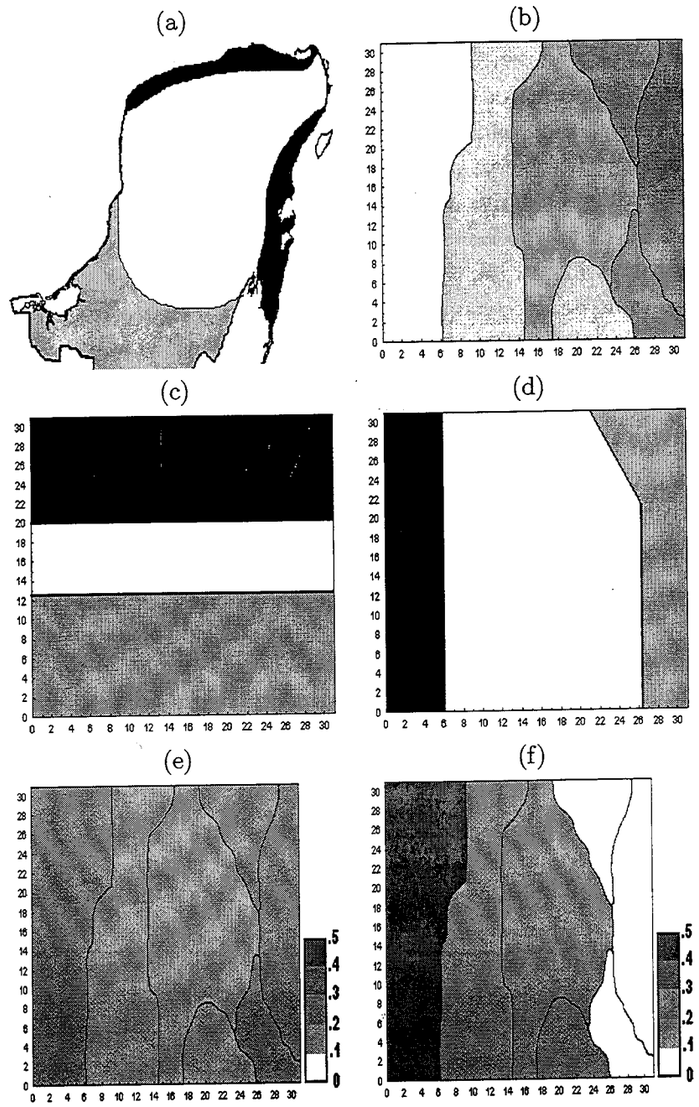

Figura 3-1: (a) Mapas *a priori* para *C. readii* proporcionados por el experto. El color negro representa la zona de presencia *a priori* y el color gris la zona de ausencia *a priori*. El color blanco denota falta de información. (b) Mapa ficticio de estudio. Los niveles de la covariable van de 1 (blanco) a 5 (obscuro) (c) Mapas *a priori* para el escenario 1. (d) Mapas *a priori* para el escenario 2. (e) Mapa de probabilidades de presencia *a priori* para el escenario 1. (f) Mapa de probabilidades de presencia *a priori* para el escenario 2.

contradicciones de acuerdo con las Definiciones 4 y 5 (ver Figura 3-1(c)).

En el escenario 1 la probabilidad *a priori* de presencia para cada f, es aproximadamente igual a .2. Esto coincide con lo observado en la Figura 3-1(c), en la que todos los valores de la covariable se encuentran en cada una de las regiones R_i. En el segundo escenario, el nivel 1 de la covariable se encuentra casi totalmente contenido en R_1, mientras que el nivel 5 lo está en R_3 (ver Figura 3-1(d)). Esto se ve reflejado en las probabilidades *a priori* de presencia, que resultaron ser .42 y .02, respectivamente, las que a su vez producen los valores $\alpha(1) = 3.44$ y $\alpha(5) = .197$ para los correspondientes parámetros de la Dirichlet.

Para validar el proceso de elicitación se procedió a observar el mapa de probabilidades *a priori* de presencia para cada uno de los escenarios. Se despliegan solamente los mapas obtenidos con $\gamma_{h,l} = 1$ para toda h y toda l. El mapa correspondiente al primer escenario se presenta en la Figura 3-1(e), donde se observa que las probabilidades *a priori* de presencia obtenidas no coinciden con las regiones proporcionadas por el experto (Figura 3-1(b)). En este caso es posible afirmar que los mapas *a priori* no aportan mucha información con respecto a las zonas de establecimiento y no establecimiento de la especie. El mapa de presencia *a priori* para el segundo escenario se presenta en la Figura 3-1(f). En este mapa se observa que las probabilidades *a priori* de presencia que se obtienen coinciden con los mapas *a priori* proporcionados por el experto. Por lo tanto, en este caso es posible afirmar que el experto aportó información relevante con respecto a las zonas de establecimiento y no establecimiento de la especie.

En vista de los resultados que se obtuvieron al comparar los mapas de probabilidades *a priori* de presencia y los mapas *a priori* proporcionados por el experto, la forma que se propone para validar el proceso de elicitación es una herramienta potencialmente útil que se sugiere utilizar antes de proceder a usar los valores elicitados en el proceso de inferencia.

3.2 Postulación de Valores

En esta sección se presenta una forma de determinar valores para las cantidades $c(s)$, $s \in R$, y para las cantidades w_i y z_i, para cada una de las especies bajo estudio. Estas cantidades se encuentran involucradas en las funciones de pérdida introducidas en el Capítulo 2. Recordando, $c(s)$ denota el costo asignado al nodo s y representa la cantidad que deberá ser invertida si el nodo s es seleccionado para ser protegido. La cantidad w_i representa el nivel de importancia que se asigna a la especie para ser protegida. Por su parte, el valor z_i se interpreta como valor de la especie en la región de interés.

3.2.1 Costo de cada Nodo: $c(s)$

La cantidad $c(s)$ representa el costo asociado al nodo s. En la práctica no es común que se cuente con un mapa donde se registre el costo de cada nodo sobre una región. Sin embargo,

existen mapas que pueden usarse como auxiliares para postular el costo de cada nodo, como se ilustra en esta sección, para el caso particular de la península de Yucatán, que se considera la región de estudio en las aplicaciones del Capítulo 4.

Ya que el objetivo es determinar la zona (o zonas) que será propuesta para proteger, y dado que proteger una zona involucra un costo, para postular un valor para cada uno de los nodos se utilizó el mapa conocido como Mapa de Regionalización Económica (MRE), introducido por García y Alonzo (1999). Este mapa proporciona zonas sobre el estado de Yucatán determinadas por el valor económico por km^2 de la producción agropecuaria y forestal. Sin embargo, no se cuenta con un mapa análogo para los estados de Campeche y Quintana Roo. Utilizando un segundo mapa y la información contenida en el MRE, es posible obtener un estimado del valor de la producción en esos dos estados.

El mapa que se utilizó como auxiliar para determinar el valor de los nodos contenidos en los estados de Campeche y Quintana Roo es el conocido como Mapa de Regiones de Especialización Productiva (MREP; Aké, Jiménez y Ruenes, 1999), que aporta información de las principales actividades productivas que se realizan en la Península. Las principales zonas de actividad que se describen en el MREP son: ganadera, maicera, henequenera, citrícola, forestal, forestal-maicera, forestal-arrocera y zonas que reportan actividades ganaderas-citrícolas-maiceras-forestales.

Ya que cada zona determinada en el MRE posee intersección con cada una de las zonas determinadas en el MREP, se propone lo siguiente. Para cada región definida del MREP se promedia el costo de los nodos contenidos en ella que se encuentran sobre el estado de Yucatán. Para un nodo localizado sobre Campeche o Quintana Roo se observará la categoría a la que pertenece de acuerdo con el MREP y se le asignará el costo promedio obtenido para esta categoría de acuerdo con el MRE. El mapa de costos que se obtuvo utilizando esta forma de proceder se presenta en la Figura 2-6(b).

3.2.2 Importancia Asignada a Proteger una Especie: w_i

Cuando se consideran I especies y se desea proponer una zona para protegerlas es razonable suponer que algunas de ellas pueden considerarse más importantes de proteger que otras. Un elemento que permite evaluar la importancia que posee una especie para ser protegida es el denominado *Índice de Rareza* de la especie. En teoría, el índice de rareza depende del tamaño del área de distribución de la especie, del tamaño de la población de la especie y de la llamada especificidad de hábitat. Los organismos con cierta especificidad de hábitat son aquellos que están adaptados a la utilización y tolerancia de rangos pequeños de gradientes ambientales (están restringidos a ciertos vectores de covariables). Ya que en general no se cuenta con información detallada acerca del tamaño de la población, ni cuál es el vector de covariables preferido por una especie, en la práctica el índice de rareza se calcula con base en el área de distribución de la especie (Arita *et al.* 1997).

Aunque el área de distribución de una especie no se conoce con certidumbre, es posible obtener de un experto información acerca de las zonas en las que se espera que la especie sea capaz de establecerse con alta probabilidad, como se describe en la Sección 3.1.1. La región que aporte el experto puede considerarse como una aproximación al área de distribución de la especie. Siguiendo la notación de la Sección 3.1.1, sea R_1^i la región de presencia potencial proporcionada por el experto para la especie i. En la Figura 4-7 se observan ejemplos de estos mapas, los cuales fueron proporcionados por un experto del Centro de Investigación Científica de Yucatán, y corresponden a las especies listadas en la Tabla 8, que se utilizan para aplicar las ideas de esta tesis (Capítulo 4).

De acuerdo con Arita *et al.* (1997), la cantidad $a_i = |R_1^i|$ proporciona el índice de rareza de la i-ésima especie. Suponga que un experto (o varios, quizá uno por especie) ha determinado la región R_1^i para cada especie y se cuenta por lo tanto con las cantidades a_i, $1 \leq i \leq I$. Con base en estas cantidades se propone calcular

$$w_i^{'} = \frac{1}{a_i}. \tag{3.3}$$

Las especies para las que se haya determinado amplias áreas de distribución potencial tendrán valores menores de $w_i^{'}$. Finalmente, se propone, mediante la normalización de las $w_i^{'}$'s, definir

$$w_i = \frac{w_i^{'}}{\sum_{j=1}^{I} w_j^{'}}. \tag{3.4}$$

En la práctica es posible que no se cuente con información de la región de presencia de alguna especie. Para asignar un valor a_i a dichas especies se propone lo siguiente. Una vez que se han obtenido los valores $a_1, ..., a_{I'}$, donde I' es el número de especies para las que se cuenta con el mapa de presencia *a priori* ($I' < I$), se procede a promediar todos los valores que sean mayores o iguales que el percentil del q% obtenido de las cantidades $a_1, ..., a_{I'}$. El promedio que resulte será asignado a todas las especies para las que el experto no fue capaz de proporcionar un mapa *a priori* de presencia. El valor q se asignará con base en consideraciones prácticas. Si por ejemplo se sabe que la especie se considera *rara*, podría postularse $q = 90$. Otra posibilidad de acción es asignar a q el valor de algún percentil, el cual se determinará con base en las cantidades a_i con que se cuenta. Una vez que se cuenta con los valores a_i, $1 \leq i \leq I$, se procede a calcular las cantidades w_i a través de las expresiones (3.3) y (3.4).

Una forma alterna de determinar valores w_i, $i = 1, ..., I$, se presenta si el usuario es capaz de ordenar a las especies de acuerdo con algún nivel de importancia. Para ordenar a las especies podría utilizarse la clasificación que la IUCN y el WWF (Fondo mundial para la naturaleza) asignan a las especies de acuerdo con el riesgo de extinción que poseen. Las categorías definidas en dicha clasificación son: en peligro crítico, en peligro, vulnerable

(genéricamente conocidas como amenazadas de extinción), dependientes de la conservación, casi amenazadas y de preocupación menor. A cada categoría i se asignará (de manera arbitraria) un peso w^*, de acuerdo con el orden definido. El valor w_i' que se asigne a cada especie será el valor w^* de la categoría a la que corresponda. Bajo este esquema puede ocurrir que se asigne el mismo valor w_i' a dos o más especies. El nivel de clasificación correspondiente a las especies que se encuentren en mayor peligro de extinción poseerá el mayor valor de w_i', mientras que el nivel correspondiente a especies en menor peligro poseerá el menor valor w_i'. Bajo este esquema se propone postular w_i mediante la expresión (3.4). En esta tesis, para obtener las cantidades w_i se utilizan los mapas proporcionados por experto y las cantidades (3.3) y (3.4).

3.2.3 Valor Biológico de una Especie: z_i

En la práctica no es tarea fácil asignar un valor que represente el costo que posee una especie de interés. La pregunta que se plantea en este contexto es ¿Cuánto se pierde si una especie se extingue, o bien, desaparece de una región de estudio? Para contestar esta pregunta se ha propuesto un método denominado Método de Valuación Contingente (CVM, por sus siglas en inglés), en el que se evalúa la disposición de una población a pagar cierta cantidad de dinero (*Willingness to Pay*) con tal de implementar acciones para conservar un bien, en nuestro caso, una especie. Para esto se propone utilizar un cuestionario diseñado para obtener un estimador de la cantidad que la población está dispuesta a pagar para no perder a una especie de interés. Por medio de este cuestionario se obtiene una cantidad que se interpreta como el costo del bien en cuestion.

Para efectos de esta tesis no fue posible realizar la estimación de los valores de las especies por medio del método CVM, por lo que se recurrió a una forma alterna para determinar el costo estimado de perder una especie en una región de interes. La forma que se propone se basa en el mapa de costos, obtenido según las ideas de la Sección 3.2.1, y en los mapas de presencia *a priori* proporcionados por el experto, según lo descrito en la Sección 3.1.1.

Para calcular la cantidad z_i para cada especie bajo estudio se propone sumar los costos de los nodos contenidos en las zonas de presencia potencial *a priori* R_1^i delimitadas por el experto. Sea Z_i la suma obtenida para la especie i, es decir, $Z_i = \sum_{s \in R_1^i} c(s)$. Seguidamente se procede a promediar dicha cantidad sobre el número de nodos de R y se define

$$z_i = \frac{Z_i}{|R|}. \tag{3.5}$$

La idea intuitiva que justifica la definición de z_i mediante la expresión (3.5) es la siguiente. La cantidad z_i se interpreta como el costo en que se incurre si se pierde la especie en el nodo s, y este costo se asume constante sobre toda la región, es decir, $z(s) = z$ para todo $s \in R$, suprimiendo la dependencia sobre s. Por su parte, la cantidad Z_i puede interpretarse como el

costo que debe pagarse para que la i-ésima especie sea preservada sobre la región de estudio, mediante la protección de la región de habitación potencial de la especie. Al promediar con respecto al número de nodos contenidos en R se está encontrando el costo promedio de preservar la especie sobre la región de estudio, lo que puede interpretarse como una estimación de la cantidad z_i.

Con esta forma de proceder se determina la cantidad que cierta población hipotética deberá pagar con tal de que la especie bajo consideración no desaparezca de la región bajo estudio. Esta forma de proceder puede conceptualizarse como un caso particular de la metodología CVM, en el que la cantidad a pagar no se determina por medio de un cuestionario, sino mediante otro tipo de información que se tiene a la mano, en este caso, la zona de distribución de la especie y los costos de proteger los nodos que la conforman.

Con respecto a especies para las que no se cuente con información *a priori* acerca de su área de distribución, se propone asignar algún valor representativo, calculado con base en los valores z_i con que se cuente. El valor representativo deberá asignarse bajo la asesoría de un experto, que si bien no conoce el área de distribución de la especie, será capaz de aportar alguna información con respecto a la situación en el que se encuentra la especie. Por ejemplo, si se sabe que la especie es considerada una especie rara, pero no se tiene idea acerca de su área de distribución, deberá asignarse a la especie un valor alto. Se propone asignar como valor a esta especie algún percentil, por ejemplo el percentil del 90% de los valores de z_i con que se cuente. Otra manera de determinar un valor biológico para las especies para las que no se cuente información es promediar los valores con que se cuente que sean mayores que un percentil determinado. El valor del percentil dependerá de la situación en la que se encuentre la especie. Así, para una especie rara se postulará un percentil alto, con lo que se promediará los valores z_i correspondientes a especies con valor biológico alto.

3.2.4 Discusión

Las cantidades propuestas en la Definición 7 no miden la calidad de los mapas *a priori*. De hecho, si se postula $\gamma_{h,l} = 1$, para toda h y toda l, las cantidades Q, P *y* S no cambiarán si los conjuntos R_1, R_2 y R_3 se rotulan en otro orden. Sin embargo, si alguna de las cantidades $\gamma_{h,l}$ es diferente de 1, las cantidades Q, P y S si cambiarán si las regiones R_1, R_2 y R_3 se rotulan de manera distinta. Por esta razón es importante que las regiones *a priori* sean proporcionadas por un experto, como se ha resaltado a lo largo de este capítulo.

Ya que la cantidad de información será medida usando las covariables en lugar de las coordenadas geográficas, el hecho de que en alguna aplicación los mapas *a priori* sean grandes con respecto a R no necesariamente implicará que se cuenta con mucha información (ver ejemplos en Sección 3.1.3). Por esta razón, deberá motivarse al experto a proporcionar los mapas *a priori* delimitando solamente las regiones en las que se encuentra muy seguro de la presencia potencial o de la ausencia de la especie.

El orden en el que los tipos de contradicciones son considerados en el proceso de eliminación es crucial. Deberá procederse primero eliminando las 3-contradicciones y posteriormente las 2-contradicciones. Ese orden de eliminación garantiza que la cantidad de información será la misma independientemente de los nodos físicamente involucrados en las contradicciones.

En este capítulo se propuso una forma de validar el proceso de elicitación que se realiza. En la práctica la validación de un procedimiento de elicitación es pedir al usuario que evalúe (de manera subjetiva) las cantidades que resultan de la elicitación, lo cual en general no es sencillo. En nuestro caso, en lugar de que el usuario evalúe directamente las cantidades elicitadas, observa un mapa obtenido con base en las cantidades elicitadas, lo que permite que la validación se realice de manera más sencilla para el experto.

Las ideas introducidas con respecto a las 3 y 2-contradicciones pueden generalizarse al caso en el que el fenómeno de interés puede asumir uno de Υ valores, con $\Upsilon \geq 2$. En este caso un experto podría aportar los mapas *a priori* $R_1, R_2, ..., R_\Upsilon$, con $\bigcup_{i=1}^{\Upsilon} R_i \subseteq R$, en cuyo caso se tendrá $R_{\Upsilon+1} = R \setminus \left[\bigcup_{i=1}^{\Upsilon} R_i\right]$. En estos mapas será posible encontrar Υ, $(\Upsilon - 1)$,..., y 2-contradicciones, cada una de las cuales se define de manera análoga a las Definiciones 4 y 5. Bajo este esquema, la expresión para Q será

$$Q = N - (\Upsilon + 1)\gamma_{1,1} d^{\Upsilon+1} - \sum_{t=2}^{\Upsilon} \sum_{f \in F} \sum_{l \in L_t} (\Upsilon - t + 2)\gamma_{t,l} d_l^{\Upsilon-t+2}(f), \tag{3.6}$$

donde $l \in L_t = \{(i_1, i_2, ..., i_t) : 1 \leq i_1 < i_2 < ... < i_t \leq \Upsilon\}$, $2 \leq t \leq \Upsilon$. Las cantidades P y S se definen de manera similar a las correspondientes de la Definición 7. Es fácil verificar que cuando $\Upsilon = 2$ la expresión (3.6) se reduce a la expresión (3.1).

La cantidad (3.6) puede utilizarse, por ejemplo, cuando el objetivo es clasificar los tipos de suelo (o tipos de cultivo) en una región específica de interés y el número total de posibles categorías, Υ, es conocido de antemano. En este caso lo mapas *a priori* R_i, $1 \leq i \leq \Upsilon$, corresponden al área caracterizada por poseer el i-ésimo tipo de suelo, de acuerdo con el conocimiento del experto.

Capítulo 4

Casos de Estudio

En este capítulo se presentan 2 aplicaciones de las ideas introducidas en los Capítulos 2 y 3 a especies de la península de Yucatán. En cada caso, los parámetros de las correspondientes distribuciones Dirichlet y las cantidades que definen las funciones de pérdida se postulan utilizando las ideas introducidas en el Capítulo 4.

La primera aplicación se realizó con una especie de palma. El mapa de probabilidades de presencia que se obtiene se compara con los mapas obtenidos con las metodologías alternas FloraMap y Domain. En cada caso también se obtuvo el mapa de certidumbre. El experto proporcionó las regiones de presencia y ausencia potencial para esta especie, de acuerdo con su conocimiento acerca de las áreas de establecimiento y no establecimiento de la especie. Utilizando el mapa de probabilidades que resultó, se procedió a encontrar una región que se propondrá para proteger a la especie, bajo diferentes escenarios.

En la segunda aplicación se consideran 12 especies, las cuales se encuentran presentes también en la península de Yucatán y se consideran especies amenazadas. Para cada especie se obtuvo el mapa de probabilidades de presencia y el mapa de certidumbre. En cada caso, se utilizó información del experto con respecto a las zonas de presencia y/o ausencia de cada especie. Utilizando los mapas de probabilidades de presencia, se procedió a proponer una región para proteger, obtenida considerando las 12 especies a la vez.

4.1 Una Especie: *Coccothrinax readii*

Esta especie pertenece a la familia *palmae* y se considera especie amenazada. Se cuenta con un total de 67 registros de presencia, algunos de ellos múltiples (más de un registro en algunos nodos), los cuales se observan en la Figura 4-1(c).

4.1.1 Mapa de Probabilidades de Presencia y de Certidumbre

El mapa de probabilidades de presencia se obtuvo incluyendo información de un experto. Las regiones *a priori* R_1 y R_2, tal como fueron proporcionadas por el experto del CICY, se presentan en la Figura 4-1(a). De esta figura se infiere que el experto considera que esta especie es capaz de establecerse a lo largo de las costas norte y este de la Península.

Para desplegar los mapas obtenidos se utilizó la misma partición que la utilizada para desplegar los mapas generados en los ejercicios de simulación (Sección 1.5), es decir, se considera una partición del intervalo $[0, 1]$ en 10 subintervalos de longitud .1 y una escala de grises.

Para evaluar si el experto aportó información coherente a través de los mapas *a priori*, se procedió a obtener el mapa de probabilidades *a priori* de presencia, el cual se presenta en la Figura 4-1(b). En esta figura se observa que los contornos de este mapa coinciden con las regiones proporcionadas por el experto. Así, las probabilidades *a priori* mayores se observan en la región señalada como de presencia *a priori*, en tanto que las probabilidades *a priori* menores coinciden con la región de ausencia proporcionada por el experto. Por esta razón se tiene evidencia de que el experto ha proporcionado información coherente acerca de las zonas de establecimiento de la especie.

El mapa de probabilidades de presencia obtenido se presenta en la Figura 4-1(c). El correspondiente mapa de certidumbre, obtenido según las ideas de la Sección 1.3, se presenta en la Figura 4-1(d). Por su parte, los mapas que resultan de utilizar los métodos FloraMap y Domain se presentan en las Figuras 4-1(e) y (f), respectivamente.

El mapa de probabilidades de presencia (Figura 4-1(c)) fue observado por el experto involucrado en el estudio de esta especie, quien comentó que consideraciones recientes sugieren que esta especie se encuentra en proceso de expandir su área de distribución. Las zonas de alta probabilidad de presencia resaltadas por el método propuesto coinciden con la apreciación que se tiene acerca de las zonas que se sospecha dicha especie podría colonizar en el futuro, que son las zonas que se observan hacia el centro de la Península. En este caso, aunque el experto tenía la sospecha de que las áreas de establecimiento potencial podían incluir una zona hacia el centro de la Península, esa zona no la proporcionó como parte de la zona de presencia *a priori*, por no contar con la suficiente seguridad.

Otro comentario relevante que surgió al observar el mapa de probabilidades se refiere al sitio de presencia aislado reportado cerca del centro de la península. La validez de dicho sitio se encuentra bajo discusión. La combinación del resultado observado en el mapa de probabilidades, el cual asigna baja probabilidad de presencia de la especie en ese sitio (Figura 4-1(c)) con el resultado observado en el mapa de certidumbre, que asigna alto nivel de certidumbre para la probabilidad de presencia (Figura 4-1(d)), sugiere que dicho sitio de presencia puede ser anómalo. Es posible que en ese sitio se haya reportado la presencia de la especie, y más aún, es posible que el ejemplar haya sobrevivido en dicho lugar. Sin embargo,

por la evidencia que aportan los mapas de probabilidades de presencia y de certidumbre se sospecha que ese sitio no es óptimo para la presencia de la especie. Note que tanto FloraMap como Domain producen una región de alto potencial alrededor de este sitio.

Con respecto al resultado obtenido con FloraMap (Figura 4-1(e)), el experto aseguró que el mapa de potencial de presencia obtenido incluye zonas que, según su apreciación, no son adecuadas para el establecimiento de la especie. De acuerdo con el experto, el método FloraMap sobrestimó la región de alto potencial para esta especie, aunque en menor grado que Domain. En la Figura 4-1(f) se observa el resultado obtenido con Domain, con el que se concluiría que casi toda la Península es de alto potencial para la presencia de la especie.

En esta aplicación se calcularon las cantidades $\pi(J \mid \mathbf{C}')$ para cada J. Para la pareja de covariables J definida por temperatura-tipo de suelo se obtuvo $\pi(J \mid \mathbf{C}') = .9889$, mientras que para la pareja J' determinada por las covariables humedad-temperatura se obtuvo $\pi(J' \mid \mathbf{C}') = .0111$. Para las otras parejas de covariables se obtuvo una probabilidad posterior menor que .0002. Con base en esas cantidades se tiene evidencia de que para el establecimiento de esta especie, la pareja de covariables que es relevante y cuyos valores son considerados cuidadosamente por la especie, es la pareja temperatura-tipo de suelo.

De los resultados que se obtuvieron y por los comentarios hechos por el experto al observar los mapas de potencial generados con cada una de las metodologías, el método que se propone en esta tesis es el que arrojó las zonas de alto potencial más adecuadas.

4.1.2 Zona para Proteger

El mapa de probabilidades que se obtuvo en la sección anterior (Figura 4-1(c)) se utilizó para encontrar la zona que se propondrá para proteger a la especie, de acuerdo con las ideas del Capítulo 2. Para aplicar la metodología que se propone es necesario definir los elementos $c(s)$, para cada $s \in R$ y z.

Con respecto al valor $c(s)$, se recurrió a definirlo utilizando los mapas auxiliares que se describen en la Sección 3.2.1. El mapa de costos obtenido y que se utiliza en esta aplicación se presenta en la Figura 2-6(b). Para fijar cantidades que representen el presupuesto con que se cuenta se consideró un mecanismo similar al descrito en la Sección 2.6, es decir, se propone definir los presupuestos con base en el costo asociado a "proteger" todos los nodos de la región, el cual se calcula como $C_T = \sum_{s \in R} c(s)$. Para esta aplicación se obtuvo $C_T = 2'703,855$ pesos. Los niveles que se consideran para el presupuesto son $B = .05C_T$ (presupuesto bajo) y $B = .3C_T$ (presupuesto alto), es decir, los presupuestos que se postulan son $B = 135,192$ pesos y $B = 405,578$ pesos.

Para determinar el valor biológico que se postula para la especie, es decir, la cantidad z, se aplicaron las ideas propuestas en la Sección 3.2.3. Así, utilizando el mapa de presencia *a priori* para esta especie dado por el experto y el mapa de costos se obtuvo el valor $z = 516$.

Para el parámetro β se consideran los valores $\beta = 0$, $\beta = 5$ y $\beta = 10$. Estos valores se

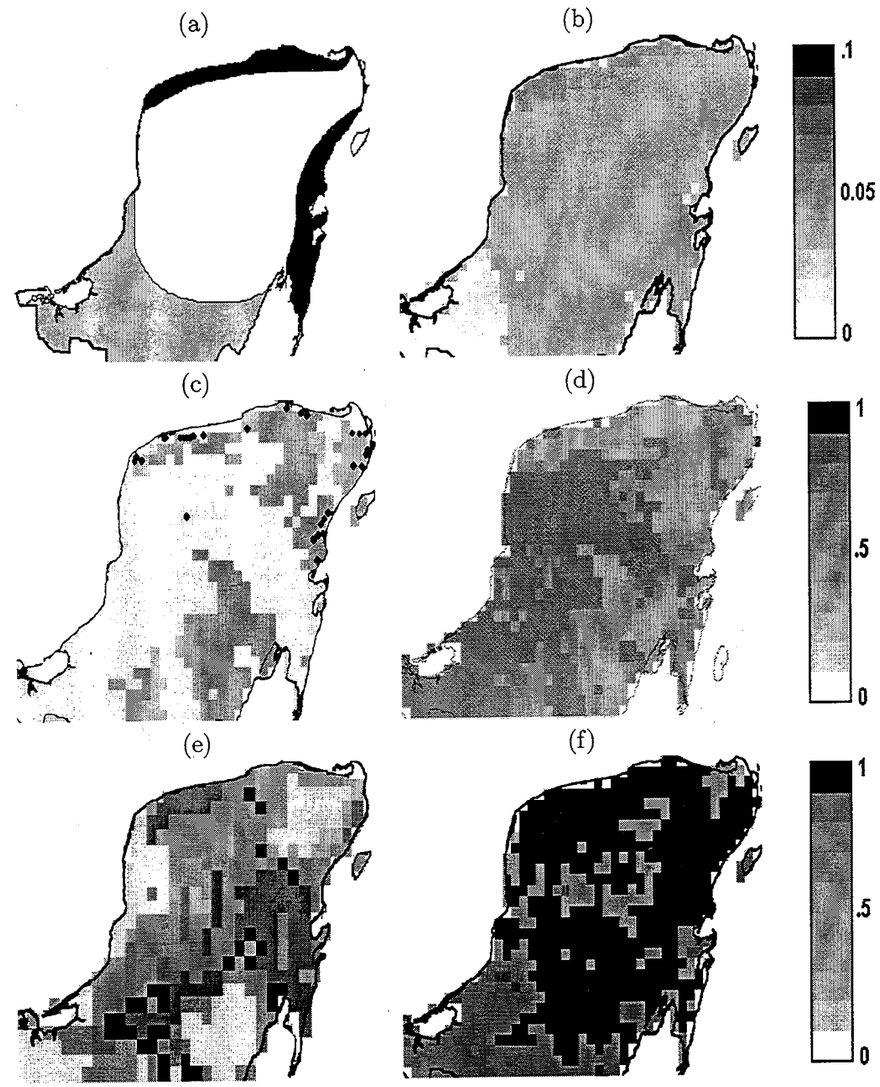

Figura 4-1: *C. readii* (a) Mapas *a priori* dados por el experto: la región en color negro indica la zona de presencia *a priori*, la región en gris indica la zona de ausencia *a priori* y la zona en color blanco indica falta de información. (b) Mapa de probabilidades de presencia *a priori*. (c) Mapa de probabilidades de presencia estimado con el método propuesto y sitios de presencia. (d) Mapa de certidumbre. (e) Mapa de Potencial estimado mediante FloraMap. (f) Mapa de Potencial estimado mediante Domain.

determinaron después de inspeccionar diferentes valores para este parámetro, como se sugiere en la Sección 2.6. Los valores que se postulan para el presupuesto y para el parámetro β permiten observar 6 escenarios. Para cada uno de ellos la búsqueda del conjunto de nodos que se propone para proteger se realizó utilizando las ideas que se proponen en la Sección 2.4.

Con respecto a los diferentes presupuestos asignados, con el presupuesto alto ($B = 405,578$ pesos) es posible adquirir una mayor cantidad de nodos para proteger a la especie, como se esperaba. Compare las Figuras 4-2(a) y (d), Figuras 4-2(b) y (e) y Figuras 4-2(c) y (f). Cada pareja de figuras corresponde a un mismo valor de β, y la primera figura citada de cada pareja corresponde al presupuesto bajo ($B = 135,192$ pesos), mientras que la segunda figura citada corresponde al presupuesto alto ($B = 405,578$ pesos). Los nodos que se seleccionaron para ser protegidos se encuentran localizados en zonas donde el costo de los nodos no es alto. Compare el mapa de costos desplegado en la Figura 2-6(b) con cada una de las regiones para proteger obtenidas desplegadas en la Figura 4-2.

Con respecto a la preferencia por proteger zonas no fragmentadas, se observa que la zona que se propone para proteger cuando se postula el valor $\beta = 10$ es menos fragmentada que la zona que se obtiene cuando se postula el valor $\beta = 0$ ó $\beta = 5$. Compare las Figuras 4-2(a), (b) y (c), las cuales se obtuvieron considerando el presupuesto $B = 135,192$ pesos, o las Figuras 4-2(d), (e) y (f), las cuales se obtuvieron considerando el presupuesto $B = 405,578$ pesos. En cada secuencia de gráficas se observa las regiones a proteger obtenidas con cada uno de los valores de β.

Para un presupuesto fijo sean A_0^*, A_5^* y A_{10}^* las regiones óptimas obtenidas con los valores postulados $\beta = 0$, $\beta = 5$ y $\beta = 10$, respectivamente. De acuerdo con las ideas introducidas en la Sección 2.6, para un presupuesto fijo es posible comparar las regiones que se obtienen al utilizar los valores $\beta = 0$, $\beta = 5$ y $\beta = 10$, comparando la pérdida esperada de las regiones A_5^* y A_{10}^* con la pérdida esperada de la región A_0^*. Las tablas que resultan para esta aplicación se presentan enseguida.

Tabla 6: Comparación de pérdida esperada, $B = 135,192$ pesos.

β	% pérdida
0	0
5	3.26%
10	3.21%

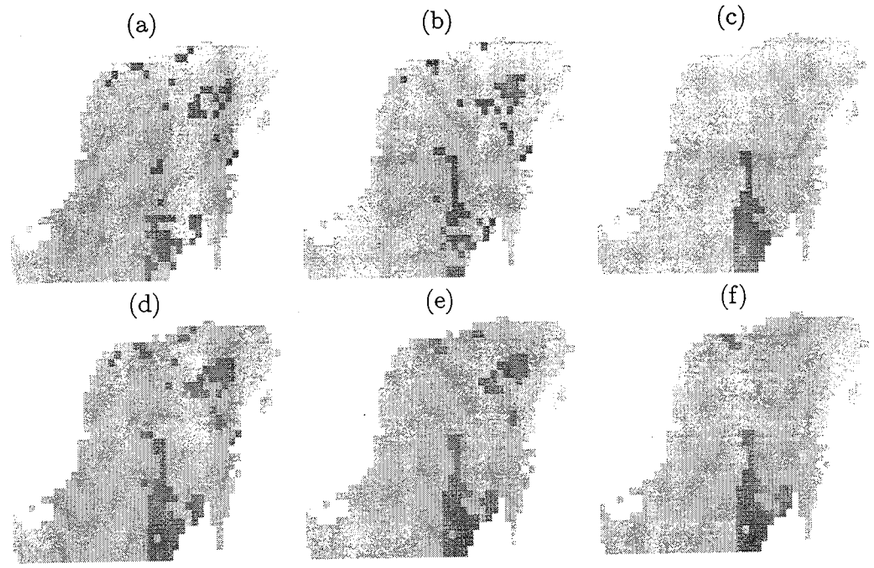

Figura 4-2: Zonas para proteger a *C. readii* obtenidas considerando (a) $B = 135,192$ y $\beta = 0$ (b) $B = 135,192$ y $\beta = 10$ (c) $B = 135,192$ y $\beta = 15$ (d) $B = 405,578$ y $\beta = 0$ (e) $B = 405,578$ y $\beta = 10$ (f) $B = 405,578$ y $\beta = 15$.

Tabla 7: Comparación de pérdida esperada, $B = 405,578$ pesos.

β	% pérdida
0	0
5	4.91%
10	6.38%

De acuerdo con la Tabla 6, para el presupuesto $B = 135,192$ pesos la región que se propondrá para proteger es A_0^*, pues con las regiones obtenidas con $\beta = 5$ y $\beta = 10$ se obtiene una pérdida esperada 3.26% y 3.21% mayores que la correspondiente a $\beta = 0$. Sin embargo, si se desea una región menos fragmentada, se podrá optar por proteger la región A_{10}^*, que representa un aumento en pérdida esperada menor que la obtenida al considerar A_5^*.

Para el presupuesto $B = 405,578$ pesos, la región que se propondrá para proteger es A_0^*, pues nuevamente el considerar una región menos fragmentada produce un aumento en la pérdida esperada. En la Tabla 7 se observa que con $\beta = 5$ se pierde 4.91% más que con $\beta = 0$, y con $\beta = 10$ se pierde 6.38% más que con $\beta = 0$. Si se desea proteger una región menos fragmentada se optará por A_5^*, que representa un menor aumento en la pérdida esperada que la obtenida con A_{10}^*.

4.2 Varias Especies

Las especies que se consideran en esta aplicación se enlistan en la Tabla 8, junto con el número de registros de presencia con que se cuenta para cada una de ellas (n_i) y los valores w_i y z_i de las que depende la función de pérdida que se utiliza.

Tabla 8: Especies bajo estudio y cantidades de interés

Especie	n_i	a_i	w_i	z_i (en pesos)
Carlowrightia myriantha	5	12	0.213	92.27
Coccothrinax readii	67	119	0.021	516.98
Furcraea cahum	25	50	0.051	92.27
Gaussia maya	28	12	0.213	24.69
Gonolobus yucatanensis	7	58	0.044	83.71
Lonchocarpus longistylus	4	114	0.022	422.07
Mammilaria gaumeri	13	28	0.092	85.06
Matelea aenea	7	17	0.150	90.54
Matelea yucatanensis	41	320	0.008	1220.97
Pterocereus gaumeri	57	40	0.064	137.19
Stenandrium nanum	6	25	0.102	79.57
Xanthosoma yucatanense	17	125	0.020	1026.74

Algunos aspectos relevantes acerca de las especies consideradas se presentan enseguida. *Carlowrightia myriantha* (Standl.) Standl. pertenece a la familia de las Acanthaceae y es considerada especie rara. *Coccotrhinax readii* H. J. Quero pertenece a la familia de las Arecaceae, y como se comenta en la Sección 4.1, se considera especie amenazada. *Furcraea cahum* Trel. pertenece a la familia de las Agavaceae. A pesar de encontrarse ejemplares de esta especie en los tres estados que conforman la península de Yucatán, se considera especie rara por ser escasa en los lugares donde se establece. La especie *Gaussia maya* (O. F. Cook) H. J. Quero & Read pertenece a la familia Palmae. Esta especie es apreciada por uso ornamental. *Gonolobus yucatanensis* (Woodson) W. D. Stevens pertenece a la familia Asclepiadaceae. Se considera que posee distribución restringida. *Lonchocarpus longistylus* Pittier pertenece a la familia Leguminosae. De su corteza se extrae, por fermentación, una bebida conocida como Balché. Esta especie es apreciada por su uso medicinal. *Mammilaria. gaumeri* (Britton & rose) Orc. pertenece a la familia Cactaceae. Es de uso ornamental y se considera que posee distribución restringida. *Matelea aenea* (Standl.) Woodson pertenece a la familia Asclepiadaceae y se considera especie rara y de distribución restringida. *Matelea yucatanensis* (Standl.) Woodson pertenece a la familia Asclepiadaceae y crece en la duna

costera. *Pterocereus gaumeri* (Britton & Rose) Th. MacDoug. & Miranda pertenece a la familia Cactaceae y se considera de distribución restringida. *Stenandrium nanum* (Standl.) T. F. Daniel pertenece a la familia Acanthaceae. Se considera especie rara. Por último, *Xanthosoma yucatanense* Engl. pertenece a la familia Araceae y se considera que posee distribución moderada. Todas estas especies se consideran amenazadas (o por lo menos vulnerables) debido a las altas tasas de deforestación que se han registrado en la Península.

4.2.1 Mapas de Probabilidades de Presencia y de Certidumbre

Los mapas de probabilidades de presencia se obtuvieron utilizando información *a priori* dada por el experto, según las ideas de la Sección 3.1.1. Los mapas *a priori* proporcionados por el experto del CICY se presentan en la Figura 4-7. Para estas especies no se presenta el mapa de potencial de presencia obtenido con los métodos alternos que existen.

Los mapas de probabilidades de presencia y de certidumbre para cada una de las especies se despliegan como sigue: *C. myriantha,* Figuras 4-3(a) y (b), *C. readii* Figuras 4-3(c) y (d), *F. Cahum* Figuras 4-3(e) y (f), *G. maya* Figuras 4-4(a) y (b), *G. yucatanensis* Figuras 4-4(c) y (d), *L. longistylus* Figuras 4-4(e) y (f), *M. gaumeri* Figuras 4-5(a) y (b), *M. aenea* Figuras 4-5(c) y (d), *M. yucatanensis* Figuras 4-5(e) y (f), *P. gaumeri* Figuras 4-6(a) y (b), *S. nanum* Figuras 4-6(c) y (d) y *X. yucatanense* Figuras 4-6(e) y (f).

De esas figuras se observa que las especies que pueden catalogarse como de amplia distribución potencial de acuerdo con los mapas de probabilidades de presencia que se obtuvieron son *C. readii, G. yucatanensis, L. longistylus, M. yucatanensis* y *X. yucatanense.* Por otro lado, las especies que resultaron con distribución restringida son *M. gaumeri* y *P. gaumeri.* La primera es una especie que habita principalmente en la duna costera y en selva baja y la segunda habita en la selva baja caducifolia. Estas especies conviven en la selva baja caducifolia con cactaceas candelabriformes. Este hecho se percibió en los mapas de probabilidades de presencia obtenidos para estas especies, en los cuales se observa que estas especies comparten la mayor parte de su área de distribución.

Las especies para las que se cuenta con mayor certidumbre acerca del mapa de probabilidades de presencia obtenido son *M. gaumeri* y *P. gaumeri.* Las especies para las que se obtiene certidumbre moderada son: *C. myriantha, C. readii, F. cahun, G. yucatanensis, L. longistylus, S. nanum, X. yucatananse* y *M. yucatanensis*, y las especies cuyo mapa de probabilidades de presencia posee la menor certidumbre son *G. maya* y *M. aenea.* El experto del CICY observó los mapas de probabilidades de presencia obtenidos y realizó los siguientes comentarios al respecto.

- *C. Myriantha.* Las zonas de alta probabilidad de presencia, las cuales se presentan principalmente en el centro de la Península de acuerdo con lo que se observa en la Figura 4-3(a), se catalogan como adecuadas. Con base en los factores ambientales que

se presentan en la Península, esta especie podría extender su distribución hacia las zonas que presentan probabilidad de presencia menores desplegadas en el mapa.

- *C. readii.* Los comentarios pertinentes con respecto al mapa de probabilidades de presencia se encuentran en la Sección 4.1.1.

- *F. cahum.* El mapa de probabilidades de presencia obtenido (Figura 4-3(e)) se considera adecuado, aunque se cree que se sobrestima la probabilidad de presencia de la especie en la región de la "subpenínsula", localizada al sureste de la Península.

- *G. maya.* El mapa de probabilidades de presencia, que se observa en la Figura 4-4(a), se califica como adecuado en las zonas donde se reporta la máxima probabilidad de presencia. El experto manifestó cierta duda de que la especie pueda extender su distribución hacia el noreste, como lo sugiere el mapa de probabilidades. Sin embargo, en estas zonas donde existe duda, se observa la menor certidumbre (Figura 4-4(b)). Este ejemplo permite valorar la información adicional que puede proporcionar el mapa de certidumbre.

- *G. yucatanensis.* Las zonas de alta probabilidad de presencia se calificaron, en su mayoría, como adecuadas, excepto en la región localizada al sureste de la Península. Sin embargo, el mapa de certidumbre observado en la Figura 4-4(d), reporta menor certidumbre para los nodos de esa zona.

- *L. longistylus.* El mapa de probabilidades de presencia (Figura 4-4(e)) se calificó como adecuado. Ya que para esta especie se cuenta únicamente con cuatro registros de presencia, las zonas de alta probabilidad potencial no alcanzan probabilidades mayores de .4. A pesar de que sólo se cuenta con 4 registros de presencia para esta especie, el mapa de certidumbre posee menor incertidumbre que el correspondiente mapa para la especie *G. maya*, para la que se cuenta con un total de 28 sitios de presencia. Como se comenta en el último párrafo de esta sección, este hecho observado en los mapas de certidumbre podría aportar evidencia acerca de la especificidad de hábitat de la especie.

- *M. gaumeri.* Las zonas de alta probabilidad de presencia obtenidas (Figura 4-5(a)) fueron calificadas como excelentes, con la sospecha de que la región de alto potencial puede extenderse hacia la costa de la porción noroeste de la Península.

- *M. aenea.* Se sospecha que es una especie endémica relativamente nueva. La zona en la que se obtuvo las mayores probabilidades de presencia, en la zona norte de la Península (Figura 4-5(c)), se califica como adecuada. La zonas con alta probabilidad de presencia localizadas al sureste de la Península se tomaron con reserva. Observe

que para estas zonas el mapa de certidumbre desplegado en la Figura 4-5(d) muestra mayor incertidumbre.

- *M. yucatanensis.* El mapa de probabilidades de presencia (Figura 4-5(e)) fué calificado como muy bueno. El mapa observado en la Figura 4-5(e) reporta un nivel alto de certidumbre para la mayoría de los nodos.

- *P. gaumeri.* El experto calificó el mapa de probabilidades de presencia, que se observa en la Figura 4-6(a), como bueno. Sin embargo, se afirma que la zona de habitación potencial de esta especie se extiende también en la región centro del Estado de Yucatán. Aunque no se obtuvo región de alta probabilidad de presencia en esa zona, el mapa de certidumbre (Figura 4-6(b)) califica esa zona como poseedora de cierta incertidumbre.

- *S. nanum.* El experto expresó cierta duda acerca de que la zona de alta probabilidad de presencia para esta especie (Figura 4-6(c)) se encuentre en la región sur de la Península. Sin embargo, este hecho no se niega de manera categórica, siendo que otras especies endémicas también presentan áreas de distribución disjuntas.

- *X. yucatanense.* Se sabe que esta especie se asocia con la presencia de cenotes, por lo que se espera que la región de mayor probabilidad de presencia se encuentre en el estado de Yucatán. Por esta razón existe cierta duda acerca de las zonas de alta probabilidad de presencia localizadas en la costa sureste de la Península (Figura 4-6(e)). Sin embargo, las zonas obtenidas como de alta probabilidad de presencia localizadas al sur se la Península se califican como adecuadas para el establecimiento potencial de la especie.

Al comparar los mapas de probabilidades de presencia se observaron las siguientes características. La especie *S. nanum* posee áreas de distribución potencial similares a las de *C. readii.* Sin embargo, la primera especie no se califica como de amplia distribución debido a que las probabilidades posteriores son sustancialmente menores que las obtenidas para la segunda especie.

Las especies *G. maya* y *G. yucatanensis* presentan zonas de alta probabilidad de presencia similares. Sin embargo, se observa mayor certidumbre en el resultado obtenido para *G. yucatanensis*, (Figura 4-4(d)) que en el resultado obtenido para *G. maya* (Figura 4-4(b)).

Las especies *M. gaumeri* y *P. gaumeri* presentan distribuciones similares. Se dice que estas dos especies tienen una alta especificidad de hábitat, lo que se concluye de que las regiones de alta probabilidad de presencia para estas especies son reducidas. Los correspondientes mapas de certidumbre permiten inferir que las zonas de alta probabilidad potencial son confiables.

Un hecho interesente se presenta para la especie *P. gaumeri.* Al delimitar los mapas *a priori* para esta especie, el experto proporcionó una región hacia el centro de la Península como área de alto potencial (ver Figura 4-7(j)). Aunque el mapa de probabilidades de

presencia no arrojó altas probabilidades en la zona referida (Figura 4-6(e)), el mapa de certidumbre (Figura 4-6(f)) presenta una certidumbre menor para la probabilidad de presencia en dicha zona. Combinando estos resultados puede concluirse que la probabilidad de presencia de dicha zona no es del todo confiable, por lo que puede deducirse que la probabilidad de establecimiento en dicha zona puede ser en realidad mayor que la obtenida.

Las zonas de mayor probabilidad de presencia obtenidas para las especies *C. readii* (Figura 4-3(c)) y *X. yucatanense* (Figura 4-6(e)) son similares, con probabilidades de presencia mayores observadas para la primera. Los mapas de certidumbre (Figuras 4-3(d) y 4-6(f)) para estas especies son relativamente similares, por lo que puede concluirse que estas especies comparten las zonas de mayor potencial.

El mapa de certidumbre puede aportar evidencia acerca de la especificidad del hábitat de una especie, de acuerdo con la siguiente observación realizada. El mapa de probabilidades de presencia para la espece *M. gaumeri*, para la que se cuenta con 13 sitios de presencia, posee mayor certidumbre que el mapa de probabilidades obtenido para la especie *G. maya*, para la que se cuenta con 28 sitios de presencia, y que el mapa probabilidades de la especie *M. yucatanensis*, para la que se cuenta con 41 sitios de presencia. Este hecho puede interpretarse de la siguiente manera: la especie *M. gaumeri* posee alta especificidad de hábitat, es decir, los pocos registros de presencia ocurren específicamente en ciertos vectores de covariables, por lo que pocos sitios de presencia son suficientes para obtener, con alta certidumbre, el área de mayor potencial de establecimiento de la especie. La especie *M. yucatanensis* se interpreta como poseedora de menor especificidad de hábitat y por lo tanto los 41 registros de presencia ocurrieron en una mayor diversidad de valores de covariables. Para esta especie sería necesario obtener un mayor número de registros de presencia para identificar los vectores de covariables preferidos por la especie, lo que permitirá obtener un mapa de probabilidades de presencia con mayor certidumbre.

4.2.2 Zona para Proteger

Utilizando las ideas introducidas en el Capítulo 2, se procedió a encontrar las regiones que se propondrán para proteger las 12 especies bajo estudio (Tabla 8). El valor $p_i(s)$ para la i-ésima especie en el nodo s es la probabilidad de presencia obtenida en la sección anterior.

El mapa en el que se observa el costo de cada nodo se presenta en la Figura 2-6(b), el cual se obtuvo utilizando las ideas introducidas en la Sección 3.2.1. Los valores w_i, $i = 1, ..., 12$, se obtuvieron utilizando la zona de presencia *a priori* proporcionada por un experto, los cuales se presentan en las Figuras 4-7(a)-(l). Las especies que según el experto deben considerarse de amplia distribución son *M. yucatanensis* y *X. yucatanense,* mientras que las especies que se consideran de distribución restringida son *G. maya, M. Gaumeri, M. aenea, P. gaumeri y S. nanum.*

El tamaño del área de distribución de cada especie, es decir, la cantidad a_i definida en la

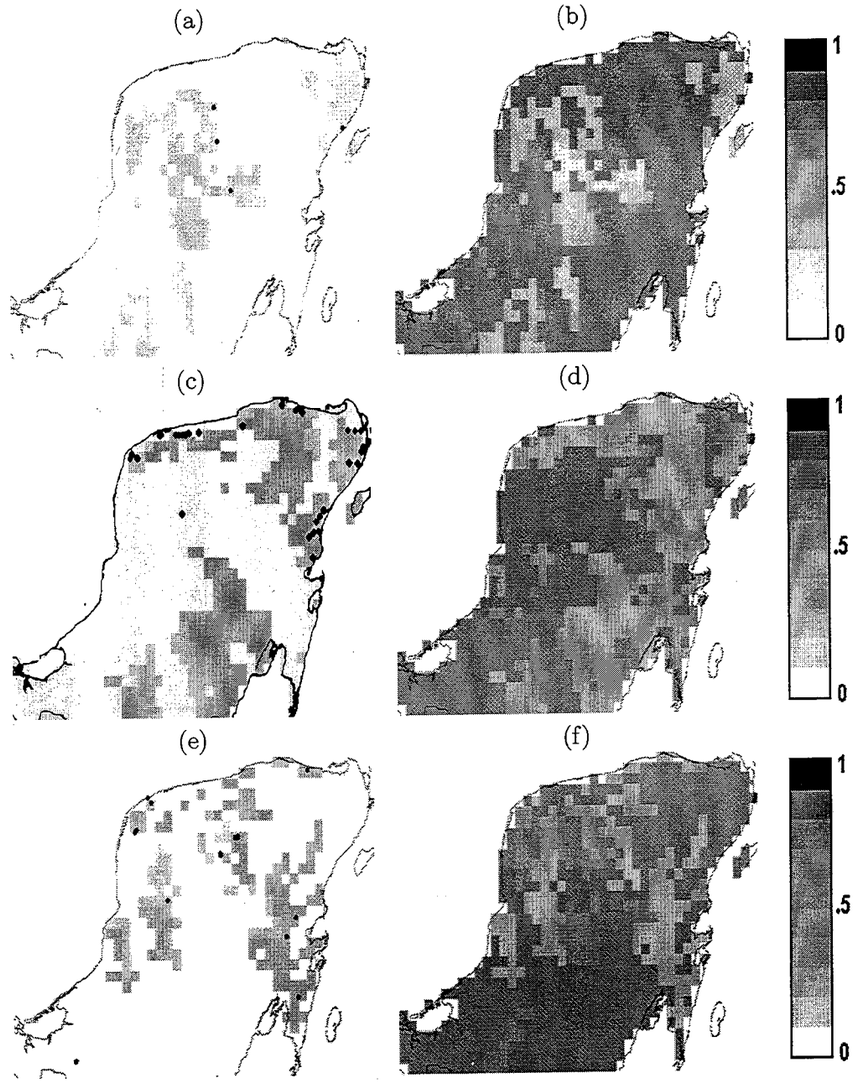

Figura 4-3: *C. myriantha*: (a) Mapa de probabilidades de presencia. (b) Mapa de certidumbre. *C. readii*: (c) Mapa de probabilidades de presencia. (d) Mapa de certidumbre. *F. cahum*: (e) Mapa de probabilidades de presencia. (f) Mapa de certidumbre.

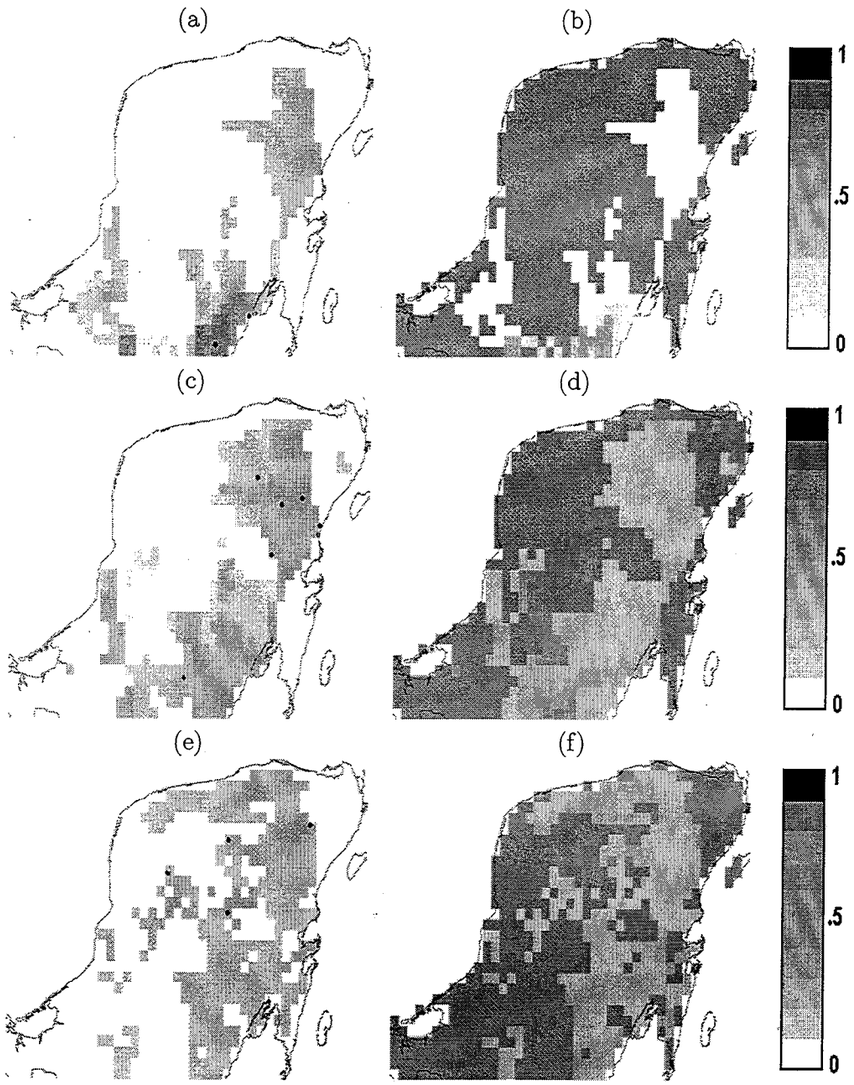

Figura 4-4: *G. maya*: (a) Mapa de probabilidades de presencia. (b) Mapa de certidumbre. *G. yucatanensis*: (c) Mapa de probabilidades de presencia. (d) Mapa de certidumbre. *L. longistylus*: (e) Mapa de probabilidades de presencia. (f) Mapa de certidumbre.

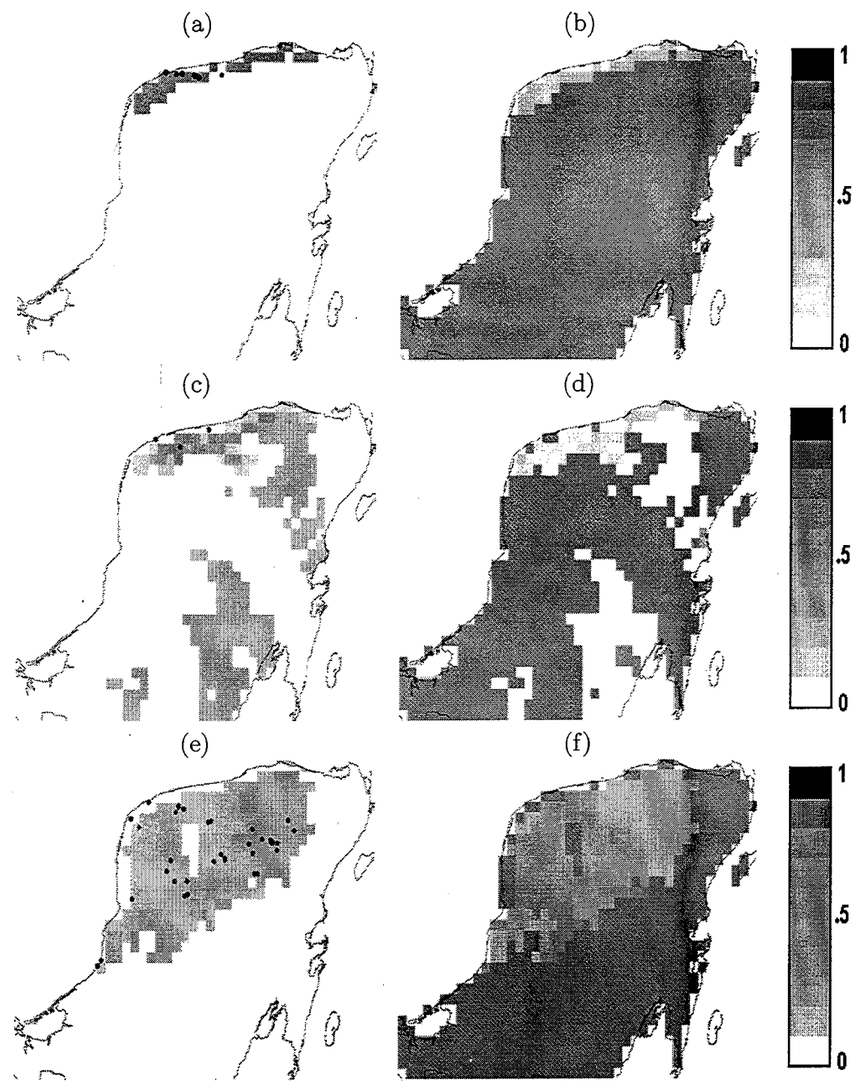

Figura 4-5: *M. Gaumeri*: (a) Mapa de probabilidades de presencia. (b) Mapa de certidumbre. *M. aenea*: (c) Mapa de probabilidades de presencia. (d) Mapa de certidumbre. *M. yucatanensis*: (e) Mapa de probabilidades de presencia. (f) Mapa de certidumbre.

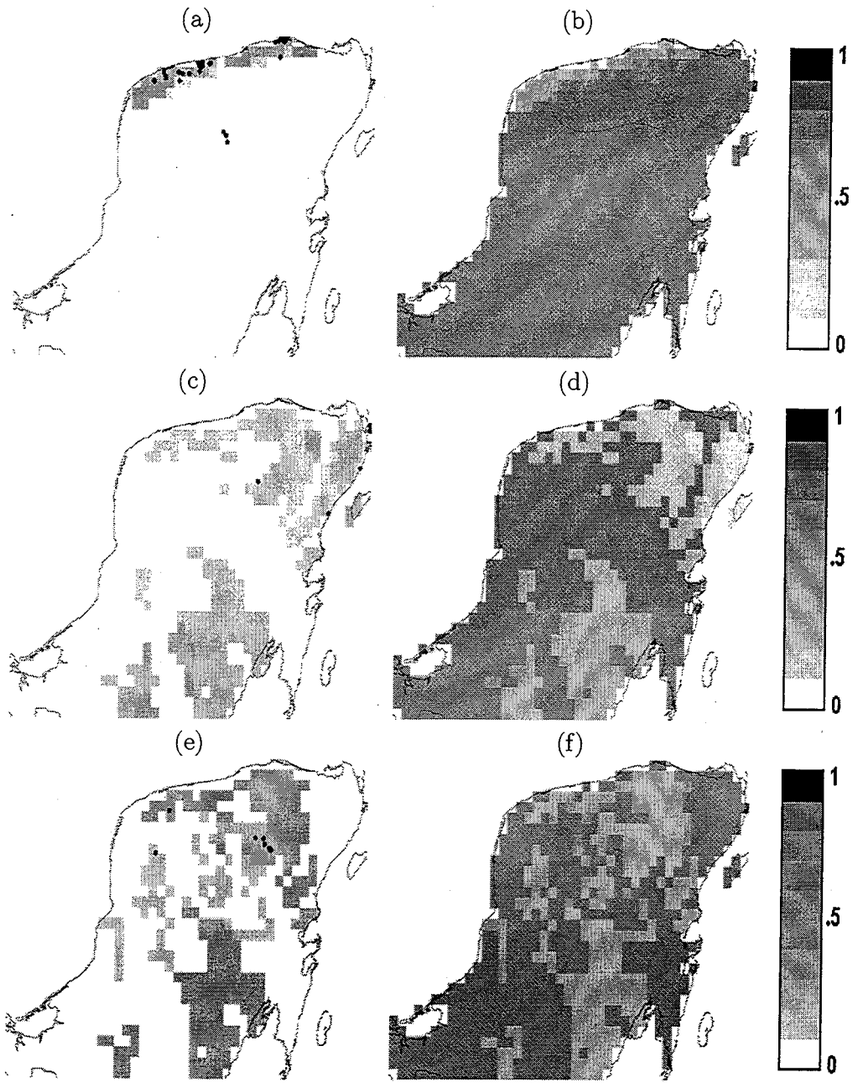

Figura 4-6: *P. gaumeri*: (a) Mapa de probabilidades de presencia. (b) Mapa de certidumbre. *S. nanum*: (c) Mapa de probabilidades de presencia. (d) Mapa de certidumbre. *X. yucatanense*: (e) Mapa de probabilidades de presencia. (f) Mapa de certidumbre.

Sección 3.2.2 se presenta en la Tabla 8, en la que también se presenta el respectivo valor w_i, $i = 1, ..., 12$, obtenido mediante las expresiones (3.3) y (3.4). De acuerdo con estos valores, las especies *C. myriantha*, *G. maya* y *M. aenea* deberán protegerse con mayor énfasis, mientras que las especies que de acuerdo con los elementos con que se cuenta se consideran menos importantes para proteger son *C. readii*, *L. longistylus*, *M. yucatanensis* y *X. yucatanense*.

Las cantidades z_i se postularon utilizando las ideas propuestas en la Sección 3.2.3. Los valores que resultaron se presentan en la última columna de la Tabla 8. Se observa que las especies que para esta aplicación se consideran más valiosas para proteger son *M. yucatanensis* y *X. yucatanense*, mientras que las especies consideradas menos valiosas para proteger son *G. maya* y *S. nanum*.

Con respecto al presupuesto con que se cuenta se consideran las cantidades $B = .05C_T$ (presupuesto bajo) y $B = .15C_T$ (presupuesto alto). La cantidad C_T es la misma que la utilizada en la Sección 4.1.2, por lo que los presupuestos que se consideran en esta aplicación son $B = 135,192$ pesos y $B = 405,578$ pesos, respectivamente.

Para el parámetro β se consideran los valores $\beta = 0$, $\beta = 5$ y $\beta = 10$, los cuales se seleccionaron después de experimentar asignando diversos valores a este parámetro, como se sugiere en la Sección 2.3. Los valores postulados para cada uno de los factores permiten observar un total de 6 escenarios, cuyos resultados se presentan en los siguientes párrafos.

Como en los resultados que se observaron en la aplicación presentada en la Sección 4.1, se observa que el parámetro β permite obtener zonas con diferentes niveles de fragmentación. En la Figura 4-8 se observa que las zonas obtenidas utilizando el valor $\beta = 10$ poseen menos nodos dispersos en la región a proteger que las zonas obtenidas utilizando el valor $\beta = 0$ ó $\beta = 5$. Compare por ejemplo las Figuras 4-8(a), (b) y (c), las cuales corresponden al presupuesto $B = 135,192$ pesos, y las Figuras 4-8(d), (e) y (f), que corresponden al presupuesto $B = 405,578$ pesos.

Al comparar las gráficas de la Figura 4-8 con los mapas de probabilidades de presencia de cada una de las especies (Figuras 4-3, 4-4, 4-5 y 4-6), se observa que estas zonas contienen nodos correspondientes a la zona de alta probabilidad de la mayoría de las especies consideradas. La especie para la que se pone menos énfasis para ser protegida es *F. cahun*. De acuerdo con la Tabla 8, esta especie es considerada poco importante para ser protegida, pues posee el valor de $w_3 = .051$, y comparada con las demás especies, fue catalogada como poseedora de valor biológico intermedio ($z_3 = 92.27$).

Por otro lado, de las especies para las que se cuenta con información *a priori* dada por un experto, la especie *G. maya* posee la mayor importancia para ser protegida, pues se postula $w_4 = .213$ para esta especie. Comparando los mapas que se proponen para proteger desplegados en la Figura 4-8 con el mapa de probabilidades de presencia para esta especie (Figura 4-4(a)), se observa que las regiones que se proponen para proteger incluyen, en cada caso, nodos que poseen alta probabilidad de presencia de establecimiento de esta especie. Estos ejemplos ilustran la manera en la que las cantidades w_i y z_i son relevantes

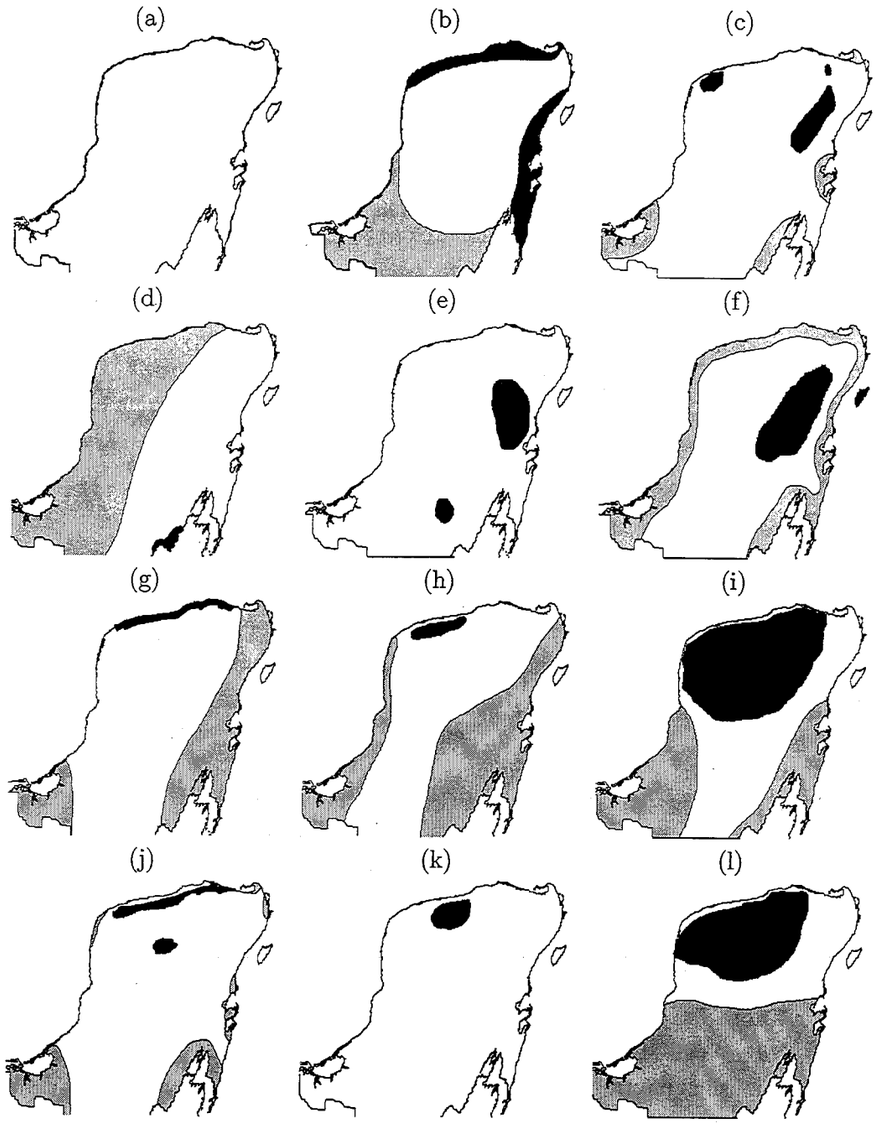

Figura 4-7: Mapas *a priori* dados por el experto: (a) *C. myriantha* (b) *C. readii* (c) *F. cahum* (d) *G. maya* (e) *G. yucatanensis* (f) *L. longistylus* (g) *M. gaumeri* (h) *M. aenea* (i) *M. yucatanensis* (j) *P. gaumeri* (k) *S. nanum* (l) *X. yucatanense*. Las regiones en color negro corresponden a las zonas de presencia *a priori*. Las regiones en color gris corresponden a las zonas de ausencia *a priori*. Las regiones en blanco indican zonas en las que el experto no aportó información.

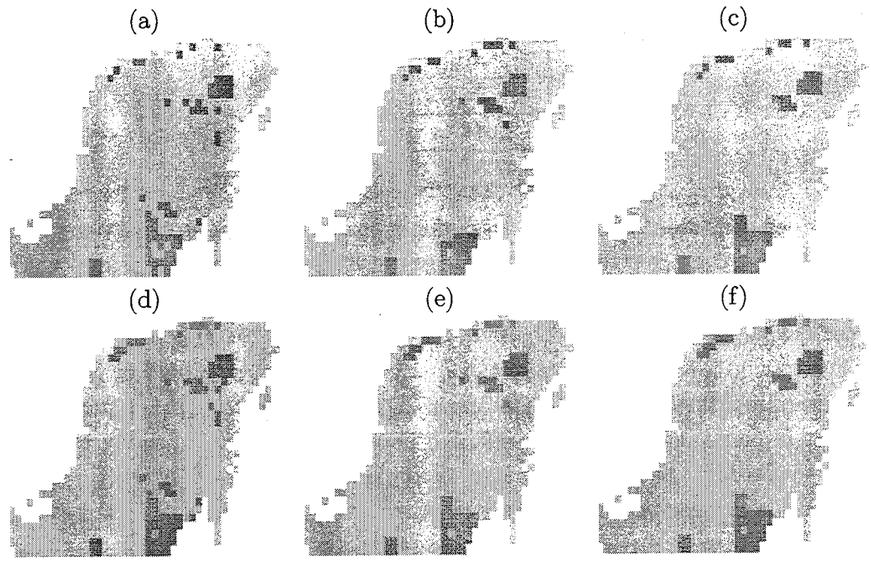

Figura 4-8: Zona para proteger obtenida para las 12 especies considerando los valores (a) $B = 135,192$, $\beta = 0$, (b) $B = 135,192$, $\beta = 5$, (c) $B = 135,193$, $\beta = 10$, (d) $B = 405,578$, $\beta = 0$, (e) $B = 405,578$, $\beta = 5$, (f) y $B = 405,578$, $\beta = 10$.

para determinar la zona que se propondrá para proteger.

En el caso del presupuesto bajo ($B = 135,192$ pesos), las especies cuya zona de alta probabilidad de presencia se encuentran mejor representadas en las zonas para proteger son *C. readii, G. maya, G. yucatanensis, M. aenea, M. yucatanensis, S. nanum y X. yucatanense* (Figuras 4-8(a), (b) y (c)). Si se considera el presupuesto alto ($B = 405,578$ pesos), las zonas a proteger incluyen, además, nodos correspondientes a las regiones de alta probabilidad de presencia de las especies *M. gaumeri* y *P. gaumeri* (Figuras 4-8(d), (e) y (f)).

Para decidir la región que se propondrá para proger se procede a comparar las regiones obtenidas por medio de las correspondientes pérdidas esperadas. Para esto se construyen las tablas de porcentaje de pérdida relativa, como se propone al final de la sección 2.6. Las tablas correspondientes al presupuesto $B = 135,192$ pesos y $B = 405,578$ pesos se presentan en las Tablas 9 y 10, respectivamente.

Tabla 9: Comparación de Pérdida Esperada, $B = 135,192$.

β	% Pérdida
0	0
5	-20.17%
10	-2.33%

Tabla 10: Comparación de Pérdida Esperada, $B = 405,578$.

β	% Pérdida
0	0
5	22.44%
10	42.86%

Para un presupuesto fijo sean A_0^*, A_5^* y A_{10}^* las regiones propuestas para proteger obtenidas con los valores $\beta = 0$, $\beta = 5$ y $\beta = 10$, respectivamente. Si se considera el presupuesto bajo, o sea, $B = 135,192$ pesos, de la Tabla 9 se observa que con la región A_5^* la pérdida esperada es 20.17% menor que la pérdida esperada obtenida con A_0^*. Por su parte, la pérdida esperada obtenida con la región A_{10}^* es 2.33% veces menor que la correspondiente pérdida obtenida con A_0^*. De estos resultados se concluye que la región A_5^* será propuesta para ser protegida.

Si se considera el presupuesto alto ($B = 405,578$ pesos), de la Tabla 10 se observa que con la región A_5^* se pierde 22.44 por ciento más que con la región A_0^*. En la misma tabla se observa que la pérdida esperada obtenida con la región A_{10}^* es 42.86% mayor que la pérdida obtenida con A_0^*. En este caso la región que se propondrá para proteger es A_0^*.

Un hecho interesante de las zonas que resultan para proteger es el siguiente. Espadas, Durán y Argáez (2003) encontraron las denominadas *áreas de endemismo* sobre la península de Yucatán con base en 162 especies, entre las cuales se encuentran las estudiadas en esta sección. Las áreas de endemismo son relevantes debido a que proporcionan conocimiento acerca de la historia evolutiva de las especies. Las zonas que se obtuvieron como zonas para proteger considerando las 12 especies bajo estudio, las cuales se observan en la Figura 4-8, se localizan principalmente en áreas de endemismo descritas en Espadas, Durán y Argáez (2003). Este hecho aporta una prueba empírica de que la metodología que se propone en el Capítulo 2 permite obtener zonas para proteger con base en información relevante acerca de las especies.

Capítulo 5

Extensiones y Conclusiones

5.1 Extensiones

Las ideas para considerar las parejas de covariables que se presentan en las Secciones 1.1 y 1.2 pueden extenderse de manera sencilla al caso de considerar, de manera explícita, interacciones de covariables de orden mayor. Tanto las ideas introducidas para realizar el proceso de elicitación como la forma de proceder para obtener las cantidades de interés pueden aplicarse si se considera cualquier orden de interacción de covariables. Sin embargo, al observar los resultados que se han obtenido en el ejercicio de simulación (Sección 1.5) y en las aplicaciones a especies de interés (Capítulo 4), no se espera que al considerar un modelo más complejo se produzca mejoría substancial en los resultados.

Aunque en esta tesis solamente se consideran explícitamente dos restricciones, es posible considerar otras restricciones o condiciones que se requiera imponer a la región que se proponga para proteger. Para esto bastará (1) restringir el espacio de posibles soluciones, o (2) agregar en la función de pérdida un término adicional por cada restricción que se considere. La forma en la que cada posible restricción se considerará en el problema de decisión dependerá de la naturaleza de las mismas. En caso de añadir un término adicional en la función de pérdida, deberá tenerse cuidado de que las unidades en las que se encuentre medido cada factor adicional sea comparable con las unidades de los términos que ya hayan sido considerados, o bien, considerar un parámetro adicional como el β de la expresión (2.6).

Las ideas introducidas en el Capítulo 2 parecen ser aplicables al siguiente problema: suponga que a un grupo de Υ expertos se les pide que evalúen a N sujetos. Cada experto determina un conjunto de atributos de una lista predeterminada que, de acuerdo con su apreciación, describe mejor al sujeto. El objetivo es medir el grado de acuerdo de los expertos en su evaluación de los sujetos. En este contexto se obtiene un acuerdo si todos los expertos coinciden en incluir o excluir un atributo de sus conjuntos de atributos seleccionados. Este problema ha sido abordado por Kupper y Hafner (1989) para el caso $\Upsilon = 2$. Para el caso general no existe un procedimiento estándar, pero sí algunas sugerencias. Por ejemplo,

Gordon (1977) sugiere debilitar la definición de lo que se entiende por acuerdo, considerando que se obtiene un acuerdo entre los Υ expertos si $\Upsilon - 1$ de ellos coinciden en incluir (o exluir) un atributo determinado de sus listas. Esta propuesta se parece a la definición de lo que es una $(\Upsilon - 1)$-contradicción bajo las ideas de la Sección 3.1.1. Por su parte, Kupper y Hafner (1989) proponen, para el caso $\Upsilon = 3$, obtener y promediar los acuerdos de todas las parejas de expertos y así obtener una medida de acuerdo global.

Si se postula que un atributo f es contradictorio si es incluido en por lo menos una lista pero no en las Υ listas proporcionadas por los expertos, un atributo se define como *contradictorio* para el conjunto de $\Upsilon - i$ expertos si no es incluido en las $\Upsilon - i$ listas dadas por los expertos, $0 \leq i \leq \Upsilon - 2$. Con esta idea es posible considerar la cantidad P de la Definición 7 como una medida del acuerdo entre expertos, con Q dada por la expresión (3.6). Cada una de las $\Upsilon - 2$, $\Upsilon - 3$, ..., 2-contradicciones puede considerarse como una definición más débil de lo que es una contradicción, de acuerdo con la idea propuesta por Gordon (1977). En este contexto, las cantidades $\gamma_{h,l}$ pueden interpretarse como el peso que se asigna a un acuerdo entre los correspondientes l expertos. Así, las ideas presentadas en la Sección 3.1.1 pueden servir como base para proponer una forma estándar de proceder para medir la coherencia entre Υ expertos.

Con respecto al problema de encontrar zonas para proteger, un trabajo a realizar en el futuro es incluir el mapa de certidumbre, que se obtiene junto con el mapa de probabilidades de presencia, en el proceso de decidir la región que se propondrá para proteger. En esta tesis se utiliza solamente el mapa de probabilidades de presencia para encontrar la región para proteger. Ya que el mapa de certidumbre califica al mapa de probabilidades, su inclusión permitirá obtener una zona para proteger a la que se dotará de alguna medida de certidumbre.

5.2 Conclusiones

La metodología que se propone para estimar las regiones de alto potencial de presencia posee varias ventajas sobre los métodos que actualmente existen. Una ventaja importante radica en que se define de manera formal el concepto de potencial. Al definir el potencial de presencia por medio de una probabilidad, los resultados son interpretables de manera natural. A pesar de que la notación puede no parecer sencilla, las ideas relevantes que sustentan la metodología son simples y permiten implementar la metodología sin mayor dificultad.

Podría argumentarse que considerar únicamente parejas de covariables para postular el modelo (1.8) es restrictivo. Sin embargo, el modelo que se propone es flexible y posee una cantidad razonable de parámetros. Más aún, intuitivamente se observa que al considerar todas las parejas de covariables se incorpora información con respecto a interacciones de covariables de orden mayor de manera indirecta.

El método que se propone en el Capítulo 1 parece sufrir menos de sobrestimación que los métodos que actualmente se usan para abordar el problema de inferir las zonas de alto

potencial de presencia. Será necesario realizar más experimentos para confirmar esta afirmación. También será necesario realizar un estudio detallado para investigar el grado de sensibilidad del método con respecto a la forma de especificar el sesgo espacial, es decir, la cantidad $\delta(s)$ para cada s en la retícula.

Una aportación relevante de esta tesis al problema de encontrar regiones de alta probabilidad de presencia de una especie es la postulación de una medida de certidumbre para los resultados que se obtienen. En la práctica esto permitirá realizar inferencias con respecto a las zonas de establecimiento potencial de la especie con base en mayor información. Más aún, la metodología que se propone permite incorporar como parte de la inferencia la información que un experto puede aportar con respecto a las zonas de establecimiento de las especies consideradas. Con respecto a la forma de proceder para encontrar una región para proteger que se aborda en el Capítulo 2, la función de pérdida que se propone funcionó de manera adecuada. Las cantidades de interés se relacionan con información relevante acerca de las especies.

Los enfoques bajo los que se aborda el problema de encontrar regiones para proteger utilizan en su mayoría la perspectiva de programación lineal. Bajo este enfoque no es posible abordar problemas con retículas con una cantidad grande de nodos, pues no será posible encontrar una solución debido a restricciones computacionales. El enfoque que se presenta en el Capítulo 2 no posee ese problema. Aunque en nuestro caso no es posible realizar la búsqueda de la solución de manera exhaustiva, los algoritmos heurísticos que se implementaron permiten encontrar una solución en relativamente poco tiempo, la cual es óptima con respecto a la función de pérdida.

Las ideas introducidas en el Capítulo 3 permiten convertir el conocimiento del experto en valores para los hiperparámetros que definen la distribución *a priori* para el parámetro de interés. El concepto de contradicción introducido es conceptualmente simple y fácil de utilizar en el proceso de elicitación.

Referencias

Aké, A., Jiménez, J. y Ruenes, M. (1999). El solar Maya. En: A. García y J. Córdoba (coords). Atlas de Procesos Territoriales de Yucatán. Facultad de Arquitectura. UADY.

Arita, H. T., Figueroa, F., Frisch, A., Rodríguez, P. y Santos del Prado, K. (1997). Geographical range size and the conservation of mexican mammals. *Conservation Biology* **11,** 92-100.

Arrow, K., Solow, R., Leamer, E., Portney, P., Randner, R. y Schuman, H. (1993). Report of the NOAA panel on contingent valuation. Federal Register **58**(10), 4602-4614.

Austin, M. P. (2002). Spatial prediction of species distribution: an inference between ecological theory and statistical modelling. *Ecological Modelling* **157,** 101-118.

Baz, A. y García-Boyero, A. (1996). The SLOSS dilemma: a butterfly case study. *Biodiversity and Conservation* **5**, 493-502.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory.* Chichester, Wiley.

Bertsimas, D. y Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science* **8**(1), 10-15.

Brown, J., Stevens, G. C. y Kaufman, D. W. (1996). The geographic range: size, shape, boundaries and internal structure. *Annual Review of Ecology and Systematics* **27**, 597–623.

Busby, J. R. (1991). BIOCLIM - A bioclimate analysis and prediction system. En *Nature conservation: cost effective biological surveys and data analysis.* Margules, C. R. y Austin, M. P. (eds) pp. 64-68 (CSIRO Australia).

Camm, J. D., Norman, S. K., Polasky, S. y Solow, A. R. (2002). Nature reserve site selection to maximize expected species covered. *Operations Research* **50,** 946-955.

Carpenter, G., Gillison, A. N. y Winter, J. (1993). Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2,** 667–680.

Cohn, H. y Fielding, M. (1998). Simulated annealing: searching for an optimal temperature schedule. *SIAM Journal of Optimisation* **9**(3), 779-802.

Costello, C. y Polasky, S. (2003). Dynamic reserve site selection. *Resource and Energy Economics* (En prensa).

DeGroot, M. (1970). *Optimal Statistical Decisions.* McGraw-Hill.

De Oliveira, V. (2000). Bayesian prediction of clipped gaussian random field. *Computational Statistics and Data Analysis* **34,** 299–314.

Drechsler, M. y Wätzold, F. (2001). The importance of economic cost in the development of guidelines for spatial conservation managenent. *Biological Conservation* **97,** 51-59.

Espadas, C., Durán, R. y Argáez, J. (2003). Phytogeographic analysis of taxa endemic to the Yucatan peninsula using geographic information systems, the Domain heuristic method and parsimony analysis of endemicity. *Diversity and Distributions* **9**(4), 313-330.

García, A. y Alonzo, A. (1999). Regionalización Económica. En: A. García y J. Córdoba (coords). Atlas de Procesos Territoriales de Yucatán. Facultad de Arquitectura. UADY.

Gaston, K. y Blackburn, T. (2000). *Pattern and Process in Macroecology.* Blackwell Science. Oxford.

Gelman, A., Carlin, J. B., Stern, H. S. y Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman & Hall.

Geman, S. y Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

Gordon, A. D. (1977). A measure of the agreement between rankins. *Biometrika* **66,** 7-15.

Heagerty, P. J. y Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93,** 1099-1111.

Heikkinen, J. y Högmander, H. (1994). Fully bayesian approach to image restoration with an application in biogeography. *Applied Statistics* **43,** 569–582.

Högmander, H. y Möller, J. (1995). Estimating distribution maps from atlas data using methods of statistical image analysis. *Biometrics* **51,** 393–404.

Jones, P. G. y Gladkov, A. (1999). FloraMap: a computer tool for predicting the distribution of plants and other organisms in the wild; version 1, 1999. Editado por Annie L. Jones. CIAT CD-ROM Series. Cali, Colombia: Centro Internacional de Agricultura Tropical.

Kupper, L. L. y Hafner, K. B. (1989). On assessing interrater agreement for multiple attribute responses. *Biometrics* **45,** 957-967.

Malcolm, S. A. (2001). Sequential land acquisition decision for nature reserves under acquisition and population uncertainty. Operational Research. Department of food and resource economics, Collegue of Agriculture and Natural Resources. University of Delaware.

McDonnell, M., Possingham, H. P., Ball, I. R. y Cousins, E. (2002). Mathematical methods for spatially cohesive reserve design. *Journal of Environmental Modelling and Assessment* **7**(1), 107-114.

Peterson, A. T. y Cohoon, K. P. (1999). Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological modelling* **117,** 159–164.

Peterson, A. T., Soberón, J. y Sánchez-Cordero, V. (1999). Conservatism of ecological niches in evolutionary time. *Science* **285,** 1265-1267.

Peterson, A. T., Stockwell, D. R. B. y Kluza, D. A. (2002). Distributional prediction based on ecological niche modeling of primary occurrence data. En: *Predicting species occurrences: issues of scale and accuracy.* Scott, J. M., Heglund, P. J. y Morrison, M. L. (eds) pp. 617-623. Island Press, Washington, D. C.

Pettitt, A., Weir, I. y Hart, A. (2002). A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing* **12,** 353-367.

Polasky S., Camm, J. D. y Garber-Yonts, B. (2001). Selecting biological reserves cost-effectively: an application to terrestrial vertebrate conservation in Oregon. *Land Economics* **77**(1), 68-78.

Robert, C. P. y Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer. New York.

Sánchez-Cordero, V. y Martínez-Meyer, E. (2000). Museum specimen data predict crop damage by tropical rodents. *Proceedings of the National Academy of Science of the United States of America* **97**(13), 7074-7077.

Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66,** 783-801.

Soberón, J., Golubov, J. y Sarukhán, J. (2001). The importance of Opuntia in Mexico and the routes of invasion and impact of *Cactoblastis cactorum. Florida Entomologist* **84,** 486-492.

Stockwell, D. R. B. y Noble, I. R. (1991). Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation* **32,** 249–254.

Stockwell, D. R. B. y Peters, D. (1999). The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13,** 143–158.

Udvardy, M. (1969). *Dynamic Zoogeography.* Van Nostrand Reinhold Company. New York.

Winkler, L. R. (1967a). The assessment of prior distributions in bayesian analysis. *Journal of the American Statistical Association* **62,** 776-800.

Winkler, L. R. (1967b). The quantification of judgment: some methodological suggestions. *Journal of the American Statistical Association* **62,** 1105-1120.

Winkler, L. R. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association* **64,** 1073-1078.