



COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

**Modelos bayesianos para la distribución de
especies con registros de solo presencias**

Bartolo de Jesús Villar Hernández

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO
2014

La presente tesis titulada: **Modelos bayesianos para la distribución de especies con registros de solo presencias**, realizada por el alumno: **Bartolo de Jesús Villar Hernández**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA ESTADÍSTICA

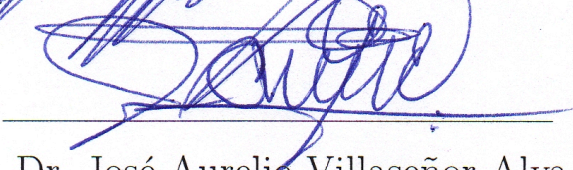
CONSEJO PARTICULAR

CONSEJERO



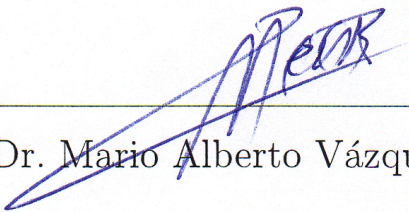
Dr. Sergio Pérez Elizalde

ASESOR



Dr. José Aurelio Villaseñor Alva

ASESOR



Dr. Mario Alberto Vázquez Peña

Montecillo, Texcoco, México, Mayo de 2014

Modelos bayesianos para la distribución de especies con registros de solo presencias.

Bartolo de Jesús Villar Hernández

Colegio de Postgraduados, 2014.

Resumen

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

Keywords and phrases : registros de solo presencia, modelos de distribución de especies, probabilidad de ocurrencia, proceso Poisson no homogéneo, Maxlike, Maxent, enfoque bayesiano.

Bayesian models for species distribution modelling with only-presence records.

Bartolo de Jesús Villar Hernández

Colegio de Postgraduados, 2014.

Abstract

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

Keywords and phrases : only-presence records, species distribution models, ocurrence probability, inhomogeneous poisson proces, Maxlike, Maxent, bayesian approach.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado para realizar mis estudios de maestría.

Al Colegio de Postgraduados, por haberme brindado la oportunidad de seguir mi formación académica y profesional en sus aulas.

A los integrantes de mi Consejo Particular:

Dr. Sergio Pérez Elizalde, por su excelente dirección y apoyo para la realización del presente trabajo.

Dr. José Aurelio Villaseñor Alva, por sus atinadas observaciones, consejos y ayuda desinteresada en la realización del presente trabajo.

Dr. Mario Alberto Vázquez Peña, por sus observaciones y apoyo desinteresado.

A mis profesores, amigos y compañeros de clases.

DEDICATORIA

Con amor inmenso para mis hijos: Aleshka y Caín.

A mi esposa.

A mis padres: Evodio y Carmen.

A mis hermanos: César, Paúl y Roberto.

Índice

1. Introducción	1
2. Objetivos	4
2.1. General	4
2.2. Particular	4
3. Modelos de distribución de especies	5
3.1. Marco conceptual para modelar distribuciones de especies	6
3.1.1. Teoría ecológica	7
3.1.2. El modelo de datos	7
3.1.3. El modelo estadístico	8
3.2. ¿Por qué es necesario modelar distribuciones de especies?	9
3.3. El espacio geográfico versus el espacio medioambiental	10
3.4. Datos biológicos y los datos mediambientales	12
3.4.1. Los datos biológicos	12
3.4.2. Los datos medioambientales	15

4. Métodos de Aprendizaje Automático	18
4.1. Árboles de Regresión y Clasificación(CART)	18
4.1.1. Un ejemplo sencillo	19
4.1.2. Árboles de regresión	21
4.1.3. Árboles de clasificación	23
4.1.4. Utilidad de los árboles de decisión	25
4.1.5. Conjunto de métodos aplicados a los árboles de decisión: bagging, boosting y random forest	26
4.2. Redes Neuronales Artificiales (ANNs)	27
4.3. Algoritmos genéticos (GAs)	29
4.4. Máquinas de soporte vectorial (SVM)	30
5. Modelos estadísticos	32
5.1. El modelo lineal	32
5.2. Modelo lineal generalizado (GLM)	33
5.2.1. El modelo de regresión logística	34
5.2.2. Supuestos en el modelo de regresión logística	35
5.2.3. Verosimilitud del modelo de regresión logística	36
5.2.4. Estimación por máxima verosimilitud en el modelo de regresión logística	36
5.2.5. Transformaciones en los predictores y selección	38
5.2.6. Implementación de modelos GLM	39

ÍNDICE

5.3. Modelos aditivos generalizados(GAM)	39
5.3.1. Ajuste de Modelos Aditivos	41
5.3.2. Uso de modelos GAM en MDEs	43
5.4. Splines de Regresión Multivariada Adaptativa (MARS)	44
5.4.1. Descripción matemática de los MARS	45
5.4.2. Algunas aplicaciones de MARS en MDEs	48
5.5. Máxima entropía (MaxEnt)	48
5.5.1. Prólogo	49
5.5.2. Covariables y sus transformaciones (features)	50
5.5.3. Explicación de MaxEnt	51
5.6. Método basado en la máxima verosimilitud (<i>Maxlike</i>)	55
5.7. Modelo de Proceso Poisson no Homogéneo (IPP) para datos de “solo presencias”	58
5.7.1. Probabilidad de ocurrencia vs tasa de ocurrencia	58
5.7.2. Notación	59
5.7.3. Descripción general del modelo	59
5.7.4. Máxima verosimilitud para el modelo IPP	60
5.8. Enfoques bayesianos en Modelos de distribución de especies	62
5.8.1. Algoritmo de Metrópolis-Hastings	63
6. Evaluación y selección de modelos en MDEs	65

ÍNDICE

6.1. Datos para la evaluación de los modelos	65
6.1.1. ¿Cómo se miden los errores?	66
6.1.2. La elección del umbral	67
6.1.3. Área bajo la curva ROC (AUC)	68
6.1.4. Evaluación en modelos de <i>solo presencias</i>	70
6.2. Selección de modelos	71
6.2.1. Criterio de Información de Akaike	71
6.2.2. Criterio de información de la Devianza	72
7. Propuesta de modelos bayesianos para modelar la distribución de especies con registros de <i>solo presencias</i>	73
7.1. Modelo <i>MaxBayes</i>	73
7.2. Modelo <i>IPPBayes</i>	74
8. Caso de estudio	76
8.1. Simulación de datos	76
8.2. Datos de género <i>Dalea</i>	77
8.3. Simulación de la distribución <i>a posteriori</i> mediante MCMC	81
9. Resultados y discusión	82
9.1. Datos de simulación	82
9.2. Género <i>Dalea</i>	87

ÍNDICE

10. Conclusiones y recomendaciones	92
10.1. Conclusiones	92
10.2. Recomendaciones	93
Referencias	93
Apéndice	98
.1. Pruebas de convergencia para el modelo <i>MaxBayes</i> e <i>IPPBayes</i>	98
.1.1. Datos simulados	98
.1.2. Datos género <i>Dalea</i>	100
.2. Código R	103

Índice de tablas

3.1. Algunos ejemplos de fuentes de información biológica y medioambiental utilizados en MDEs.	17
4.1. Datos del ejemplo árboles de decisión	19
6.1. Matriz de confusión para dos clases, presencia-ausencia de la especie.	67
6.2. Medidas de precisión en función del umbral para datos binarios. . . .	67
8.1. Nombre de covariables medioambientales	80
8.2. Covariables medioambientales	81
9.1. Resumen de <i>MaxBayes</i> para distintos n (simulación).	83
9.2. Prevalencia estimada por <i>MaxBayes</i> bajo diferentes n (simulación). .	83
9.3. Resumen de <i>IPPBayes</i> para distintos n_1 (simulación).	86
9.4. <i>MaxBayes</i> vs <i>IPPBayes</i> en términos del DIC (Simulación)	87
9.5. Resumen de <i>MaxBayes</i> y <i>Maxent</i> para distintos tamaños de muestra.	87
9.6. Resumen del modelo <i>IPPBayes</i>	90
9.7. <i>MaxBayes</i> vs <i>IPPBayes</i> en términos del DIC	90

ÍNDICE DE TABLAS

.1. Prueba de convergencia de Gelman y Rubin en el modelo <i>MaxBayes</i> (simulación).	100
.2. Prueba de convergencia de Gelman y Rubin en el modelo <i>IPPBayes</i> (simulación).	100
.3. Prueba de convergencia de Gelman y Rubin en los modelos <i>MaxBayes</i> e <i>IPPBayes</i>	102

Índice de figuras

3.1. Diagrama que muestra los componentes de un MDE.	6
3.2. Espacio geográfico vs espacio medioambiental.	11
3.3. Ejemplo de registros de <i>solo presencias</i>	14
4.1. Ejemplo de árboles de clasificación describiendo la relación entre la presencia/ausencia de <i>P. menziesii</i> y las variables explicativas.	20
4.2. Medida de la impureza de los nodos para la clasificación de dos clases como una función de la proporción p en la clase 2.	24
4.3. Diagrama esquemático de una red neuronal aplicada a MDEs.	28
4.4. Diagrama que ilustra un hiperplano en el caso de clases separables.	30
5.1. Forma de la función de respuesta (determinada usando un spline cúbico como <i>scatterplot suavizador</i>) entre la razón de log-verosimilitudes de presencia de la especie (eje y , etiquetada como “s(X)”) y las variables predictoras (eje x) estimados mediante un modelo GAM usando la distribución binomial (<i>logit link</i>) para datos de presencia-ausencia.	43
5.2. Las funciones base $(x - t)_+$ (línea en naranja) y $(t - x)_+$ (línea azul punteada) usados por MARS.	45
5.3. La función $h(X_1, X_2) = (X_1 - x_{51})_+ \cdot (x_{72} - X_2)_+$	47

ÍNDICE DE FIGURAS

5.4. Interfaz gráfica del software <i>MaxEnt</i>	50
5.5. Ejemplo esquemático de las densidades de probabilidad en el modelo <i>MaxEnt</i> utilizando dos variables, los datos de presencias y el <i>background</i> (tomada de Elith <i>et al.</i> (2011)).	52
6.1. Representación gráfica de Sensibilidad, Especificidad, kappa y AUC	69
8.1. Zona de estudio y Reserva Cuicatlán-Tehuacán.	78
8.2. Covariables medioambientales	79
9.1. Distribuciones <i>a posteriori</i> de los parámetros de <i>MaxBayes</i> para diferentes n (simulación).	84
9.2. Área bajo la curva ROC (AUC) de <i>MaxBayes</i> (simulación).	85
9.3. Distribuciones <i>a posteriori</i> de los parámetros de <i>IPPBayes</i> para diferentes n (simulación).	86
9.4. Distribución potencial del género <i>Dalea</i> obtenidos mediante los modelos <i>Maxent</i> , <i>MaxBayes</i> e <i>IPPBayes</i>	88
9.5. Distribuciones <i>a posteriori</i> de los parámetros del modelo <i>MaxBayes</i>	89
9.6. Distribuciones <i>a posteriori</i> de los parámetros del modelo <i>IPPBayes</i>	89
.1. Cadenas simuladas del modelo <i>MaxBayes</i> (ejemplo simulación).	99
.2. Cadenas simuladas del modelo <i>IPPBayes</i> (ejemplo simulación).	99
.3. Cadenas simuladas de las distribuciones <i>a posteriori</i> del modelo <i>MaxBayes</i>	101
.4. Cadenas simuladas de las distribuciones <i>a posteriori</i> del modelo <i>IPPBayes</i>	102

Capítulo 1

Introducción

Uno de los temas centrales en ecología es el estudio de la distribución geográfica de especies tanto de flora como de fauna. Hoy en día, se han estado desarrollando modelos cuyo objetivo a grandes rasgos es modelar la distribución espacial de las especies de interés. Estos Modelos de Distribución de Especies (MDEs) se han utilizado para diversos propósitos, por ejemplo, han sido aplicados para estudiar la propagación de especies intrusas ([Thuiller *et al.*, 2005](#)), para investigar los impactos del cambio climático en la extinción de ciertas especies ([Thomas *et al.*, 2004](#)), para conocer la diversidad biológica de una zona en particular ([Graham *et al.*, 2006](#)), por citar solo algunos. En todas las aplicaciones de los MDEs, el problema central es utilizar la información de donde las especies han sido observadas (y donde no) y asociar ésta información con un conjunto de covariables medioambientales para determinar la probabilidad [o algún índice proporcional a ésta] de que una determinada especie pueda estar presente o no en sitios no muestreados ([Latimer *et al.*, 2006](#)).

El desarrollo de un MDE comienza con observaciones de presencia o ausencia de determinada especie [en determinado intervalo de tiempo] y de variables medioambientales asociadas a dichos registros, que influyen en la aptitud del hábitat y por lo tanto, en la distribución de la especie ([Franklin, 2009](#)). Con los modernos Sistemas de Información Geográfica (GIS), el investigador tiene acceso a estas variables medioambientales en toda una zona de interés y las utiliza como variables predictivas para estimar la probabilidad de ocurrencia de la especie en dicha área. Generalmente, el área de interés se divide en una malla de celdas del mismo tamaño, donde la elección del tamaño de las celdas quedará determinado según la resolución deseada

1. Introducción

por el investigador, y la probabilidad de presencia se generalizará para toda la celda.

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

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

En el presente trabajo se proponen dos metodologías, en el marco de inferencia bayesiana, que se han denominado *MaxBayes* e *IPPBayes*. Estas dos alternativas se han construido a partir de los modelos *Maxlike* e *IPP*. Para ilustrar de forma práctica dichas propuestas, se implementaron dos ejemplos de estudio: (1) se simuló re-

1. Introducción

registros de presencia-ausencia mediante un ensayo *Bernoulli* en una zona ficticia de 10,000 celdas en donde la prevalencia fue de 0.38, de las cuales, una vez descartadas las ausencias, se muestreo aleatoriamente las presencias que se utilizaron para ajustar los modelos y estimar sus respectivos parámetros, y en el caso de *MaxBayes* comparar la prevalencia estimada contra la real, y (2) se utilizaron registros de presencia del género *Dalea* provenientes de la zona de la reserva de la biosfera Tehuacán-Cuicatlán y se compararon las distribuciones potenciales arrojadas de ambos modelos con el software *Maxent*.

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

Capítulo 2

Objetivos

2.1. General

Proponer una nueva metodología estadística en el contexto de inferencia bayesiana para modelar distribuciones potenciales de especies con registros de *solo presencias*, a partir de un modelo logístico estimado por máxima verosimilitud y de un proceso Poisson no homogéneo, que permitan incorporar el conocimiento *a priori* acerca de la especie estudiada, para así estimar *a posteriori*, la probabilidad de presencia y la intensidad de ocurrencia, respectivamente.

2.2. Particular

- Ilustrar la implementación práctica de los modelos propuestos y comparar los resultados obtenidos, mediante dos ejemplos: utilizando un conjunto de datos generados mediante simulación y empleando un conjunto de datos reales del género *Dalea* en la reserva de la biósfera Tehuacán-Cuicatlán. En el último caso, los resultados también se compararán con los obtenidos con el software *Maxent*.

Capítulo 3

Modelos de distribución de especies

Los términos como “nicho ecológico” o “modelo de nichos de especies” se han utilizado para describir a los Modelos de distribución de especies (MDEs). Algunos autores incluso distinguen entre modelos de nichos ecológicos y modelos de distribución de especies. El primer término lo aplican para describir un modelo de distribución potencial (esto es, excluyendo la competencia biótica, interacciones y transformación del hábitat natural), mientras que el segundo lo utilizan para referirse a la distribución actual de la especie. ([Franklin, 2009](#)).

Los MDEs también han sido referidos como “modelos de aptitud de hábitat” y que describen que tan ideal es un sitio para albergar cierta especie. El concepto de “aptitud de un hábitat” está íntimamente relacionado con la idea de “función de selección de recursos” en biología, y que puede aplicarse tanto a flora como a fauna. Una función de selección de recursos (FSR) es cualquier función que es proporcional a la probabilidad de que un hábitat esté habitado por un organismo ([Manly *et al.*, 2002](#)). Si una “función de selección de recursos” es proporcional a la probabilidad de que un hábitat esté ocupado, entonces un MDE se podría decir que predice la verosimilitud de que un evento (especie) ocurra en una ubicación determinada, es decir, la probabilidad de presencia de la especie ([Franklin, 2009](#)).

Cuando un MDE es aplicado a un conjunto de datos de ocurrencias de especies, y a mapas que representan a las covariables medioambientales, como resultado se ob-

3.1. Marco conceptual para modelar distribuciones de especies

tiene un mapa de predicciones que representa la distribución geográfica potencial de la especie. A estos mapas resultantes se les conoce como “superficies de respuesta ecológica”, “modelos biogeográficos de distribución de especies”, “predicción espacial de distribución de especies”, entre muchos otros; una de las aplicaciones más importantes de las FSRs es la obtención de mapas de distribuciones de especies (Franklin, 2009).

3.1. Marco conceptual para modelar distribuciones de especies

Para Austin (2002), son tres los componentes que conforman un modelo de distribución de especies: el modelo ecológico, el modelo de datos y el modelo estadístico. El modelo ecológico consta del conocimiento ecológico y la teoría a utilizarse o por probarse en el estudio; puede contener suposiciones que necesitan incorporarse en el análisis o en las hipótesis que están a prueba. El modelo de datos involucran decisiones en relación a cómo los datos deben colectarse. El modelo estadístico implica la elección del método estadístico en función de los dos puntos anteriores.

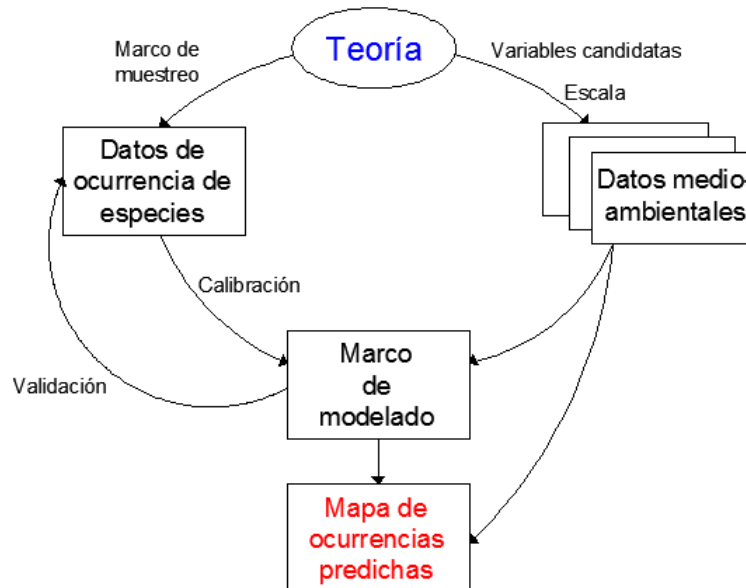


Figura 3.1: Diagrama que muestra los componentes de un MDE.

3.1. Marco conceptual para modelar distribuciones de especies

3.1.1. Teoría ecológica

En la mayoría de los artículos recientes sobre el tema, la teoría ecológica generalmente está implícita, es decir, se asume que las distribuciones de las especies están determinadas al menos en parte por las variables medioambientales y que se pueden realizar estimaciones razonables para estas variables. La respuesta de las especies también dependerá de la naturaleza de las variables medioambientales y de los procesos ecológicos asociados. Sin embargo, en la mayoría de los estudios las variables predictivas medioambientales se seleccionan en base a la disponibilidad de éstas y a la experiencia del investigador en el sentido de conocer que variables están correlacionadas con la distribución de cierta especie ([Austin, 2006](#)).

3.1.2. El modelo de datos

El modelo de los datos tiene muchos componentes, sin embargo, cuatro son los componentes básicos: el propósito, la escala del estudio, la disponibilidad de los datos y la selección de los predictores ([Austin, 2006](#)). La escala a la cual los datos están disponibles puede limitar el propósito para el cual los datos pueden usarse. A su vez, dos aspectos importantes de la escala son la extensión y la resolución; la extensión se refiere al área sobre el cual un estudio se lleva a cabo, mientras que la resolución se refiere al tamaño de la unidad de muestreo en la que los datos se registran. Por ejemplo, si el propósito es investigar el nicho ecológico, entonces la extensión del área de estudio debe extenderse más allá de los límites ambientales donde las especies han sido observadas.

Por otra parte, en la elección de los predictores debe de tomarse en consideración, cómo los predictores que han sido seleccionados, dependen de los procesos ecológicos y biofísicos a través de su influencia en la biota, por lo que se debe de tener en cuenta la naturaleza directa o indirecta de los posibles predictores y si se dispone del conocimiento ecofisiológico adecuado para la elección de los mismos. Por ejemplo, las variables indirectas tales como la altitud y la latitud pueden estar únicamente correlacionadas con los organismos a través de su correlación con variables directas tales como la temperatura y la precipitación y que pueden tener un impacto fisiológico en los organismos, por otra parte, la precipitación aunque tiene un efecto directo sobre las plantas, es una variable distante, sin embargo, está relacionada con el agua

3.1. Marco conceptual para modelar distribuciones de especies

disponible en la zona radicular de las plantas.

Un aspecto al que debe darse una gran importancia es a los errores en las variables medioambientales ya que pueden tener un gran impacto en los resultados que arrojan los modelos. [Van Neil *et al.* \(2004\)](#) estudiaron la influencia de los errores en Modelos de Elevación Digital (DEM). Variables tales como la pendiente se calculan a partir de un DEM y luego son incorporados a un sistema de información geográfica (SIG) desde donde son recuperados al momento de modelar. Los autores concluyeron que las variables directas (tales como la radiación solar) son menos propensas a errores que las variables indirectas de las cuales se calculan variables como la pendiente. La mayoría de los estudios obtienen sus predictores medioambientales de un GIS. Los errores que se generan al producir el GIS necesitan una evaluación cuidadosa antes de utilizar los predictores para un modelo de distribución de especies ([Austin, 2006](#)).

3.1.3. El modelo estadístico

Una de las tareas cruciales al momento de modelar distribuciones de especies es elegir el modelo estadístico que se empleará. Ésta tarea, está íntimamente relacionada con el modelo de los datos y la teoría ecológica que estamos asumiendo. Sin embargo, en la mayoría de las veces no se dispone de un modelo de datos perfectamente estructurado, en parte porque no se ha colectado la información de interés en un marco de muestreo consistente con la teoría ecológica que se quiere probar. Por tanto, se elegirá el modelo estadístico que mejor se adapte tanto a los objetivos del investigador, como a los datos disponibles que alimentaran el modelo.

Otro aspecto de importancia relevante es la evaluación de los modelos usados en predicción espacial tanto de especies de flora como de fauna, de aquí nacen las siguientes tres interrogantes: (1) ¿cómo comparar los numerosos modelos estadísticos disponibles? (2) ¿cómo se debe evaluar el ajuste de los modelos? (3) ¿cómo se debe evaluar la compatibilidad del modelo estadístico con el modelo ecológico?

La comparación de los diferentes métodos es una tarea complicada debido a que continuamente se han estado introduciendo nuevos métodos. Por citar solo un ejemplo, [Leathwick *et al.* \(2005\)](#) realizaron una comparación entre los métodos GAM y MARS para modelar distribuciones de peces de agua dulce y concluyeron que ambos

3.2. ¿Por qué es necesario modelar distribuciones de especies?

métodos proporcionan resultados similares, sin embargo, el algoritmo MARS tiene ventajas computacionales con respecto a GAM que hacen más fácil su implementación. La comparación de distintos métodos realizados por diferentes autores, son raramente o nunca comparables. Por ejemplo, [Araujo *et al.* \(2005\)](#) y [Elith *et al.* \(2006\)](#) compararon 4 y 16 métodos respectivamente, pero solo dos de ellos eran similares. Algunos modelos se alimentan de registros binarios (presencias-ausencias), mientras que otros están diseñados para alimentarse de registros de solo presencias provenientes de museos y herbarios; lo anterior hace que cualquier posible comparación se basa en el procedimiento en particular no en el método general.

Otro aspecto importante, como lo enfatiza [Araujo *et al.* \(2005\)](#), es la necesidad de utilizar datos independientes para la validación de los modelos. Dos métodos para validación (o evaluación como se acostumbra a llamar en la literatura) se describen someramente enseguida: *resustitución* cuando los mismos datos se utilizan para calibrar el modelo y para medir su ajuste y *división de los datos* cuando los datos son divididos en dos grupos aleatoriamente, uno para calibración y el otro para validación, o bien un grupo totalmente independiente de los dos anteriores (de una región distinta) se utiliza para validación.

3.2. ¿Por qué es necesario modelar distribuciones de especies?

Una razón es comprender la relación entre una especie y su entorno abiótico y biótico sobre la base de observaciones con el fin de realizar inferencia ecológica, o para probar la hipótesis ecológicas y biogeográficas sobre la distribución de las especies y sus rangos ([Franklin, 2009](#)).

Los MDEs hoy día, son ampliamente utilizados para interpolar y extrapolar a partir de puntos de observaciones sobre el espacio geográfico, similar a un pronóstico en el tiempo, con el fin de *predecir* la ocurrencia de una especie para ubicaciones donde se carece de datos del muestreo. Los mapas de probabilidades de presencia (mapas predictivos), también son requeridos para muchos aspectos de manejo de recursos y planes de conservación incluso para evaluación de la biodiversidad, diseño de reservas, manejo de hábitats y restauración, estudio de especies invasivas y evaluación de los

3.3. El espacio geográfico versus el espacio medioambiental

riesgos que representan, restauración ecológica, y para predecir el efecto del cambio climático sobre las especies y los ecosistemas (Franklin, 2009).

3.3. El espacio geográfico versus el espacio medioambiental

Generalmente cuando se habla de ocurrencia de una especie se le relaciona inmediatamente al espacio geográfico, es decir, que la distribución de la especie es visualizada en un mapa. Para entender los modelos de distribución de especies es importante también entender que la ocurrencia de una especie puede verse desde el espacio medioambiental, el cual es un espacio conceptual definido por las variables medioambientales a las cuales la especie responde. El concepto de espacio medioambiental tiene sus cimientos en la teoría del nicho ecológico, y de acuerdo a la definición de Hutchinson (1957), el nicho fundamental de una especie lo constituye el conjunto de condiciones medioambientales dentro de las cuales la especie puede sobrevivir y persistir.

Visualizar la distribución de una especie tanto en el enfoque geográfico como en el espacio medioambiental nos permite definir algunos conceptos básicos cruciales cuando se modela la distribución de una especie (ver Figura 3.2). Note que los puntos de ocurrencia constituye todo lo que se sabe acerca de la *distribución actual* de la especie; es probable que la especie ocurra en otras áreas en las cuales aún no ha sido detectada (Figura 3.2, área A). Si la distribución actual se representa en el espacio medioambiental, entonces podemos identificar que parte del espacio medioambiental está ocupado por la especie y definimos a este espacio como *nicho ecológico* o nicho ocupado (Pearson, 2007). Hutchinson describe al nicho ecológico como aquella área que abarca la porción del nicho fundamental en la cual una especie no es excluida por la competencia biótica, por lo tanto, el nicho ecológico refleja todas las restricciones impuestas a la distribución actual, incluyendo restricciones espaciales relacionadas con la capacidad de dispersión, y las múltiples interacciones con otros organismos.

Si las condiciones medioambientales encapsuladas dentro del nicho fundamental se representa en el espacio geográfico, entonces lo que se tiene es la *distribución potencial*. Note que algunas regiones de la distribución potencial no pueden ser habitados por la especie (figura 3.2, áreas B y C), ya sea porque la especie ha sido excluida debido a

3.3. El espacio geográfico versus el espacio medioambiental

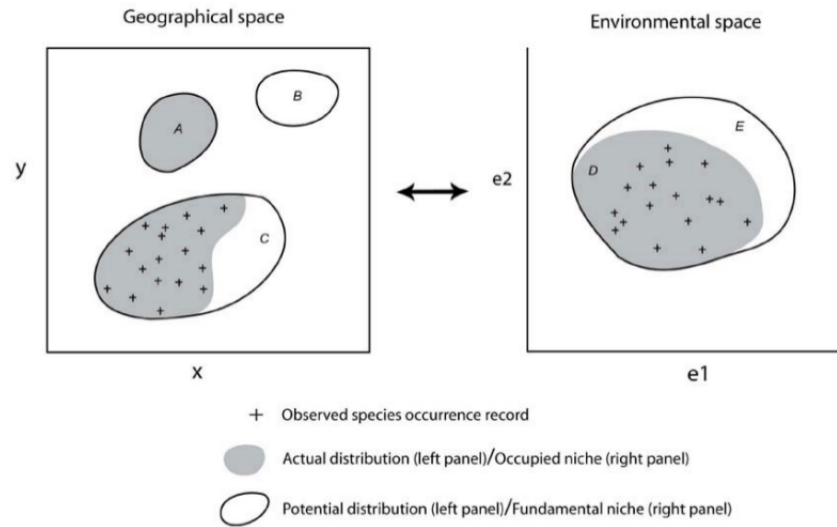


Figura 3.2: Espacio geográfico vs espacio medioambiental.

interacciones bióticas (por ejemplo la presencia de un competidor o la falta de fuente de alimentación); o bien porque la especie aún no se ha dispersado por el área (por ejemplo una barrera geográfica tal como una montaña o porque el tiempo no ha sido el suficiente aún para la dispersión); o bien porque la especie ha sido extinta del área (por ejemplo por la modificación humana del paisaje) (Pearson, 2007).

Cuando hablamos de modelos de distribución de especies, debemos tener en mente algunas consideraciones adicionales de suma importancia. Es importante apreciar que las variables medioambientales para la construcción de modelos es imposible que contenga todas las dimensiones posibles del espacio medioambiental, por tanto, las variables disponibles para modelar solo representan un subconjunto de los factores medioambientales que influyen en la distribución de las especies. Otro factor de suma importancia es considerar aquellos casos donde se observan especies en medioambientes inadecuados, como en el caso de especies de fauna que se trasladan de un lugar a otro y en ocasiones visitan zonas aledañas fuera del nicho fundamental; lógicamente uno esperaría que estos eventos ocurrieran con mayor frecuencia en especies con alta capacidad de dispersión, tales como las aves (Pearson, 2007).

Desde el punto de vista práctico, al implementar un modelo estadístico para encontrar una distribución potencial, por ejemplo *Maxent*, no existe razón para considerar al espacio geográfico x diferente del espacio medioambiental z , si se asume que z es muestreado aleatoriamente en lugar de x y comúnmente se denota en los artículos

3.4. Datos biológicos y los datos mediambientales

como $z(x)$ (Royle *et al.*, 2012).

3.4. Datos biológicos y los datos mediambientales

Los MDEs requieren básicamente dos tipos de información, el primero lo constituyen los registros de las especies que nos proporciona información acerca de la distribución conocida de la especie y denominados como datos biológicos. El segundo tipo de datos lo conforman los datos medioambientales que describen el paisaje donde la especie se encuentra. En esta sección discutiremos acerca de este tipo de información.

3.4.1. Los datos biológicos

Los datos que describen la distribución conocida de una especie puede provenir de una variedad de fuentes que podemos resumir de la forma siguiente:

1. *Colecciones personales*: los registros de ocurrencia pueden ser obtenidos mediante recorridos de campo por un pequeño grupo de investigadores.
2. *Muestreos extensos*: la información puede provenir de muestreos formales proveniente de un amplio grupo de investigación respaldado por instituciones de investigación y por cientos de voluntarios.
3. *Colecciones de museos y herbarios*: los registros pueden provenir de colecciones en museos de historia natural o herbarios. En México se dispone por ejemplo del herbario de la UNAM, el herbario del CICY, por citar un par de ejemplos.
4. *Recursos en línea*: una amplia fuente de información está disponible actualmente en internet, como por ejemplo www.gbif.org o concretamente en México en el portal de la base de datos REMIB de la CONABIO (<http://www.conabio.gob.mx/>).

3.4. Datos biológicos y los datos mediambientales

Registros de presencias-ausencias versus registros de solo presencias

Los datos de registros de especies pueden ser dos tipos: de presencia-ausencia (es decir, se disponen de registros donde la especie fue observada e información con cierto grado de fiabilidad de donde no) y registros de *solo presencias* (únicamente registros puntuales donde la especie fue observada). Diferentes modelos se aplican para modelar distribuciones en función de los datos disponibles, así por ejemplo, el modelo logístico es de los más socorridos cuando se trabajan con datos de presencias-ausencias, mientras que *Maxent* goza de mucha popularidad cuando los investigadores trabajan con datos de solo presencias.

Tal como menciona [Pearson \(2007\)](#), aunque en ocasiones se disponen de registros de ausencias, no siempre son del todo confiables; por ejemplo, a pesar de que la especie esté presente en una determinada área, aún no haya sido detectada. Este tipo de registros se le conoce como “falsas ausencias” y el modelo interpretará a estas falsas ausencias como verdaderas. La inclusión de falsas ausencias en un modelo debe tomarse con cautela ya que pueden sesgar el análisis de la información.

En la mayoría de las ocasiones, de lo único que se disponen son de registros de presencias. Estos registros de *solo presencias* tienen tres componentes: (i) una colección de ubicaciones puntuales donde la especie ha sido observada, (ii) un área de interés dividida por una malla de celdas, que el investigador considera como disponible para la especie (es decir, que la especie puede migrar a todas las ubicaciones dentro de esa área), y (iii) covariables medioambientales para cada una de las celdas de la malla ([Merow y Silander, 2014](#)). Las ubicaciones de los registros se asocian a la celda en la cual caen. Los modelos que trabajan con datos de *solo presencias* asumen que el grid de celdas han sido muestreadas aleatoriamente para detectar la presencia, y por lo tanto, los registros se producen en proporción a las preferencias de uso de hábitat de la especie. Independientemente de si se registran más de un ejemplar de la especie, un solo registro es asociado a la celda.

Un ejemplo de este tipo de datos se muestra en la Figura (3.3). Esta figura muestra todas las ubicaciones donde el género *Agave* fue observado entre los años 1980-2000. Note que estos registros no constituyen el total de ubicaciones donde el género se encuentra, más bien consiste de las ubicaciones donde se ha informado que se encuentra la especie.

3.4. Datos biológicos y los datos mediambientales

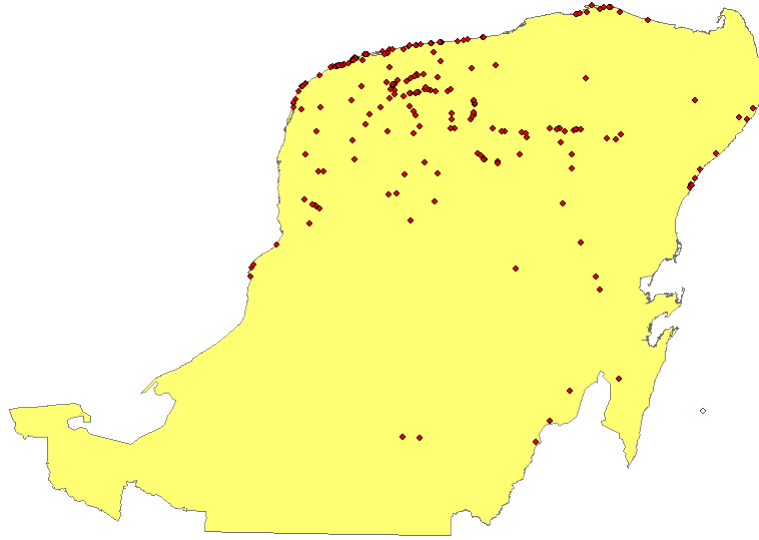


Figura 3.3: Ejemplo de registros de *solo presencias*.

Cuando se trabaja con registros de solo presencias, se dispone además de un conjunto de puntos elegidos al azar dentro del área total de estudio. En algunos modelos estos puntos son tratados como “seudo-ausencias” (por ejemplo, en modelo logístico tipo *naïve*). Otro enfoque es tratar al conjunto de puntos provenientes de toda el área de estudio (o una muestra aleatoria de estos) como *background*. El *background* está compuesto por tanto, de una muestra de celdas donde se conocen las condiciones medioambientales y éstas se comparan con aquellas celdas donde existen registros de presencias para determinar si las condiciones medioambientales son utilizadas de manera desproporcionada a su disponibilidad (Merow y Silander, 2014).

El sesgo en la colección de los datos

Un aspecto de gran importancia a tomar en cuenta es el sesgo en los datos, que a su vez se traducen en errores al momento de modelar la distribución de una especie en particular. Los datos de *solo presencias* son los más susceptibles de presentar sesgo debido a que la mayoría de estos datos son colectados en zonas de fácil acceso, por ejemplo, a orillas de caminos, carreteras, ríos, ciudades y estaciones de monitoreo biológico. En algunas ocasiones el sesgo geográfico puede conducir al sesgo medioambiental, donde las muestras medioambientales no son representativas, aunque no siempre se da el caso (Pearson, 2007). Algunos autores como Phillips *et al.* (2009) proponen que en el caso de modelos que trabajan con datos de *solo presencias*, una forma de abordar

3.4. Datos biológicos y los datos mediambientales

el sesgo en los datos es elegir muestras del *background* con el mismo tipo de sesgo. Cuando se conoce la distribución del muestreo, una forma sencilla es elegir puntos del *background* con el mismo patrón de los registros de presencias, sin embargo, rara vez se conoce dicha distribución. Los mismos autores señalan que una forma práctica de abordar el sesgo en los registros de *solo presencias*, es utilizar como *background* la información asociada a los datos de otras especies con registros de *solo presencias*, pues seguramente compartirán un sesgo similar si es que fueron colectados de la misma forma.

3.4.2. Los datos medioambientales

Un amplio rango de variables medioambientales se han empleado como variables predictivas en MDEs. Las más comunes son las variables relacionadas con el clima (por ejemplo, temperatura, precipitación, radiación solar), topográficas (por ejemplo la elevación), el tipo de suelo y la cobertura del mismo. Se debe ser cuidadoso al momento de seleccionar que variables medioambientales están relacionadas con la especie de flora o fauna en estudio ya que una mala elección nos conducirá a predicciones erróneas del modelo. Por ejemplo, algunas especies no responden directamente a la elevación, en lugar de ello si lo hacen a los cambios en temperatura y presión atmosférica que a su vez son afectadas por la elevación.

Las variables medioambientales pueden abarcar tanto datos continuos como categóricos, éstos últimos pueden expresarse como variables *dummy*. Existe amplia información acerca de las variables medioambientales en la internet, por ejemplo, en el sitio <http://www.worldclim.org/bioclim> se encuentra una base de datos global de datos climáticos codificadas como BIO1-BIO19. Sin embargo, se debe ser muy cuidadoso al emplear la información proveniente de estos sitios dado que no siempre se conoce el mecanismo utilizado para interpolar valores de variables climáticas a través de toda una área de interés. Es conveniente que el investigador comprenda la teoría detrás de cada método de interpolación utilizado y el tipo de correlación espacial y tendencia que puedan presentar los datos utilizados. A menudo se utiliza el método del inverso de la distancia para interpolación, sin embargo, este método no toma en cuenta la variabilidad espacial de los datos, y las predicciones de un modelo que se alimenta con datos erróneos conduce a predicciones fuera de la realidad en la mayoría de las ocasiones.

3.4. Datos biológicos y los datos mediambientales

En la Tabla (3.1) se encuentra información concerniente a distintos tipos de datos biológicos y medioambientales, así como su fuente de información.

Tabla 3.1: Algunos ejemplos de fuentes de información biológica y medioambiental utilizados en MDEs.

Tipo de datos	Fuente
<i>Distribución de especies</i>	
- Datos de un amplio número de organismo en distintas regiones del mundo.	Global Biodiversity Information Facility: www.gbif.org
- Datos correspondientes a una gama de organismos, en su mayoría raras o en peligro de extinción.	Nature serve: www.natureserve.org
- Datos de especies de flora y fauna para México.	Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO): http://www.conabio.gob.mx/
<i>Clima</i>	
- Datos climáticos a nivel mundial a 1km de resolución.	Worldclim: http://www.worldclim.org/
- Escenarios de clima futuros.	Intergovernmental Panel on Climate Change (IPCC): http://www.ipcc-data.org/
- Datos de reconstrucción de paleoclimas.	NOAA: http://www.ncdc.noaa.gov/data-access/paleoclimatology-data
- Normales climáticas para México.	Comisión Nacional del Agua: smn.conagua.gob.mx
<i>Topografía</i>	
- Elevación y variables relacionadas para todo el globo a 1 km de resolución.	USGS: http://www.usgs.gov/
- Elevación para la república mexicana.	INEGI: Continuo de elevaciones mexicano http://www.inegi.org.mx
<i>Imágenes satelitales</i>	
- Conjuntos de datos de cobertura del suelo.	Global Landcover Facility: http://www.landcover.org/
- Diversos productos atmosféricos y terrestres. MODIS	NASA: http://modis.gsfc.nasa.gov/data/
<i>Suelos</i>	
- Tipos de suelo a nivel global.	UNEP: http://www.grid.unep.ch/data/data.php?category=lithosphere

Capítulo 4

Métodos de Aprendizaje Automático

En este capítulo se discutirá acerca de los métodos de aprendizaje automático o supervisado que han sido aplicados para modelar distribuciones de especies. Entre estos métodos se encuentran los Árboles de Decisión (CART), las Redes Neuronales Artificiales (ANNs), los Algoritmos Genéticos (GAs) y las Máquinas de Soporte Vectorial (SVM).

4.1. Árboles de Regresión y Clasificación(CART)

Los árboles de decisión o árboles de clasificación conforman un método muy útil de clasificación supervisada cuando la variable respuesta es una variable categórica con muchos (más de dos) niveles y cuando los predictores incluyen tanto variables discretas como continuas ([Franklin, 2009](#)). El objetivo en el modelado de los árboles de decisión es el de particionar los datos en subgrupos que sean homogéneos, es decir, donde las variables respuestas tengan valores similares o sean miembros de la misma clase, basándose en los rangos de los valores de las variables predictoras. Lo anterior se lleva a cabo en tres etapas: la construcción del árbol o crecimiento, el parado del árbol y la poda del árbol o selección óptima del árbol ([Olden *et al.*, 2008](#)). La creación de particiones es análoga a la selección de variables en la regresión.

4.1. Árboles de Regresión y Clasificación(CART)

Para comenzar, una simple partición o división se hace con una variable explicatoria. La variable y la ubicación de la división son elegidas minimizando la impureza en el nodo. Hay muchas formas de minimizar la impureza de cada nodo y son conocidas como reglas de división o de particionado. Cada una de las dos regiones resultantes son nuevamente divididas bajo el mismo criterio y el árbol continua creciendo hasta que ya no sea posible crear nuevas particiones o que el proceso se detenga mediante alguna regla de parada definida por el usuario. El árbol puede reducirse de tamaño mediante un proceso que se conoce como *pruning* o podado (Moisen, 2008).

Cuando se trabaja con una variable continua como respuesta, la reducción en la varianza o devianza (suma de cuadrados) es una medida de la homogeneidad y el árbol resultante se llama árbol de regresión. En el caso de variables de respuesta categóricas, por lo general alguna medida de la homogeneidad o “pureza” de la pertenencia a una clase en los subconjuntos resultantes es utilizada y el árbol resultante se llama árbol de clasificación (Franklin, 2009).

4.1.1. Un ejemplo sencillo

Como ejemplo, considerese el problema de modelar la presencia o ausencia de la especie de árbol *Pseudotsuga menziesii*(abeto Douglas) en las montañas del norte de Utah usando como variables predictoras la elevación (ELEV) y el aspecto (ASP) tomado de (Moisen, 2008). Los datos se ordenan como se ilustra en la Tabla (4.1) y constan de 1544 registros.

Tabla 4.1: Datos del ejemplo árboles de decisión

<i>Estatus</i>	<i>Elevación(m)</i>	<i>Aspecto</i>
Ausente	2045	E
Presente	2885	SE
Presente	2374	NE
Ausente	2975	S
...

En la Figura (4.1) se muestra el árbol de clasificación para nuestro presente ejemplo. Note que al principio existen 1544 observaciones en la raíz, los 393 casos que caen por debajo de los 2202 m no son clasificados como abeto Douglas. Si la elevación es mayor que 2202 m, entonces se necesita mayor información. La siguiente división ocurre a una elevación de 2954 m. Elevaciones mayores del punto de corte también

4.1. Árboles de Regresión y Clasificación(CART)

se clasifican como lugares donde el abeto Douglas no está presente. Las restantes 928 observaciones de elevación media necesitan una tercera división, la cual ocurre a los 2444 m. Las 622 observaciones de elevación moderadamente alta (> 2444 m) se clasifican como lugares de presencia del abeto Douglas. La última división utiliza el aspecto del lugar para determinar si el abeto es probable que crezca en los restantes 306 sitios con elevación moderadamente baja, prediciendo la presencia del abeto en las zonas frías, zonas húmedas del norte y las laderas del este; y prediciendo ausencias en las zonas más cálidas.

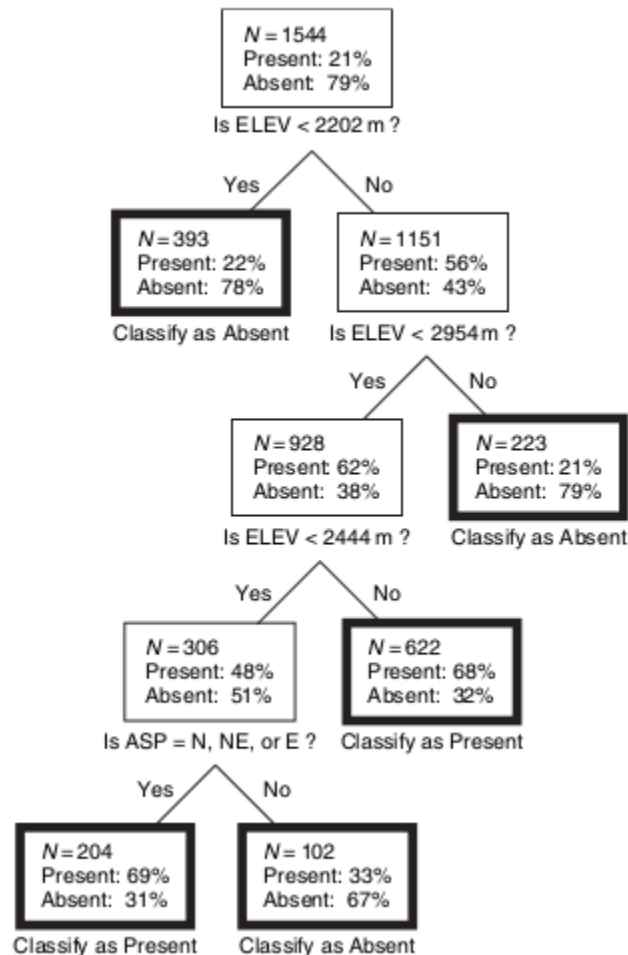


Figura 4.1: Ejemplo de árboles de clasificación describiendo la relación entre la presencia/ausencia de *P. menziesii* y las variables explicativas.

4.1. Árboles de Regresión y Clasificación(CART)

4.1.2. Árboles de regresión

En este apartado abordaremos la interrogante de cómo hacer crecer un árbol de regresión. Los datos consisten en p entradas y una respuesta, para cada una de las N observaciones, es decir, (x_i, y_i) para $i = 1, 2, \dots, N$, con $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. El algoritmo debe decidir automáticamente sobre las variables y los puntos de división, y de igual manera también que topología (forma) que el árbol debe tener. Supóngase primero que se tiene una partición en M regiones R_1, \dots, R_M , y que se modela la respuesta como una constante c_m en cada región:

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m) \quad (4.1)$$

Si se adopta como criterio de minimización a la suma de cuadrados $\sum (y_i - f(x_i))^2$, es fácil ver que el mejor \hat{c}_m es el promedio de y_i en la región R_m :

$$\hat{c}_m = \text{prom}(y_i \mid x_i \in R_m) \quad (4.2)$$

Para encontrar la mejor partición binaria en términos del mínimo de la suma de cuadrados computacionalmente hablando, resulta imposible. Por lo tanto, se procede mediante un algoritmo, comenzando con todos los datos, se considera una variable de división j y punto de corte s , y se define el par de semiplanos

$$R_1(j, s) = \{X \mid X_j \leq s\} \quad \text{y} \quad R_2(j, s) = \{X \mid X_j > s\}. \quad (4.3)$$

Luego se busca la variable de división j y el punto de división s que resuelve

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]. \quad (4.4)$$

Para cualesquiera j y s , la minimización interior se resuelve por

4.1. Árboles de Regresión y Clasificación(CART)

$$\hat{c}_1 = \text{prom}(y_i \mid x_i \in R_1(j, s)) \quad \text{y} \quad \hat{c}_2 = \text{prom}(y_i \mid x_i \in R_2(j, s)). \quad (4.5)$$

Para cada variable de división, la determinación del punto de división s puede hacerse muy rápidamente y por lo tanto, mediante la exploración de todas las entradas, la determinación del mejor par (j, s) es factible. Una vez encontrada la mejor división, la partición de los datos conlleva a dos regiones resultantes y el proceso de división se repite en cada una de las dos regiones. El proceso descrito se repetirá en todas las regiones resultantes (subregiones)([Hastie et al., 2009](#)).

En este punto nace una interrogante, ¿Qué tan grande debe crecer el árbol? Claramente un árbol muy grande podría sobreajustarse a los datos, mientras que un árbol pequeño podría no capturar la estructura importante.

El tamaño del árbol es un parámetro de ajuste que rige la complejidad del modelo y el tamaño del árbol debe ser elegido de forma adaptativa a partir de los datos. Un enfoque consistiría en dividir los nodos del árbol únicamente si la disminución en la suma de cuadrados debido a la división excede cierto umbral. La estrategia principal es hacer crecer un árbol grande T_0 , el proceso de división se detiene solo si se alcanza un cierto tamaño mínimo de nodos (por ejemplo 5). Entonces el árbol se poda tomando en cuenta el costo y la complejidad del modelo, tal como se describe enseguida.

Se define un subárbol $T \subset T_0$ que sea cualquier árbol que se pueda obtener por poda de T_0 , es decir, colapsando cualquier número de sus nodos internos. Denotaremos los nodos terminales por m , donde el nodo m representa la región R_m . Sea $|T|$ quien representa el número de nodos terminales en T . Haciendo

$$\begin{aligned} N_m &= \#(x_i \in R_m), \\ \hat{c}_m &= \frac{1}{N_m} \sum_{x_i \in R_m} y_i, \\ Q_m(T) &= \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2, \end{aligned} \quad (4.6)$$

4.1. Árboles de Regresión y Clasificación(CART)

se define el criterio costo-complejidad

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|. \quad (4.7)$$

La idea es encontrar, para cada α , el subárbol $T_\alpha \subseteq T_0$ que minimize $C_\alpha(T)$. El parámetro de ajuste $\alpha \geq 0$ regula el equilibrio entre el tamaño del árbol y la bondad de ajuste a los datos. Valores grandes de α conllevan a pequeños árboles T_α e inversamente, para valores pequeños valores de α .

Para cada α se puede demostrar que existe un único subárbol más pequeño T_α que minimiza a $C_\alpha(T)$. Para encontrar T_α se usa la poda por el enlace más débil: sucesivamente se colapsa el nodo interno que produzca el más pequeño incremento por nodo en $\sum_m N_m Q_m(T)$ y se continua así hasta que se produzca un árbol de un solo nodo (raíz). Lo anterior conlleva a una secuencia finita de subárboles, y se puede probar que la secuencia debe contener a T_α (Hastie *et al.*, 2009). La estimación de α se consigue entre cinco y diez validaciones cruzadas: se elige el valor $\hat{\alpha}$ que minimize la suma de cuadrados de la validación cruzada. El árbol final es $T_{\hat{\alpha}}$.

4.1.3. Árboles de clasificación

Si el objetivo de la clasificación es un resultado que tome valores $1, 2, \dots, K$, el único cambio necesario en el algoritmo corresponde al criterio de división de nodos y podado del árbol. Para regresión se usa el error cuadrático en el nodo como medida de impureza $Q_m(T)$ definida en la ecuación (4.6), pero este no es adecuado para clasificación. En un nodo m , representando una región R_m con N_m observaciones, sea

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

la proporción de k observaciones de clase en el nodo m . Se clasifican las observaciones en el nodo m a la clase $k(m) = \arg \max_k \hat{p}_{mk}$, la clase mayoritaria en el nodo m . Existen diferentes medidas $Q_m(T)$ de la impureza en los nodos, entre las cuales se incluyen las siguientes:

4.1. Árboles de Regresión y Clasificación(CART)

$$\begin{aligned}
 \text{Error de clasificación: } & \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}. \\
 \text{Índice de Gini: } & \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}). \\
 \text{Entropía cruzada o devianza: } & - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}. \tag{4.8}
 \end{aligned}$$

Para dos clases, si p es la proporción en la segunda clase, éstas tres medidas son $1 - \max(p, 1 - p)$, $2p(1 - p)$ y $-p \log p - (1 - p) \log (1 - p)$ respectivamente. Las tres medidas se muestran en la Figura (4.2). Las tres son muy similares, sin embargo, la entropía cruzada y el índice Gini son diferenciables, y por lo tanto más susceptible de optimización numérica. Comparando las ecuaciones (4.4) y (4.6), se observa que se necesita ponderar la medida de la impureza del nodo por el número N_{m_L} y N_{m_R} de las observaciones en los dos nodos hijos creados al dividir el nodo m .

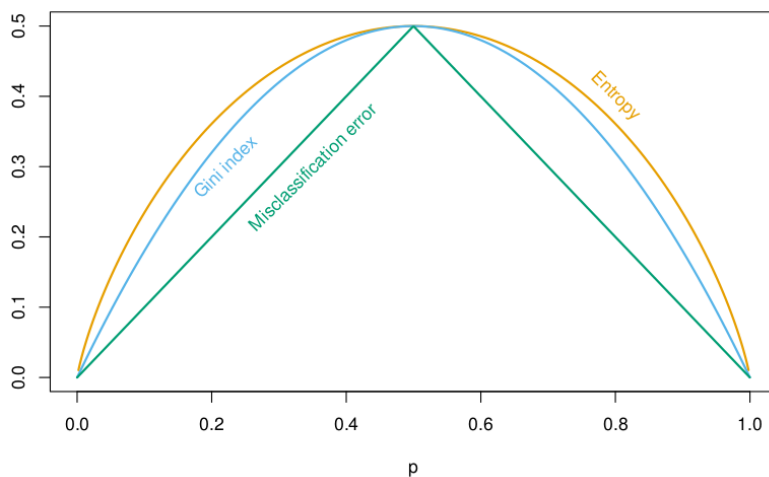


Figura 4.2: Medida de la impureza de los nodos para la clasificación de dos clases como una función de la proporción p en la clase 2.

Además, la entropía cruzada y el índice Gini son más sensibles a los cambios en las probabilidades de los nodos que la razón de clasificación errónea. Por ejemplo, en un problema de dos clases con 400 observaciones en cada clase (denotada por (400,400)), suponga que una división crea los siguientes nodos (300,100) y (100,300), mientras que otra crea (200,400) y (200,0). Ambas divisiones producen un índice

4.1. Árboles de Regresión y Clasificación(CART)

de error de clasificación de 0.25, pero la segunda división produce un nodo puro y por tanto es preferible. Tanto el índice Gini como el índice de entropía cruzada son menores en la segunda división. Por ésta razón, tanto el índice Gini como el de entropía cruzada se deben de utilizar cuando el árbol está en crecimiento. Cualquiera de las tres medidas de impureza en los nodos deben utilizarse al momento de podar un árbol, sin embargo, el más utilizado es el índice de error de clasificación.

4.1.4. Utilidad de los árboles de decisión

Los árboles de decisión es una herramienta poderosa que se ha aplicado en diversos campos de la ciencia, entre ellos, la biología, caso concreto en MDEs. Su utilidad radica en las ventajas para tratar cierta clase de datos y problemas:

1. *Predictores categóricos*: Los predictores categóricos, tales como el tipo de suelo, tipo de vegetación, etc., pueden ser difíciles de parametrizar e interpretar en un modelo lineal si existen muchas categorías; en el contexto de los árboles de decisión los predictores de este tipo son fácilmente asociados con respuestas de tipo categórica tales como presencia o ausencia de la especie.
2. *Interacciones jerárquicas*: Los árboles de decisión modelan efectos no aditivos y no lineales entre predictores y respuestas de una manera muy fácil; al igual para las respuestas de tipo jerárquicas (cuando la respuesta de un predictor está condicionado por otro).
3. *Respuestas a umbrales*: Los árboles de decisión caracterizan los efectos umbral de variables predictivas sobre la respuesta, ya que el método de particionamiento recursivo es muy efectivo al caracterizar el efecto umbral de una variables medioambiental sobre la respuesta de la especie.
4. *Salidas informativas*: Los árboles son efectivos al explorar, gráfica y analíticamente, complejas relaciones en datos multivariados y son útiles para identificar patrones en los datos (si la elevación es superior a X y la pendiente está entre m_1 y m_2 y el tipo de suelo es A , entonces la especie está presente).

4.1. Árboles de Regresión y Clasificación(CART)

4.1.5. Conjunto de métodos aplicados a los árboles de decisión: bagging, boosting y random forest

Nuevos métodos que son computacionalmente intensivos han sido desarrollados para abordar las deficiencias existentes en los métodos de clasificación y en los árboles de decisión. Estos métodos involucran la estimación de un número grande de árboles basados en subconjuntos de los datos y posteriormente se promedian los resultados obtenidos, por ello se consideran como un tipo de modelo “promedio” (Franklin, 2009).

Una de estas técnicas es la denominada como *bagging* la cual trabaja muestreando los datos con reemplazo repetidamente (*bootstrapping*) y desarrollando un árbol para cada subconjunto de los datos usando alguna regla de parada pero sin podar el árbol. Normalmente 1/3 de los datos quedan fuera de cada muestra y son utilizados para evaluar el modelo. Posteriormente las predicciones resultantes de todos los árboles se promedian. En otras palabras, cuando se quiere realizar una predicción para un nuevo conjunto de datos, cada uno de los árboles es utilizado y al final se promedian las predicciones (Franklin, 2009).

Otra variación llamada *boosting* (Ridgeway, 1999) es similar a la técnica *bagging*, a excepción de que cada observación, en lugar de tener la misma probabilidad de ser seleccionada en muestras subsecuentes, es ponderada para que tenga una mayor probabilidad de ser seleccionada si se trata de un “problema” de observación (que tienden a ser mal clasificados por los modelos anteriores).

Una variante de la técnica *boosting* llamada “*stochastic gradient boosting*” (SGB) (Friedman, 2002) construye muchos árboles pequeños de manera secuencial producto de los residuales de los árboles anteriores. La técnica *boosting*, en particular SGB, ha sido utilizada en MDEs y en predicción espacial de otras variables ecológicas y ha mostrado un mejor desempeño en comparación con los árboles de decisión ordinarios (Franklin, 2009).

Otra técnica es conocida como “*random forest*” (RF) y es una especie de *bagging* que construye un gran número de árboles no correlacionados (Hastie *et al.*, 2009) y luego las promedia. De manera similar que en el *bagging*, muchos árboles se desarrollan con subconjuntos de los datos, pero además, cada división en cada árbol se desarrolla con

4.2. Redes Neuronales Artificiales (ANNs)

un subconjunto aleatorio de variables predictoras. Un número grande (500-2000) de árboles crecen hasta alcanzar su tamaño máximo (sin poda) y después los predictores resultantes se promedia. La técnica *random forest* ha comenzado a utilizarse en MDEs y al igual que la técnica SGB, se ha demostrado que tiene una mejor precisión en la predicción comparado con los árboles de decisión ordinarios (Franklin, 2009).

4.2. Redes Neuronales Artificiales (ANNs)

Las redes neuronales artificiales (ANNs), llamadas así ya que fueron desarrolladas como modelos para el cerebro humano, representan otro enfoque de aprendizaje automático que ha sido ampliamente utilizado para clasificación de imágenes en teledetección (Benediktsson *et al.* (1993); Civco (1993)) y en otras aplicaciones relacionadas.

El principio básico consiste en derivar características (variables respuesta) que son combinaciones lineales de los predictores (entradas) y después modelar las salidas (respuestas) como funciones no lineales de esas características (Franklin, 2009). Las redes neuronales comprenden un grupo de modelos bastante amplio, aunque como ejemplo podemos describirla en términos de una red de una sola capa oculta (o perceptrón de una sola capa), y que corresponde al tipo de red neuronal utilizado frecuentemente en ecología (Olden *et al.*, 2008). Esta es una clasificación de dos etapas y para clasificar K clases, existen K unidades en la capa de entrada. Para una sola clase (ocurrencia de la especie) o una variable respuesta continua, existe comúnmente una unidad en la capa de salida. La red neuronal tiene una capa oculta que contiene características derivadas, que son combinaciones lineales de las variables predictoras (entradas) escaladas por una “función de activación” que no es lineal (frecuentemente logística o sigmoidea). La variable respuesta, o salida en la terminología de las ANNs, es una combinación ponderada de las características derivadas en cada etapa oculta (Figura 4.3)

El sobreajuste de los datos de entrenamiento usando redes neuronales se evita limitando el número de iteraciones del procedimiento de estimación usando validación cruzada (Moisen y Frescino, 2002). Sin embargo, para Hastie *et al.* (2009) existe todo un arte para estimar estos modelos. El análisis deberá decidir los valores iniciales para los pesos, el número de unidades ocultas (y las capas), la escala de las entradas (frecuentemente se estandarizan para tener las observaciones con media cero y desviación

4.2. Redes Neuronales Artificiales (ANNs)

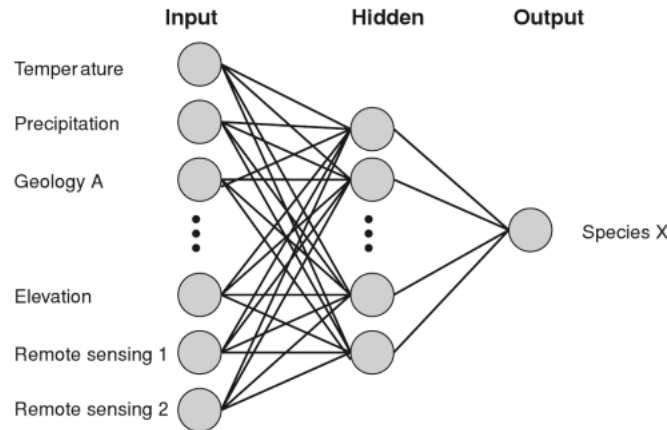


Figura 4.3: Diagrama esquemático de una red neuronal aplicada a MDEs.

estándar uno), un parámetro de decaimiento de los pesos, y así sucesivamente.

A pesar de que las redes neuronales han sido extensamente utilizadas en otras áreas de la investigación, no han sido aplicadas de manera extensiva en MDEs, sin embargo, existen algunos trabajos pioneros aunque difíciles de encontrar ([Fitzgerald y Lees, 1992](#)). Los detalles de los clasificadores desarrollados usando redes neuronales no son obvios e interpretables, aunque recientemente se han desarrollado algunas herramientas que permiten examinar las contribuciones de las variables predictoras, que corresponden a la magnitud y dirección (signo) de los pesos, que análogamente corresponden a los coeficientes en regresión ([Olden *et al.*, 2008](#)). Sin embargo, cuando las redes neuronales han sido comparadas con otros modelos tales como GLMs, GAMs y CARTs no han resultado mejores e incluso han resultado inferiores ([Thuiller *et al.* \(2006\)](#); [Benito Garzón *et al.* \(2006\)](#)). Desde una perspectiva práctica los modelos ANNs resultan más difíciles de usar que otros métodos utilizados en MDEs ([Moisen y Frescino, 2002](#)). La “curva de aprendizaje” para implementar modelos ANNs puede ser la clave de algunos resultados aparentemente contradictorios cuando las redes neuronales han sido aplicadas para modelar distribuciones de especies. En estudios donde la experiencia y habilidad por parte del investigador en el uso de ANNs es aceptable, estos modelos han obtenido alto rendimiento predictivo, en contraste, en otros estudios las redes neuronales no han realizado un mejor desempeño que los métodos estadísticos y de aprendizaje automático ([Franklin, 2009](#)).

4.3. Algoritmos genéticos (GAs)

Los algoritmos genéticos (GAs) corresponden a otro enfoque más de aprendizaje automático de clasificación supervisada que fué desarrollada en los 70's. Se llama así debido a que se genera una población de reglas de clasificación y estas reglas evolucionan mediante un proceso similar a la selección natural (mutaciones aleatorias y selecciones basadas en la aptitud) hasta que una solución óptima sea alcanzada (Franklin, 2009).

Al igual que en las redes neuronales, los algoritmos genéticos son útiles cuando hay un gran espacio de búsqueda para una solución y donde existen complejas relaciones entre variables. Las reglas se desarrollan mediante la búsqueda de las probabilidades condicionales correspondientes (Franklin, 2009), y el modelo resultante se expresa en términos de reglas de decisión condicional. Por ejemplo: “la especie *A* está presente si la precipitación anual > 20 cm., y el promedio de la temperatura de julio $< 14^{\circ}C$ ”. En ese sentido los algoritmos genéticos son semejantes a los árboles de decisión, excepto que en el caso de los árboles de decisión se derivan particionando los datos de manera recursiva.

Los algoritmos genéticos se han estado usando extensivamente en MDEs a través de un software llamado GARP (*genetic algorithm for rule-set production*), la cual arroja un resultado de tipo estocástico, por lo que el modelo se corre muchas veces y al final se promedia un subconjunto de los mejores modelos. Este software ha sido ampliamente utilizado para modelar distribuciones de especies principalmente en grandes extensiones donde solo se tienen registros de presencias. Técnicamente GARP es un “super algoritmo” para clasificación binaria que primero genera un población de reglas para clasificar a una determinada especie como “presente” o “ausente” mediante cuatro diferentes métodos: (a) *atomic rules* utiliza variables simples que representan a las variables medioambientales, por ejemplo, “si el suelo es tipo *A*, entonces la especie está presente”; (b) *Bioclim-type rules* utiliza rangos de variables bioclimáticas; (c) *range rules* es una generalización de la anterior; y (d) *logit rules* es una adaptación de la regresión logística donde los coeficientes se estiman de acuerdo al peso de las variables predictoras para predecir una probabilidad de presencia (*logit rules* predice presencia si la probabilidad es > 0.75). GARP maneja como equivalentes los conceptos de ausencias y pseudo-ausencias, por tanto, GARP maneja aquellos sitios donde no se dispone de registros de presencia como pseudo-ausencias. GARP ha sido muy utilizado en MDEs donde solo se disponen de datos de *presencias* (Franklin, 2009).

4.4. Máquinas de soporte vectorial (SVM)

Otro enfoque de aprendizaje supervisado que ha sido empleado en MDEs son las llamadas Máquinas de Soporte Vectorial (SVM). En la teoría de las SVM, dado un conjunto de puntos en el que cada uno de ellos pertenece a una de dos posibles categorías o clases (algoritmo de dos clases), un algoritmo basado en una SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría se desconoce) pertenece a una categoría o a otra. Como en la mayoría de los métodos de clasificación supervisada, los datos de entrada (los puntos) son vistos como un vector p -dimensional. La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase de otra (presencias-ausencia de la especie, en el contexto de los MDEs).

Matemáticamente, el conjunto de datos de entrenamiento en un modelo de dos clases consiste de N pares $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, con $x_i \in \mathbb{R}^P$ y $y_i \in \{1, -1\}$. El hiperplano está definido por

$$\{x : f(x) = x^T \beta + \beta_0 = 0\} \quad (4.9)$$

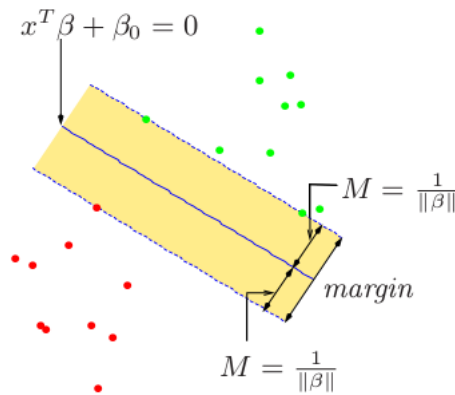


Figura 4.4: Diagrama que ilustra un hiperplano en el caso de clases separables.

donde β es un vector unitario $\|\beta\| = 1$. Una regla de clasificación inducida por $f(x)$ es

$$G(x) = \text{sign}[x^T \beta + \beta_0].$$

4.4. Máquinas de soporte vectorial (SVM)

$f(x)$ en la ecuación (4.9) proporciona la distancia desde el punto x al hiperplano $f(x) = x^T \beta + \beta_0 = 0$. Ya que las clases son separables, podemos encontrar una función $f(x) = x^T \beta + \beta_0$ con $y_i f(x) > 0 \forall i$. Por lo tanto, estamos en condiciones para encontrar el hiperplano que crea el mayor margen entre los puntos de entrenamiento para las clases 1 y -1 (ver Figura 4.4). El problema de optimización

$$\max_{\beta, \beta_0, \|\beta\|=1} M$$

sujeto a $y_i(x_i^T \beta + \beta_0) \geq M, i = 1, \dots, N$, captura este concepto.

En el caso de MDEs, las SVM han sido aplicadas a problemas donde se tiene únicamente registros de presencias, y en el caso de SVM de una sola clase se ha demostrado que tienen buen desempeño cuando se trabajan con pocos registros (alrededor de 40). Existen estudios donde se han comparado el rendimiento de las SVM con otras técnicas tales como la regresión logística, GARP y árboles de decisión que han mostrado que las SVM producen patrones espaciales muy similares a la hora de predecir (Franklin, 2009).

Capítulo 5

Modelos estadísticos

En el presente capítulo se revisan algunos aspectos teóricos de los modelos estadísticos que han sido aplicados para modelar distribuciones de especies, tal es el caso de los Modelos Lineales Generalizados (GLM), los Modelos Aditivos Generalizados (GAM) y los Splines de Regresión Multivariada Adaptativa (MARS). Estos modelos han sido ampliamente documentados en la literatura y se han utilizado para modelar distribuciones de especies utilizando registros binarios de presencia-ausencia.

Otros métodos estadísticos que se abordan son el enfoque de Máxima Entropía (Maxent) (Phillips *et al.*, 2004), el Modelo de Máxima Verosimilitud para datos de solo presencias (Maxlike) (Royle *et al.*, 2012) y el modelo de proceso poisson no homogéneo (IPP) (Warton y Shepherd, 2010). Estos métodos tienen la particularidad de que han sido aplicados para registros de *solo presencias*.

5.1. El modelo lineal

La regresión lineal es una de las técnicas estadísticas más antiguas, y que ha sido extensamente utilizado en investigaciones biológicas (Guisan *et al.*, 2002). El modelo básico de regresión lineal es de la siguiente forma:

$$Y = \alpha + X^T \beta + \varepsilon \tag{5.1}$$

5.2. Modelo lineal generalizado (GLM)

donde Y denota la variable respuesta, α es una constante llamada intercepto, $X = (X_1, \dots, X_p)$ es un vector de p variables predictoras, $\beta = (\beta_1, \dots, \beta_p)$ es el vector de p coeficientes de regresión (uno por cada predictor), y ε es el error. Al ajustar un modelo de regresión se intenta minimizar la variación no explicada a través de técnicas de estimación tal como el algoritmo de mínimos cuadrados.

Aún cuando el análisis de regresión es una herramienta poderosa cuando se aplica correctamente, sin embargo, está limitada por los siguientes supuestos :

1. los errores ε_i se asumen idénticos e independientemente distribuidos; esto incluye el supuesto de que la varianza de Y es constante a través de las observaciones;
2. para fines de pruebas de hipótesis, los errores ε_i se asumen que siguen una distribución normal;
3. la función de regresión es lineal en los predictores.

La violación del primer supuesto constituye una limitación en la aplicación de la mayoría de los modelos estadísticos paramétricos, y está relacionado directamente con el muestreo de datos. Por otra parte, muchos datos en ecología no son normales y por lo tanto no tienen una varianza constante. La violación del último supuesto ha sido abordado de manera tradicional aumentando los predictores con términos polinomiales, contemplando interacciones y otras transformaciones no lineales de los predictores originales, conduciendo a un modelo no lineal en X_j , pero lineal en los parámetros (Guisan *et al.*, 2002).

5.2. Modelo lineal generalizado (GLM)

Los datos ecológicos frecuentemente violan los supuestos del modelo lineal. Los modelos lineales generalizados (GLM) son extensiones del modelo lineal que pueden trabajar con variables respuestas que no provienen de una distribución normal (Franklin, 2009).

En el modelo GLM, las variables predictoras $\mathbf{X}_i (i = 1, \dots, p)$ son combinadas para producir un predictor lineal el cual se relaciona con el valor esperado $\boldsymbol{\mu} = \mathbf{E}(Y)$ de la variable respuesta Y a través de una función de enlace $g()$, tal como:

5.2. Modelo lineal generalizado (GLM)

$$g(\mathbf{E}(Y)) = \mathbf{X}_i^T \boldsymbol{\beta} \quad (5.2)$$

donde $\mathbf{X}_i^T \boldsymbol{\beta} = (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$. La función $g(\cdot)$ tiene que ser una función monótona y diferenciable.

Las principales mejoras del modelo GLM sobre el modelo de regresión general son por lo tanto:

1. la capacidad de manejar una clase más grande de distribuciones para la variable respuesta Y . Aparte de la distribución normal, otras distribuciones son la binomial, la Poisson y la Gamma (Guisan *et al.*, 2002).
2. la relación de la variable respuesta Y con el predictor lineal a través de la función de enlace $g(\mathbf{E}(Y))$. Además de garantizar la linealidad, ésta es una forma eficiente de condicionar las predicciones a estar dentro de un rango de valores posibles para la variable respuesta (por ejemplo, entre 0 y 1 para probabilidades de presencia) (Guisan *et al.*, 2002).

Un modelo GLM con una familia binomial (liga logit) es conocido como regresión logística y es muy utilizado cuando se modelan distribuciones de especies con respuestas binarias (presencia-ausencia).

5.2.1. El modelo de regresión logística

En algunas ocasiones se disponen de registros binarios para modelar distribuciones de especies, estos registros denominados también de presencia-ausencia son tratados como éxitos y fracasos respectivamente. La variable respuesta se asocia con un conjunto de variables predictivas y se considera que las respuestas provienen de una secuencia de ensayos independientes *Bernoulli*, donde cada ensayo tiene su propia probabilidad de éxito que depende de los valores de las variables predictoras.

Definimos la variable respuesta $y_i = 1$ cuando el i -ésimo ensayo resulta exitoso, y $y_i = 0$ cuando no lo es. Por tanto, y_i tiene la distribución *binomial*(1, p_i), donde p_i es la probabilidad de éxito en el i -ésimo ensayo.

5.2. Modelo lineal generalizado (GLM)

Dicho lo anterior, se requiere encontrar un modelo de regresión para predecir los éxitos utilizando la variable predictora, sin embargo, note que una función lineal de la variable predictora $\mathbf{X}_i^T \boldsymbol{\beta}$ estará en el intervalo $-\infty$ a ∞ mientras que p_i , la probabilidad de éxito, estará siempre entre 0 y 1. Por lo tanto no podemos igualar a la función lineal del predictor directamente con la probabilidad de éxito. En lugar de ello, es necesario encontrar una función de la probabilidad de éxito que también oscile entre $-\infty$ a ∞ que podamos ligar con la función lineal del predictor.

Supongamos que decidimos usar el logaritmo de la razón de probabilidades conocida como función de enlace *logit*. La igualamos a la función lineal del predictor

$$\log \left(\frac{p_i}{1 - p_i} \right) = \mathbf{X}_i^T \boldsymbol{\beta} \quad (5.3)$$

y resolvemos para p_i como función de \mathbf{X}_i . Note que p_i es la cantidad que interesa al investigador, es decir, la probabilidad de ocurrencia de la especie, véase (5.5).

5.2.2. Supuestos en el modelo de regresión logística

1. La *i*-ésima observación tiene la distribución *binomial*(1, p_i). Cada una con su propia probabilidad de éxito.
2. El *logit* está vinculado al predictor lineal, una función lineal desconocida de las variables predictoras.

$$\log \left(\frac{p_i}{1 - p_i} \right) = \mathbf{X}_i^T \boldsymbol{\beta} \quad (5.4)$$

Resolviendo para p_i se obtiene

$$p_i = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}}} \quad (5.5)$$

que relaciona la probabilidad de éxito con el valor de las variables predictoras.

3. Las observaciones son independientes unas de otras.

5.2. Modelo lineal generalizado (GLM)

5.2.3. Verosimilitud del modelo de regresión logística

La verosimilitud de una sola observación y_i es la probabilidad de una *binomial*(1, p_i) donde p_i es función de $p + 1$ parámetros $\beta = (\beta_0, \dots, \beta_p)$. Está dado por

$$\begin{aligned} f(y_i | \beta_0, \dots, \beta_p) &= \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) \\ &= (e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}})^{y_i} \times \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}} \right) \end{aligned}$$

donde x_{ij} es el valor de la j -ésima variable predictora para la i -ésima observación. Todas las observaciones son independientes por lo que la distribución conjunta de la muestra es el producto de las distribuciones individuales. Está dada por

$$f(y_1, \dots, y_n | \beta_0, \dots, \beta_p) = \prod_{i=1}^n f(y_i | \beta_0, \dots, \beta_p) \quad (5.6)$$

$$\begin{aligned} f(y_1, \dots, y_n | \beta_0, \dots, \beta_p) &= \prod_{i=1}^n \left(\frac{(e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}})^{y_i}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}} \right) \\ &= e^{\beta_0 \sum y_i + \sum \beta_j \sum x_{ij} y_i} \prod_{i=1}^n \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}} \right) \end{aligned}$$

5.2.4. Estimación por máxima verosimilitud en el modelo de regresión logística

Para encontrar los estimadores de máxima verosimilitud bajo el enfoque frecuentista, se debe de solucionar el sistema de ecuaciones dado por,

$$\frac{\partial \log_e f(y_1, \dots, y_n | \beta_0, \dots, \beta_p)}{\partial \beta_i} = 0$$

5.2. Modelo lineal generalizado (GLM)

para $i = 0, \dots, p$. En general, puede ser complicado encontrar la solución al sistema de ecuaciones, por ello [Nelder y Wedderburn \(1972\)](#) mostraron que en el modelo GLM, los estimadores de máxima verosimilitud pueden encontrarse iterativamente por mínimos cuadrados ponderados. Sea el vector de observaciones y el de los parámetros

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

respectivamente. El vector fila de las variables predictoras para la i -ésima observación es

$$\mathbf{x}_i = \begin{pmatrix} x_{i0} & x_{i1} & \dots & x_{ip} \end{pmatrix}$$

donde $x_0 = 1$ es el coeficiente para el intercepto, por lo tanto, la matriz de las variables predictoras para todas las observaciones resulta

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

[Pawitan \(2001\)](#) mostró que para la regresión logística, la solución de las ecuaciones de máxima verosimilitud se pueden encontrar mediante los siguientes pasos hasta que los parámetros convergan. Sea $\boldsymbol{\beta}^{(n-1)}$ el vector de parámetros en el paso $n - 1$.

1. Dado que la media y la varianza de una observación $binomial(1, p_i)$ es p_i y $p_i(1 - p_i)$, respectivamente, primero actualizamos las medias por

$$\mu_i = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}}{1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}}$$

y las varianzas por

5.2. Modelo lineal generalizado (GLM)

$$\Sigma_{ii} = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}}{1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}} \left(1 - \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}}{1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}} \right) = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}}}{\left(1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}^{(n-1)}} \right)^2}$$

2. Entonces calculamos las observaciones linealizadas por

$$\mathbf{Y}_i^n = \mathbf{X}_i \boldsymbol{\beta}^{(n-1)} + \left(\frac{y_i - \mu_i}{\Sigma_{ii}} \right)$$

3. Posteriormente se actualiza el vector de parámetros en el paso n por

$$\boldsymbol{\beta}^{(n)} = (\mathbf{X} \boldsymbol{\Sigma}^{-1} \mathbf{X}')^{-1} \mathbf{X} \boldsymbol{\Sigma}^{-1} \mathbf{Y}^{(n)}$$

El vector de estimadores de máxima verosimilitud es

$$\hat{\boldsymbol{\beta}}_{ML} = \lim_{n \rightarrow \infty} \boldsymbol{\beta}^{(n)}$$

el límite al cual el método iterativo de cuadrados medios reponderados converge. \mathbf{V}_{ML} es la “matriz de covarianzas” del vector de EMV donde su inversa es

$$\mathbf{V}_{ML}^{-1} = \mathbf{X} \boldsymbol{\Sigma}^{-1} \mathbf{X}'$$

para $\boldsymbol{\Sigma}$ calculado en $\hat{\boldsymbol{\beta}}_{ML}$.

5.2.5. Transformaciones en los predictores y selección

Respuestas de tipo no lineal pueden obtenerse incluyendo transformaciones adicionales de los predictores, tales como la inclusión de términos polinomiales, interacciones (efectos multiplicativos, no aditivos), funciones lineales por partes del tipo $(X > t)(X - t)$ donde t es un valor umbral de X , *splines* paramétricos tales como las funciones beta y predictores categóricos que son tratados como variables *dummy* (Franklin, 2009). En lo que respecta a la selección de variables puede abordarse manualmente o con un proceso automatizado de selección de variables tales como eliminación *backward*, selección *forward* y procedimientos *stepwise*.

5.3. Modelos aditivos generalizados(GAM)

5.2.6. Implementación de modelos GLM

Los modelos GLM han sido ampliamente documentados y utilizados en MDEs, tanto desde el punto de vista clásico como bayesiano. Una discusión acerca de la implementación práctica desde el punto de vista bayesiano se encuentran en [Latimer et al. \(2006\)](#). Dichos autores ilustran la construcción de modelos más elaborados partiendo del modelo GLM. Así por ejemplo, para modelar la dependencia espacial de la presencia o ausencia de determinada especie a partir de ubicaciones vecinas (o a partir de celdas vecinas), se aborda añadiendo un efecto aleatorio espacial al modelo presentado en (5.3), por lo que se tiene

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}'_i\boldsymbol{\beta} + \rho_i. \quad (5.7)$$

En la ecuación (5.7), cada celda del grid tiene un efecto aleatorio asociado ρ_i , el cual ajusta la probabilidad de presencia de la celda i dependiendo de los valores de ρ en las celdas vecinas, donde

$$\rho_i|\rho_j \approx N\left(\frac{\sum_{j \in \delta_i} a_{ij}\rho_j}{a_{i+}}, \frac{\sigma_\rho^2}{a_{i+}}\right) \quad j \neq i \quad (5.8)$$

donde a_{i+} denota el número de celdas las cuales son vecinas de i , $a_{ij} = 1$ si las celdas i y j comparten límites, 0 de otro modo. Note que σ_ρ^2 es un hiperparámetro al que se le asocia una distribución *a priori* al igual que para el vector de parámetros $\boldsymbol{\beta}$, donde ambas pueden ser no informativas. Este modelo puede ser fácilmente implementado en OpenBugs.

5.3. Modelos aditivos generalizados(GAM)

Los modelos aditivos generalizados (GAM) son extensiones semiparamétricas del modelo GLM ([Guisan et al., 2002](#)). Los modelos GAM conforman un enfoque flexible y automatizado para la identificación y descripción de relaciones no lineales entre las variables predictoras y variables respuestas ([Yee y Mitchel, 1991](#)).

5.3. Modelos aditivos generalizados(GAM)

En los modelos GAM, la relación entre la variable respuesta y las explicativas no tienen una estructura paramétrica, sino que se ajustan de forma local mediante funciones gráficas. Es decir, no existe una ecuación que represente la relación entre variables de manera constante, sino que tal ecuación varía según el entorno de los valores de respuesta de interés (Hastie y Tibshirani, 1990).

Un modelo aditivo generalizado tiene la siguiente forma

$$E(Y | X_1, X_2, \dots, X_p) = \hat{\beta}_0 + f_1(X_1) + f_2(X_2) + \dots + f_p(X_p) \quad (5.9)$$

Como de costumbre X_1, X_2, \dots, X_p representan los predictores; las f_j 's son funciones de suavizado no especificadas (no paramétricas). Se ajusta cada función usando un gráfico de dispersión suavizado (*scatterplot smoother*), por ejemplo, un *spline* de suavización cúbica o un núcleo suave (Hastie et al., 2009).

En general, la media condicional $\mu(X)$ de la respuesta Y se relaciona a una función aditiva de los predictores vía la función de enlace g :

$$g[\mu(X)] = \hat{\beta}_0 + f_1(X_1) + f_2(X_2) + \dots + f_p(X_p) \quad (5.10)$$

A diferencia del modelo de regresión logístico, en el *modelo de regresión logístico aditivo*, se reemplaza cada término lineal por una forma funcional más general

$$\log \left(\frac{\mu(X)}{1 - \mu(X)} \right) = \hat{\beta}_0 + f_1(X_1) + f_2(X_2) + \dots + f_p(X_p) \quad (5.11)$$

donde de nuevo, cada f_j es una función de suavizado no especificada. El modelo de regresión logístico aditivo es un ejemplo de modelo aditivo generalizado, frecuentemente utilizado para modelar probabilidades binomiales.

Las funciones f_j se estiman de manera flexible, usando un algoritmo que cuyo componente básico es un *scatterplot smoother*. Las funciones estimadas \hat{f}_j pueden revelar relaciones no lineales en el efecto de X_j . No todas las funciones f_j tienen que ser no lineales. Se puede mezclar fácilmente relaciones lineales y otras formas paramétricas

5.3. Modelos aditivos generalizados(GAM)

con términos no lineales, algo muy frecuente cuando algunas de las variables predictoras son cualitativas. Los términos no lineales no se restringen a los efectos principales; pueden tenerse componentes no lineales en dos o más variables o curvas separadas en X_j para cada nivel del factor X_k (Hastie *et al.*, 2009).

Algunos ejemplos de funciones de enlace son los siguientes:

- $g(\mu) = X^T\beta + \alpha_k + f(Z)$ - un modelo semiparamétrico, donde X es el vector de predictores a modelar linealmente, α_k es el efecto para el k -ésimo nivel de una entrada cualitativa V , y el efecto del predictor Z es modelado de forma no paramétrica.
- $g(\mu) = f(X) + g_k(Z)$ - nuevamente k indica los niveles de la entrada cualitativa V , y por lo tanto crea un término de interacción $g(V, Z) = g_k(Z)$ para los efectos V y Z .
- $g(\mu) = f(X) + g(Z, W)$ donde g es una función no paramétrica en dos características.

5.3.1. Ajuste de Modelos Aditivos

En esta sección describiremos un algoritmo para ajustar modelos aditivos y sus generalizaciones. El componente básico es el *scatterplot* más suave para ajustar efectos no lineales de una manera flexible. Concretamente se utiliza como *scatterplot* suavizador al suavizador *spline* cúbico .

El modelo aditivo tiene la siguiente forma

$$Y = \beta + \sum_{j=1}^p f_j(X_j) + \varepsilon \quad (5.12)$$

donde el término del error ε tiene media cero. Dadas las observaciones x_i, y_i , un criterio como la suma de cuadrados penalizada se puede especificar para este problema,

5.3. Modelos aditivos generalizados(GAM)

$$PRSS(\alpha, f_1, f_2, \dots, f_p) = \sum_{i=1}^N \left(y_i - \beta - \sum_{j=1}^p f_j(x_{ij}) \right)^2 + \sum_{j=1}^p \lambda_j \int f_j''(t_j)^2 dt_j \quad (5.13)$$

donde los parámetros de tuning $\lambda_j \geq 0$.

Afortunadamente existe un procedimiento iterativo muy simple para encontrar la solución. Se fija $\hat{\beta} = \frac{1}{N} \sum_{i=1}^N y_i$ y esto nunca cambia. Aplicamos un suavizador *spline* cúbico S_j a $\left\{ (y_i - \hat{\beta} - \sum_{k \neq j} \hat{f}_k(x_{ik}))_1^N \right\}$, como una función de las x_{ij} , para obtener una nueva estimación de \hat{f}_j . Lo anterior se realiza para cada predictor en turno, usando las estimaciones actuales de las otras funciones \hat{f}_k al calcular $y_i - \hat{\beta} - \sum_{k \neq j} \hat{f}_k(x_{ik})$. El proceso continua hasta que las \hat{f}_j se estabilizan. Este procedimiento se detalla en el Algoritmo (1) y es conocido como Algoritmo *backfitting*.

Algorithm 1 Algoritmo Backfitting para Modelos Aditivos (Hastie *et al.*, 2009)

▷ Inicio: $\hat{\beta} = \frac{1}{N} \sum_{i=1}^N y_i, \hat{f}_j \equiv 0, \forall i, j$

for $j = 1, 2, \dots, p, \dots, 1, 2, \dots, p, \dots$, **do**
 $\hat{f}_j \leftarrow S_j \left[\left\{ y_i - \hat{\beta} - \sum_{k \neq j} \hat{f}_k(x_{ik}) \right\}_1^N \right]$,
 $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^N \hat{f}_j(x_{ij})$
end for

▷ hasta que las funciones \hat{f}_j cambien menos que una cantidad fijada.

Este mismo procedimiento puede adaptarse a otros métodos de ajuste, especificando apropiadamente el operador de suavizamiento S_j .

En el caso del modelo de regresión logístico y de otros modelos aditivos generalizados, el criterio apropiado es la log-versosimilitud penalizada. Para maximizarla, el procedimiento *backfitting* se usa de manera conjunta con un maximizador de la versosimilitud. El método de Newton-Raphson para maximizar la log-verosimilitud en los modelos lineales generalizados puede modificarse como un algoritmo IRLS (método iterativo de mínimos cuadrados reponderados) (Hastie *et al.*, 2009).

5.3. Modelos aditivos generalizados(GAM)

5.3.2. Uso de modelos GAM en MDEs

Los modelos GAM, según Franklin (2009), se propusieron originalmente como un método gráfico de gran alcance para la detección y descripción de funciones de respuesta no lineales (por ejemplo, gaussianos, piecewise, curvilínea, etc.) y a partir de ello construir modelos paramétricos, por ejemplo un modelo GLM.

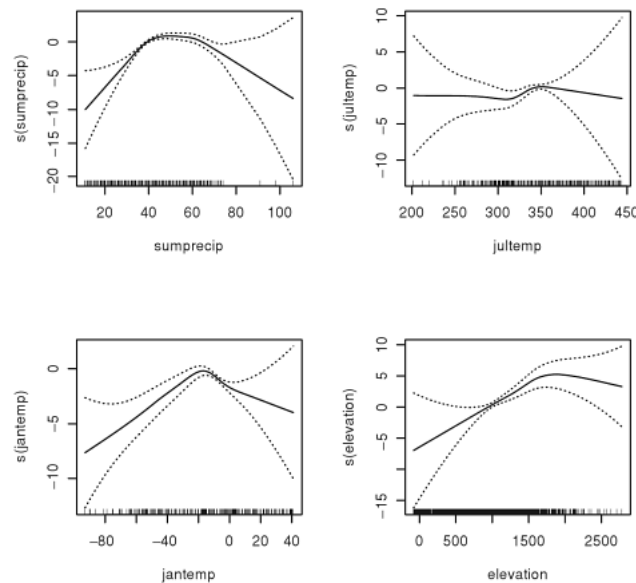


Figura 5.1: Forma de la función de respuesta (determinada usando un spline cúbico como *scatterplot suavizador*) entre la razón de log-verosimilitudes de presencia de la especie (eje y , etiquetada como “ $s(X)$ ”) y las variables predictoras (eje x) estimados mediante un modelo GAM usando la distribución binomial (*logit link*) para datos de presencia-ausencia.

Sin embargo, los modelos GAM no se pueden usar para calcular los parámetros de respuesta de las especies tales como el óptimo y la tolerancia, que otros métodos sí pueden. Otra de las limitaciones importantes es que tanto para la exploración como identificación, es que los modelos son aditivos y esto dificulta la incorporación de interacciones en el modelo. Otra desventaja importante, es precisamente la necesidad de disponer de elementos gráficos para su implementación.

Los modelos GAM se han utilizado en diversos estudios, muchos de ellos citados por Franklin (2009), entre los cuales destacan los encaminados para fines de conservación, examinando la dinámica de los hábitat temporales, variación regionales en las preferencias de hábitat de especies raras, predicción espacial de los impactos del cambio climático en las especies, para estudiar el efecto de la composición y estructura de

5.4. Splines de Regresión Multivariada Adaptativa (MARS)

bosques sobre distribuciones de aves, etc.

5.4. Splines de Regresión Multivariada Adaptativa (MARS)

Tal como se describe en [Hastie *et al.* \(2009\)](#), el enfoque *MARS* puede verse como generalización de la regresión lineal por partes, muy adecuado para los problemas con un gran número de variables predictoras, también puede verse como una modificación del enfoque CART.

Según [Leathwick *et al.* \(2006b\)](#), los modelos MARS son un método de regresión no lineal, donde las llamadas funciones base están definidas en pares, a ambos lados de un nudo. Un nudo es un valor de una variable que define un punto de inflexión a lo largo del rango del predictor, por tanto, dos nudos definen una sección. Los coeficientes se estiman de tal forma que definen la pendiente de cada sección, muchos nudos potenciales se identifican automáticamente (ésta es la parte adaptativa del enfoque MARS).

Los nudos y sus correspondientes pares de funciones base, así como sus productos se seleccionan de tal forma que proporcionen el mayor decremento en la suma de cuadrados de los residuales, y se evalúan mediante validación cruzada generalizada. Como resultado obtenemos un modelo *grande* que posteriormente es podado de forma iterativa, removiendo las funciones base que menos contribuyen en el modelo. Este aspecto de los modelos MARS es similar al enfoque CART ([Franklin, 2009](#)).

Los modelos MARS tienen muchas ventajas potenciales en MDEs que han sido enfatizadas en pocos estudios. Una de las ventajas, es por ejemplo, que los MARS son computacionalmente más rápidos que los GAM, y por tanto, son modelos prácticos para trabajar con grandes bases de datos, donde se tengan muchas observaciones y un gran número de variables predictoras. Otra ventaja de los MARS versus GAM, es que en estos últimos la predicción espacial resulta difícil, mientras que los primeros son mucho más fáciles de usar ([Franklin, 2009](#)).

Un desafío práctico de usar MARS es que comúnmente se utiliza software basados en ajuste por mínimos cuadrados (por ejemplo, la librería *mda* en R), que son apropiados

5.4. Splines de Regresión Multivariada Adaptativa (MARS)

para datos donde los errores tienen distribución normal, sin embargo, éste enfoque no es apropiado para datos binomiales (es decir: presencia/ausencia). Otra dificultad técnica, es que la mayoría del software disponible solo permiten el uso de variables predictoras continuas o cuantitativas (no categóricas). Así, algunos estudios no han considerado ninguna variable categórica (Franklin, 2009).

5.4.1. Descripción matemática de los MARS

Los MARS utilizan expansiones de funciones de base lineal por partes de la forma $(x - t)_+$ y $(t - x)_+$. El signo “+” representa la parte positiva, por lo que

$$(x - t)_+ = \begin{cases} x - t, & \text{si } x > t \\ 0, & \text{de otro modo} \end{cases}$$

y

$$(t - x)_+ = \begin{cases} t - x, & \text{si } x < t \\ 0, & \text{de otro modo} \end{cases}$$

Como ejemplo, la función $(x - 0.5)_+$ y $(0.5 - x)_+$ se muestran en la Figura (5.2).

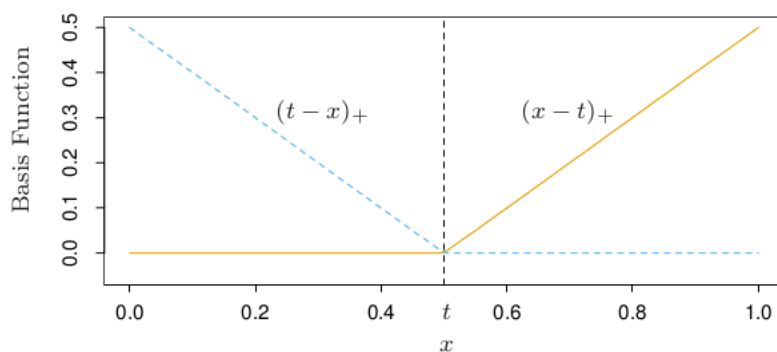


Figura 5.2: Las funciones base $(x - t)_+$ (línea en naranja) y $(t - x)_+$ (línea azul punteada) usados por MARS.

Cada función es lineal por partes con un nudo en el valor t , lo que también se conoce como splines lineales. Se les llama a las dos funciones como un par reflejado. La idea

5.4. Splines de Regresión Multivariada Adaptativa (MARS)

es formar pares reflejados para cada entrada X_j con nudos para cada valor observado x_{ij} de dicha entrada. Por lo tanto, la colección de funciones base:

$$C = \{(X_j - t)_+, (t - X_j)_+\}_{t \in \{x_{1j}, x_{2j}, \dots, x_{Nj}\}} \text{ con } j = 1, 2, \dots, p \quad (5.14)$$

La estrategia para construir el modelo se asemeja a una regresión lineal por partes, pero en lugar de utilizar las entradas originales, se permite utilizar las funciones de la serie C y sus productos. Así el modelo tiene la forma

$$f(X) = \beta_0 + \sum_{m=1}^M \beta_m h_m(X) \quad (5.15)$$

donde cada $h_m(X)$ es una función en C , o un producto de dos o más de tales funciones. Los coeficientes β_m se estiman minimizando la suma de cuadrados de los residuales, como en la regresión lineal estándar. Sin embargo, el verdadero arte radica en la construcción de las funciones h_m . Se puede comenzar con la función constante $h_0(X) = 1$ en el modelo y por tanto, todas las funciones en el conjunto C son funciones candidatas.

En cada etapa se considera como un nuevo par de función base a todos los productos de una función h_m en el modelo del conjunto M con uno de los pares reflejados en C . Añadimos al modelo M el término de la forma

$$\hat{\beta}_{M+1} h_l(X) \cdot (X_j - t)_+ + \hat{\beta}_{M+2} h_l(X) \cdot (t - X_j)_+, \quad h_l \in M$$

que produce la mayor reducción en el error de entrenamiento. Aquí $\hat{\beta}_{M+1}$ y $\hat{\beta}_{M+2}$ son los coeficientes estimados por mínimos cuadrados, junto con todos los demás $M + 1$ coeficientes en el modelo. Después los productos ganadores se añaden al modelo y el proceso continua hasta que el modelo conjunto M contenga un número de términos predefinidos.

Por ejemplo, en la primera etapa consideremos adicionar al modelo una función de la forma $\beta_1(X_j - t)_+ + \beta_2(t - X_j)_+$ $t \in \{x_{ij}\}$, ya que al multiplicar por una función constante solo produce la misma función. Suponemos que la mejor elección es $\hat{\beta}_1(X_2 -$

5.4. Splines de Regresión Multivariada Adaptativa (MARS)

$x_{72})_+ + \hat{\beta}_2(x_{72} - X_2)_+$. Este par de funciones base se adicionan al conjunto M , y en la siguiente etapa se incluyen un par de productos de la forma de

$$h_m(X) \cdot (X_j - t)_+ \quad \text{y} \quad h_m(X) \cdot (t - X_j), t \in \{x_{ij}\}$$

donde para h_m se tienen las opciones:

$$\begin{aligned} h_0(X) &= 1, \\ h_1(X) &= (X_2 - x_{72})_+, \quad \text{o} \\ h_2(X) &= (x_{72} - X_2)_+. \end{aligned}$$

La tercera opción produce funciones tales como: $(X_1 - x_{51})_+ \cdot (x_{72} - X_2)_+$, representada en la Figura (5.3)

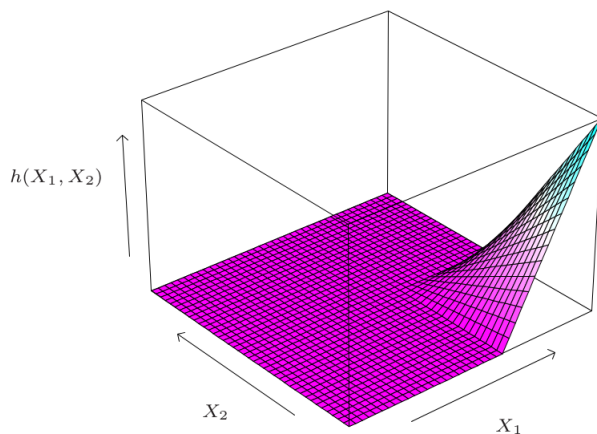


Figura 5.3: La función $h(X_1, X_2) = (X_1 - x_{51})_+ \cdot (x_{72} - X_2)_+$.

Al final de este proceso se tiene un modelo grande de la forma de la ecuación (5.15). Este modelo generalmente se sobreajusta a los datos y es necesario aplicar el procedimiento de selección de variables *backward*. El término cuya extracción cauce el menor incremento en el error cuadrático de los residuales se elimina del modelo en cada etapa, produciendo un modelo estimado \hat{f}_λ . Para estimar el valor óptimo de λ se utiliza el criterio de validación cruzada generalizada. Este criterio se define como

$$GCV(\lambda) = \frac{\sum_{i=1}^N (y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2} \quad (5.16)$$

5.5. Máxima entropía (MaxEnt)

El valor $M(\lambda)$ es el número efectivo de parámetros en el modelo: esto representa tanto el número de términos en el modelo, más el número de parámetros utilizados en la selección de las posiciones óptimas de los nudos.

5.4.2. Algunas aplicaciones de MARS en MDEs

La aplicación de los modelos MARS en MDEs es relativamente reciente, y derivado de ello son pocos los estudios donde han explorado esta herramienta estadística para modelar distribuciones de especies. Muñoz y Felicísimo (2008) mostraron que el enfoque MARS y los árboles de clasificación son mejores a la hora de predecir cuando se le compara con la regresión logística múltiple. La predicción espacial resulta mucho más simple cuando se utiliza el enfoque MARS en comparación con los árboles de clasificación.

5.5. Máxima entropía (MaxEnt)

El método de Máxima entropía (Phillips *et al.*, 2004) ha sido clasificado por algunos autores como un método de aprendizaje automático, sin embargo, algunos otros los han visto desde la perspectiva de un método estadístico. Este método ha cobrado gran importancia y aceptación por parte de investigadores en MDEs gracias a que es un método que tiene la bondad de trabajar con datos de *solo presencias*. Hoy en día es el método más popular para datos de *solo presencias* y abundan estudios de aplicación y de análisis del método en si.

El principio de máxima entropía establece que la mejor aproximación a una distribución no conocida es alguna cuya entropía sea máxima (la más dispersa) sujeta a restricciones conocidas. Estas restricciones se definen en función del valor esperado de la distribución, la cual es estimada por medio de un conjunto de observaciones de presencia de una determinada especie (Elith *et al.*, 2011).

Una de las diferencias importantes entre MaxEnt y otros modelos, por ejemplo el modelo logístico, es que en MaxEnt, las ubicaciones donde no existen registros de ocurrencia no son interpretados como pseudo-ausencias, sino más bien como la representación de una muestra de fondo o *background* en la terminología MaxEnt. El

5.5. Máxima entropía (MaxEnt)

enfoque MaxEnt únicamente requiere de datos de presencia aunada a la información medioambiental del área de estudio. Otra de las características importantes de este método es que es muy robusto cuando se tienen cantidades limitadas de datos de entrenamiento (muestras pequeñas), de aquí que MaxEnt es un método para estimar una densidad que se desconoce y por tanto no es un método de regresión (Elith *et al.*, 2011).

5.5.1. Prólogo

El modelo asume que únicamente existen datos de presencia, es decir, un conjunto de ubicaciones dentro de L , donde L es el área de estudio de interés. La variables aleatoria Y puede tomar por tanto dos valores, $y = 1$ (presencia) y $y = 0$ (ausencia), \mathbf{z} representa un vector de covariables medioambientales y se asume que están disponibles para toda L y *background* como todas las ubicaciones dentro de L (o una muestra aleatoria de los mismos).

Se define $f(\mathbf{z})$ como la densidad de probabilidad de las covariables en L , $f_1(\mathbf{z})$ representa la densidad en L donde las especies están presentes y de manera similar $f_0(\mathbf{z})$ representa la densidad en L donde las especies están ausentes. La cantidad que desea estimarse es la probabilidad de que una determinada especie condicionada por el medioambiente esté presente en un sitio determinado $Pr(y = 1 | \mathbf{z})$.

Estrictamente los datos de presencia solo permiten modelar $f_1(\mathbf{z})$ que por si sola no puede aproximar la probabilidad de presencia, sin embargo, los datos de presencia más el *background* permiten modelar $f_1(\mathbf{z})$ y $f(\mathbf{z})$, luego empleando la *regla de Bayes* se tiene que:

$$Pr(y = 1 | \mathbf{z}) = \frac{f_1(\mathbf{z})Pr(y = 1)}{f(\mathbf{z})} \quad (5.17)$$

El único término no conocido en la ecuación (5.17) es $Pr(y = 1)$ que corresponde a la *prevalencia* de la especie (proporción de sitios ocupados) en el área de interés, lo que significa que no puede ser determinada exactamente, independientemente del tamaño de la muestra, lo que es una limitación cuando se trabaja con datos de solo presencia (Elith *et al.*, 2011). Aunque como veremos más adelante, dicha prevalencia puede

5.5. Máxima entropía (MaxEnt)

calcularse a partir de datos de las covariables asociadas a las muestras de presencias y del *background* según el enfoque sugerido por Royle *et al.* (2012).

5.5.2. Covariables y sus transformaciones (features)

El enfoque de máxima entropía ha sido aterrizado en un paquete multiplataforma que corre mediante Java y que en lo sucesivo lo denotaremos como *Maxent* (en “cursivas”). *MaxEnt*, trabaja con transformaciones de las variables originales a las cuales denomina *features* y que aquí nos referiremos a ellas como *transformaciones*. En *MaxEnt* la variable respuesta y las covariables medioambientales son formadas detrás de escena de la misma forma como en el análisis de regresión, donde la matriz diseño es aumentada por términos específicos en el modelo (ejemplo: polinomios, interacciones). *MaxEnt* actualmente cuenta con seis tipos de transformaciones: lineal, producto, cuadrática, umbral por arriba, umbral por abajo, y categórica. Una descripción somera de estas transformaciones se da enseguida:

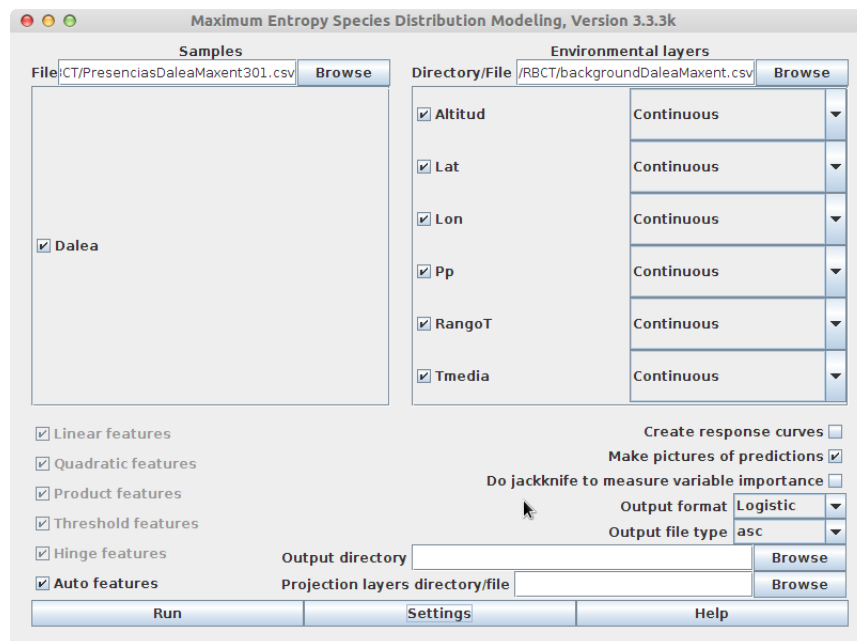


Figura 5.4: Interfaz gráfica del software *MaxEnt*.

- *Lineal*: limita la distribución resultante de tal forma que cada especie tenga la misma esperanza para cada una de las variables continuas, tal como en la

5.5. Máxima entropía (MaxEnt)

muestra. Una transformación lineal es simplemente una de las variables medioambientales continuas.

- *Cuadrática*: (cuando se utiliza junto con la transformación lineal) limita la distribución resultante de tal forma que tenga la misma esperanza y varianza de la muestra. Una transformación cuadrática es el cuadrado de una de las variables continuas.
- *Producto*: corresponde al producto de dos covariables medioambientales; cuando se usan conjuntamente con las transformaciones lineales o cuadráticas, ésta transformación limita la distribución resultante de tal forma que dos pares de variables tengan la misma covarianza, como la muestral.
- *Umbral por arriba*: ésta transformación se deriva de una variable continua. Para un umbral v , la transformación es binaria (toma valores de 0 y 1) y es 1 cuando la variable es mayor que el umbral v . El efecto de esta transformación es hacer que la probabilidad total del grid de celdas cuyos valores sean mayores que v sea igual a la fracción de ubicaciones muestrales cuyos valores son mayores que v .
- *Umbral por abajo*: es similar a la transformación lineal, solo que es constante por debajo de un umbral v .
- *Catagórica*: se selecciona automáticamente para cada variable catagórica seleccionada. Una característica se realiza para cada posible valor de cada variable catagórica: la transformación para un valor v es binaria (tomando valores 0 y 1) y es 1 cuando la variable tiene un valor v . El efecto de esta transformación es hacer que la probabilidad total del grid de celdas con un valor particular de una variable catagórica sea igual a la fracción de ubicaciones muestrales con ese valor.

5.5.3. Explicación de MaxEnt

La ecuación (5.17) muestra que si conoce la densidad condicional de las covariables en los sitios de presencia, $f_1(\mathbf{z})$, y la densidad marginal de las covariables en toda el área de estudio $f(\mathbf{z})$, entonces únicamente resta conocer la prevalencia $P_r(y = 1)$, para calcular la probabilidad condicional de ocurrencia.

5.5. Máxima entropía (MaxEnt)

El software primero calcula la razón $f_1(\mathbf{z})/f(\mathbf{z})$, denominada *raw output*, una especie de salida en bruto que proporciona una idea relativa de que tan idóneo es un sitio frente a otro. Debido a que no se dispone de la información de la prevalencia por tanto no puede calcularse la probabilidad condicional de ocurrencia, como ya se ha dicho, por ello en *MaxEnt* se ha planteado una solución alternativa denominada salida logística de *MaxEnt*, la cual realiza una transformación monótona de $f_1(\mathbf{z})/f(\mathbf{z})$ como $\eta(\mathbf{z}) = \log(f_1(\mathbf{z})/f(\mathbf{z}))$, y calibra el intercepto de tal forma que implique la probabilidad de presencia en “condiciones típicas” para las especies, que es el parámetro τ .

Descripción del modelo

MaxEnt utiliza los datos de las covariables asociadas a los registros de presencia y del *background* para estimar la razón $f_1(\mathbf{z})/f(\mathbf{z})$. Esto se hace al estimar $f_1(\mathbf{z})$ que es consistente con los datos de ocurrencia; muchas de éstas distribuciones son posibles, pero se elige aquella que esté lo más cercana a $f(\mathbf{z})$ (Elith *et al.*, 2011). La distancia existente entre $f_1(\mathbf{z})$ y $f(\mathbf{z})$ se denomina entropía relativa (también conocida como divergencia de *Kullback-Leibler*) y lo que desea es minimizar dicha distancia. La entropía relativa proporciona la distancia entre dos distribuciones de probabilidad diferentes, por tanto, la entropía relativa es siempre positiva y es cero si y solo si las dos distribuciones son la misma e incrementa en la medida que las distribuciones divergen.

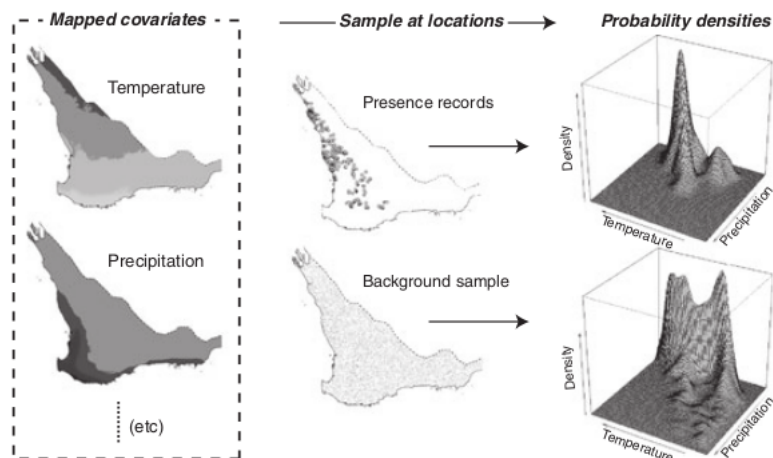


Figura 5.5: Ejemplo esquemático de las densidades de probabilidad en el modelo *MaxEnt* utilizando dos variables, los datos de presencias y el *background* (tomada de Elith *et al.* (2011)).

5.5. Máxima entropía (MaxEnt)

Usando los datos del *background* obtenemos información acerca de $f(\mathbf{z})$, la densidad de las covariables en la región, la cual sirve para comparar con la densidad de las covariables asociadas a los registros de presencias, es decir, $f_1(\mathbf{z})$ (Figura 5.5). Las restricciones se imponen de tal forma que la solución sea la que refleja la información de los registros de presencia. Por ejemplo, si una de las covariables es la lluvia de verano, entonces las restricciones impuestas aseguran que la media de la precipitación en verano para estimar $f_1(\mathbf{z})$ este cerca de su media a través de las ubicaciones con presencias observadas, por lo tanto, la distribución de la especie se estima minimizando la distancia entre $f_1(\mathbf{z})$ y $f(\mathbf{z})$ sujeta a la restricción de que la media de la precipitación en verano estimada por f_1 (y la media de las otras covariables) estén cerca de la media en los sitios de presencia.

Se introduce en este punto $h(\mathbf{z})$ como el vector de covariables mediambientales y β como el vector de coeficientes. Al minimizar la entropía relativa resulta en la distribución de Gibbs el cual es un modelo de la familia exponencial:

$$f_1(\mathbf{z}) = f(\mathbf{z})e^{\eta(\mathbf{z})} \quad (5.18)$$

donde $\eta(\mathbf{z}) = \alpha + \beta \cdot h(\mathbf{z})$ y α es una constante de normalización que nos asegura que al integrar $f_1(\mathbf{z})$ sume 1.

Note que el objetivo de un modelo MaxEnt es $e^{\eta(\mathbf{z})}$, el cual estima la razón $f_1(\mathbf{z})/f(\mathbf{z})$. Es un modelo log-lineal similar en forma a un modelo GLM, y que depende tanto de la muestra de presencias como de la muestra de fondo.

Mecánica de la solución

Para llegar a una solución, *Maxent* necesita encontrar los coeficientes β' s que darán lugar a las restricciones que deben satisfacerse, pero que no estén tan cercanos entre ellos y no produzcan un sobreajuste. *Maxent* aborda este problema estableciendo un límite para el error de la muestra (empírica) mediante funciones. *Maxent* primero reescala todas las covariables transformadas de tal forma que estén dentro del rango 0 – 1. Después, el límite para el error se calcula para cada transformación (λ_j en la ecuación 5.19)

5.5. Máxima entropía (MaxEnt)

$$\lambda_j = \lambda \sqrt{\frac{s^2[h_j]}{m}} \quad (5.19)$$

donde λ_j es el parámetro de regularización por cada transformación h_j . La varianza de esta transformación es $s^2[h_j]$ sobre los m sitios de presencia y tiene un parámetro de ajuste λ . Conceptualmente λ_j corresponde al ancho del intervalo de confianza, donde λ_j es el parámetro de regularización para cada transformación h_j . Las λ 's en la ecuación (5.19) permiten la regularización, es decir, el suavizamiento de la distribución, haciendola más regular. El objetivo de la regularización es negociar el ajuste del modelo (primer término de la ecuación 5.20) y la complejidad del modelo (segundo término de la ecuación 5.20). En ese sentido *Maxent* se ajusta a un modelo de máxima verosimilitud penalizada, estrechamente relacionada con otras penalizaciones, tales como el criterio de información de Akaike.

Al maximizar el logaritmo de la verosimilitud penalizada es equivalente a minimizar la entropía relativa sujeta a las restricciones impuestas por la cota del error:

$$\max_{\alpha, \beta} \frac{1}{m} \sum_{i=1}^m \ln (f(\mathbf{z}_i) e^{\eta(\mathbf{z}_i)}) - \sum_{j=1}^n \lambda_j |\beta_j| \quad (5.20)$$

sujeta a $\int_L f(\mathbf{z}) e^{\eta(\mathbf{z})} dz = 1$ donde \mathbf{z} es el vector de *features* o transformaciones de las variables originales para el punto de ocurrencia i de m lugares y para $j = 1, \dots, n$ características.

Salida logística de *Maxent*

De la ecuación (5.17) podemos apreciar que un enfoque para estimar $P_r(y = 1 | \mathbf{z})$ sería simplemente multiplicar a $e^{\eta(\mathbf{z})}$ por una constante que estime la prevalencia, sin embargo, este enfoque tiene la desventaja de que $e^{\eta(\mathbf{z})}$ puede ser arbitrariamente grande, lo que a su vez implica que podemos obtener una estimación de los parámetros de tal manera que $P_r(y = 1 | \mathbf{z})$ exceda 1 (Elith *et al.*, 2011).

Los modelos exponenciales tienen en general mal comportamiento cuando se aplica a nuevos datos, por ejemplo, cuando se extrapolan a nuevos paisajes. Para evitar

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

estos problemas y para esquivar la no identificabilidad de la prevalencia $P_r(y = 1)$, la salida logística de *Maxent* transforma el modelo perteneciente a la familia exponencial (ecuación 5.18) a un modelo logístico:

$$P_r(y = 1 \mid \mathbf{z}) = \frac{\tau e^{\eta(\mathbf{z})-r}}{1 - \tau + \tau e^{\eta(\mathbf{z})-r}} \quad (5.21)$$

donde $\eta(\mathbf{z})$ es el score lineal de la ecuación (5.18), r es la entropía relativa que estima *MaxEnt* de $f_1(\mathbf{z})$ con respecto a $f(\mathbf{z})$ y τ es la probabilidad de presencia en sitios condiciones “típicas” para las especies (es decir, donde $\eta(\mathbf{z}) =$ promedio de los valores de $\eta(\mathbf{z})$ bajo f_1).

El valor por default para τ es 0.5, sin embargo, *Maxent* permite cambiar este valor a criterio del investigador, y siempre que se tenga información confiable de τ . Note entonces que la salida logística de *Maxent* no es una probabilidad en el sentido de que se asume conocida la prevalencia, pero si un índice que va de 0 – 1 y que mide que tan idóneo es un sitio para albergar a determinada especie.

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

Supóngase que Y representa una variable aleatoria que denota la presencia ($y = 1$) o ausencia ($y = 0$) de la especie y X representa alguna covariable medioambiental asociada a estos registros. Según [Royle et al. \(2012\)](#), cuando se tiene un conjunto de ubicaciones para las cuales $y = 1$, surgidos al descartar previamente aquellos celdas donde no hay registro de la especie de un conjunto de datos derivados de un muestreo aleatorio, es decir, cuando se obtuvo una muestra aleatoria x_1, \dots, x_N y como registros $y(x_1), \dots, y(x_N)$ y de los cuales solo se consideran aquellos sitios x_1, \dots, x_n para las cuales $y(x) = 1$, entonces la prevalencia puede estimarse bajo el enfoque de máxima verosimilitud.

La característica principal en los datos de *solo presencias* es que la variable y ya no es aleatoria, debido a que $y = 1$ con probabilidad 1 para todas las observaciones. En lugar de ello, x es una cantidad aleatoria, y el conjunto de n ubicaciones x_1, \dots, x_n

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

son los datos sobre los cuales se basa la inferencia. Los valores de x que aparecen en la muestra representan el sesgo en la selección sobre todos los valores posibles \mathcal{X} favoreciendo aquellos donde $y = 1$ (Royle *et al.*, 2012).

En la notación de Royle *et al.* (2012), $\pi()$ representa la distribución de probabilidad de x , y $\psi()$ representa la distribución de probabilidad de y . Para encontrar la verosimilitud, se necesita identificar la distribución de probabilidad condicional $\pi(x|y = 1)$, de x para los cuales $y = 1$. Aplicando la regla de Bayes se tiene:

$$\pi(x|y = 1) = \frac{\psi(y = 1|x)\pi(x)}{\psi(y = 1)} \quad (5.22)$$

Note que la ecuación (5.22) está expresada en términos del espacio geográfico, sin embargo, puede expresarse en términos del espacio medioambiental (donde Z es la variable aleatoria que denota la covariable medioambiental), siempre que las z sean muestreadas aleatoriamente.

La distribución de probabilidad $\pi(x)$ es la que describe los resultados posibles de la variable aleatoria x (pixel identidad). Consideremos que el espacio de valores de x es discreto y que tiene M elementos únicos equiprobables, y por lo tanto, $\pi(x) = 1/M$. Por otra parte, $\psi(y = 1|x)$ es la probabilidad de que $y = 1$ condicionada por x y que se denomina como *probabilidad de ocurrencia*. Note que $\psi(y = 1)$ es la probabilidad marginal de que un pixel albergue a la especie, que por definición es

$$\psi(y = 1) = \sum_{x \in \mathcal{X}} \psi(y = 1|x)\pi(x) \quad (5.23)$$

y que es el promedio espacial de probabilidad de ocurrencia y que en la literatura se denomina *prevalencia* (Royle *et al.*, 2012).

La Verosimilitud

Note que $\psi(y_i = 1|x_i)$ corresponde a la probabilidad de ocurrencia y que depende de algunos parámetros β asociados a las covariables medioambientales, por lo que podemos escribirla como $\psi(y_i = 1|x_i; \beta)$, por lo tanto, $\pi(x_i|y_i = 1)$ es

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

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

sin embargo, $\pi(x_i)$ es constante y por tanto se cancela del numerador y del denominador resultando en

$$\pi(x_i|y_i = 1) = \frac{\psi(y_i = 1|x_i; \boldsymbol{\beta})}{\sum_{x \in \mathcal{X}} \psi(y_i = 1|x; \boldsymbol{\beta})} \quad (5.24)$$

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

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \frac{\psi(y_i = 1|x_i; \boldsymbol{\beta})}{\sum_{x \in \mathcal{X}} \psi(y_i = 1|x; \boldsymbol{\beta})} \quad (5.25)$$

Los parámetros $\boldsymbol{\beta}$ de la ecuación (5.25) pueden estimarse maximizando la verosimilitud usando métodos estándar. El denominador de la ecuación corresponde a la probabilidad marginal de ocurrencia a través del área de estudio, que se calcula sumando sobre todos los elementos de $x \in \mathcal{X}$ donde \mathcal{X} (o sobre una muestra aleatoria de \mathcal{X} , comúnmente denominada *background* y denotada como \mathfrak{B}) corresponde al espacio de valores de x .

Note que cuando y depende únicamente de x a través de \mathbf{z} (donde \mathbf{z} representa el vector de valores de las covariables medioambientales), podemos expresar la ecuación (5.22) como

$$\pi(\mathbf{z}|y_i = 1) = \frac{\psi(y_i = 1|\mathbf{z})\pi(\mathbf{z})}{\psi(y = 1)}.$$

Note que por la *Ley de la probabilidad total*, la probabilidad marginal $\psi(y = 1|\mathbf{z})$ puede calcularse si se dispone de una muestra aleatoria de \mathbf{z} independiente de y . Al introducir covariables en el modelo, podemos modelar la relación entre $\psi(y_i = 1|\mathbf{z})$ y las covariables, por ejemplo, mediante la función liga *logit*

5.7. Modelo de Proceso Poisson no Homogéneo (IPP) para datos de “solo presencias”

$$\text{logit}\{\psi(y_i = 1|\mathbf{z}; \boldsymbol{\beta})\} = \frac{\psi(y_i = 1|\mathbf{z}; \boldsymbol{\beta})}{1 - \psi(y_i = 1|\mathbf{z}; \boldsymbol{\beta})} = \mathbf{z}'\boldsymbol{\beta}$$

donde $\mathbf{z}'\boldsymbol{\beta} = \beta_0 + z_1 + \dots + \beta_j z_j$ y como se aprecia incluye el intercepto β_0 y los otros β 's son los coeficientes asociados con cada una de las j covariables (del mismo modo que en el modelo GLM).

5.7. Modelo de Proceso Poisson no Homogéneo (IPP) para datos de “solo presencias”

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

5.7.1. Probabilidad de ocurrencia vs tasa de ocurrencia

Generalmente cuando se modela una especie en determinada región, se proporciona un mapa de probabilidades. El mapa de probabilidades refleja las ubicaciones más o menos favorecidas por la especie dependiendo del color en función de la probabilidad calculada. Dependiendo de la resolución del estudio se realizan las generalizaciones, es decir, si el tamaño de celda es del 1 km^2 la probabilidad se asocia para toda la celda y se interpreta que si uno recorre la celda, se espera observar con probabilidad p al menos un ejemplar (o algún rastro que indique presencia temporal en el caso de fauna) de la especie, por lo tanto, según ([Fithian y Hastie, 2013](#)), la definición misma de “probabilidad de ocurrencia” en un estudio de presencia-ausencia depende crucialmente del esquema de muestreo específico utilizado para recoger los datos de presencia-ausencia.

En estudios donde se trabajan con datos de presencias únicamente, es más natural

5.7. Modelo de Proceso Poisson no Homogéneo (IPP) para datos de “solo presencias”

estimar una *tasa* o *intensidad* de ocurrencia, es decir, una cantidad con unidades del inverso del área (por ejemplo, $1/km^2$) correspondiente al número esperado de especímenes por unidad de área. En el modelo IPP discutido en esta sección, el especificar la tasa de ocurrencia es equivalente a especificar la probabilidad de ocurrencia de manera simultánea para cualquier tamaño de celda (Fithian y Hastie, 2013).

5.7.2. Notación

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

5.7.3. Descripción general del modelo

El modelo *IPP* es un modelo simple para un conjunto de puntos aleatorios \mathbf{W} que caen dentro algún dominio D y puede definirse por su función de intensidad como

$$\lambda : D \longrightarrow [0, \infty). \quad (5.26)$$

Para cualquier subconjunto $A \subseteq D$, al integrar $\lambda(w)dw$ se obtiene el número de registros de presencias en A , y se expresa como

$$\Lambda(A) = \int_A \lambda(w)dw \quad (5.27)$$

donde la única restricción impuesta es que la integral en (5.27) sea finita y se asume que $\Lambda(D) < \infty$.

Existen dos formas de expresar a un modelo IPP con intensidad λ . El primer enfoque consiste en que del número de puntos es una variable aleatoria Poisson con media $\Lambda(D)$ y condicionada sobre el número de puntos, sus ubicaciones son independientes e idénticamente distribuidas (i.i.d) con densidad $p_\lambda(w) = \lambda(w)/\Lambda(D)$. El otro enfoque

5.7. Modelo de Proceso Poisson no Homogéneo (IPP) para datos de “solo presencias”

consiste en pensar a un IPP como un límite continuo de un modelo de conteo Poisson independiente para una discretización muy fina de D . Si $N(A) = \#(\mathbf{W} \cap A)$ es el número de puntos que caen en el conjunto A , entonces

$$N(A) \sim Poisson(\Lambda(A)). \quad (5.28)$$

En el caso de un dominio finito y discreto $D = \{w_1, w_2, \dots, w_m\}$, el modelo IPP se reduce a un modelo Poisson discreto con $N(w_i) \sim Poisson(\lambda(w_i))$. En este sentido el modelo IPP puede verse como una discretización muy fina (es decir, de celdas muy pequeñas) de D .

Según [Warton y Shepherd \(2010\)](#), los supuestos del modelo puntual de Proceso Poisson no Homogéneo son:

1. Las ubicaciones de los n_1 eventos puntuales (y_1, \dots, y_n) son independientes.
2. La intensidad en el punto w_i [$\lambda(w_i)$ denotado como λ_i por conveniencia] que limita el número esperado de presencias por unidad de área, puede modelarse como función de j covariables explicativas. Asumiendo un modelo log-lineal

$$\lambda(w_i) = e^{\alpha + \beta' \mathbf{z}} \quad (5.29)$$

donde \mathbf{z} representa a las covariables medioambientales y puede incluir términos polinomiales, interacciones, o cualesquiera otro tipo de transformaciones de las variables originales.

Al interpretar el modelo IPP como una muestra *i.i.d.* de tamaño aleatorio, podemos apreciar que α únicamente multiplica a $\lambda(w)$ por una constante, la cual no tiene efecto sobre $p_\lambda(w) = \lambda(w)/\Lambda(D)$. Por otro lado, los demás parámetros β 's determinan por completo a p_λ .

5.7.4. Máxima verosimilitud para el modelo IPP

La log-verosimilitud en términos de la muestra de presencias en el modelo IPP es

5.7. Modelo de Proceso Poisson no Homogéneo (IPP) para datos de “solo presencias”

$$l(\alpha, \beta, \mathbf{y}) = \sum_{i:y_i=1} (\alpha + \beta' \mathbf{z}_i) - \int_D e^{\alpha + \beta' \mathbf{z}} dw - \log n_1! \quad (5.30)$$

Al diferenciar con respecto a α se obtiene

$$n_1 = \int_D e^{\alpha + \beta' \mathbf{z}} dw = \Lambda(D) \quad (5.31)$$

Note que para cualquier $\hat{\beta}$, $\hat{\alpha}$ juega el rol de constante de “normalización” que garantiza que $\lambda(w)$ integre a n_1 , es decir, el total de registros de presencias, y que por tanto, si n_1 no es de interés para el investigador tampoco lo será $\hat{\alpha}$. Al resolver para α en (5.31) e ignorando las constantes, se obtiene la log-verosimilitud parcial maximizada

$$l^*(\beta) = \sum_{i:y_i=1} \left(\beta' \mathbf{z}_i - \log \int_D e^{\beta' \mathbf{z}} dw \right) = \sum_{i:y_i=1} \log p_\lambda(w_i), \quad (5.32)$$

que corresponde a la misma log-verosimilitud que se obtiene al condicionar sobre n_1 y tratando a las w_i como una muestra aleatoria con densidad $p_\lambda(w) = \frac{e^{\beta' \mathbf{z}}}{\int_D e^{\beta' \mathbf{z}}}$.

Finalmente, al diferenciar (5.32) con respecto a β y dividiendo por n_1 resulta en:

$$\frac{1}{n_1} \sum_{i:y_i=1} \mathbf{z}_i = \frac{\int_D e^{\beta' \mathbf{z}} \mathbf{z} dw}{\int_D e^{\beta' \mathbf{z}} dw} = \mathbb{E}_{p_\lambda} \mathbf{z}. \quad (5.33)$$

Por lo tanto, la máxima verosimilitud para un modelo IPP log-lineal puede encontrarse mediante el algoritmo siguiente:

1. Estimar la densidad p_λ : encontrar $\hat{\beta}$ para la cual $\mathbb{E}_{\hat{p}_\lambda} \mathbf{z}$ coincide con la media empírica de la muestra de presencias \mathbf{z} .
2. Multiplicar \hat{p}_λ por n_1 : encontrar $\hat{\alpha}$ para la cual $\hat{\lambda}(w) = n_1 \cdot \hat{p}_\lambda(w)$.

5.8. Enfoques bayesianos en Modelos de distribución de especies

Evaluación numérica de la integral

Según Fithian y Hastie (2013), cuando no sea posible evaluar analíticamente la integral en (5.30), ésta puede evaluarse numéricamente con base en el *background*, por lo tanto, una aproximación de (5.30) es:

$$l(\alpha, \beta, \mathbf{y}) \approx \sum_{i:y_i=1} (\alpha + \beta' \mathbf{z}) - \frac{|D|}{n_0} \sum_{i:y_i=0} e^{\alpha + \beta' \mathbf{z}} - \log n_1! \quad (5.34)$$

donde $|D| = \int_D 1dw$ representa el área total de la región de estudio. Los puntos del *background* pueden ser tanto una muestra uniforme de D o bien un grid regular. Al repetir la derivación en que resultaron en las ecuaciones (5.30-5.33) derivamos:

$$\frac{|D|}{n_0} \sum_{i:y_i=0} e^{\alpha + \beta' \mathbf{z}} \approx n_1, \quad \frac{\sum_{i:y_i=0} e^{\beta' \mathbf{z} \mathbf{z}}}{\sum_{i:y_i=0} e^{\beta' \mathbf{z}}} \approx \frac{1}{n_1} \sum_{i:y_i=1} \mathbf{z} \quad (5.35)$$

En la práctica, ajustar el modelo IPP implica resolver la ecuación (5.35) para alguna muestra del *background*.

5.8. Enfoques bayesianos en Modelos de distribución de especies

Los métodos bayesianos brindan un enfoque alternativo para hacer inferencia estadística que difiere del enfoque clásico o frecuentista. La inferencia bayesiana estima la probabilidad de que una hipótesis sea verdadera dados los datos, y define dicha probabilidad como el grado de creencia en la verosimilitud de un evento. En el enfoque bayesiano, los parámetros de un modelo se asumen como variables aleatorias, para los cuales se incorpora conocimiento previo por medio de una distribución *a priori* que es independiente de los datos.

En los estudios acerca de distribución de especies que se han abordado, las probabilidades *a priori* de observar determinada especie (basándose en la literatura, en un estudio previo, o en una suposición) se han combinado con las probabilidades de

5.8. Enfoques bayesianos en Modelos de distribución de especies

ocurrencia condicional sobre los valores de los predictores medioambientales usando el teorema de Bayes:

$$P(H|Y) = \frac{F(Y|H)\pi(H)}{P(Y)}$$

Donde $P(H)$ corresponde a la probabilidad de la hipótesis dados los datos (denominada probabilidad posterior), $F(Y|H)$ es la verosimilitud dados los valores de las predictores medioambientales, y $\pi(H)$ es la probabilidad *a priori*. El denominador corresponde a la probabilidad marginal de los datos y se emplea como una constante de normalización. Sin embargo, en la mayoría de las ocasiones no se dispone de conocimiento previo por lo que generalmente se emplean distribuciones *a priori* no informativas (o planas).

Desde el punto de vista bayesiano, generalmente es necesario emplear algún algoritmo para simular cadenas de Markov Monte Carlo (MCMC) para obtener las densidades *a posteriori* de los parámetros involucrados en los modelos.

5.8.1. Algoritmo de Metrópolis-Hastings

El algoritmo de Metrópolis-Hastings es un algoritmo ampliamente utilizado en estadística y en física estadística. Es un método de cadenas de Markov Monte Carlo (MCMC) para obtener una secuencia de muestras aleatorias de una función de densidad de probabilidad de la cual el muestreo directo se dificulta. Esta secuencia se utiliza para aproximar la distribución *a posteriori* en el caso de inferencia bayesiana. La teoría de este algoritmo puede consultarse en [Casella y Robert \(2004\)](#) o en [Liang et al. \(2010\)](#).

El algoritmo construye una cadena de Markov apropiada definiendo las probabilidades de transición de la siguiente manera. Sea $Q(\theta^*|\theta)$ una distribución de transición (arbitraria) y definamos

$$\alpha(\theta^*, \theta) = \min \left\{ \frac{p(\theta^*|x)Q(\theta|\theta^*)}{p(\theta|x)Q(\theta^*|\theta)}, 1 \right\}.$$

5.8. Enfoques bayesianos en Modelos de distribución de especies

Algoritmo: Dado un valor inicial $\theta^{(0)}$, la t -ésima iteración consiste en:

1. generar una observación θ^* de $Q(\theta^*|\theta^{(t)})$;
2. generar una variable $u \sim U(0, 1)$;
3. si $u \leq \alpha(\theta^*, \theta^{(t)})$, hacer $\theta^{(t+1)} = \theta^*$; en caso contrario, hacer $\theta^{(t+1)} = \theta^{(t)}$.

Este procedimiento genera una cadena de Markov con distribución de transición

$$P(\theta^{(t+1)}|\theta^{(t)}) = \alpha(\theta^{(t+1)}, \theta^{(t)})Q(\theta^{(t+1)}|\theta^{(t)}).$$

La probabilidad de aceptación $\alpha(\theta^*, \theta)$, solo depende de $p(\theta|x)$ a través de un cociente, por lo que la constante de normalización no es necesaria.

Capítulo 6

Evaluación y selección de modelos en MDEs

En este capítulo se discutirá acerca de la evaluación y la selección de modelos en el campo de los MDEs. La primera sección se enfoca exclusivamente a la evaluación de los modelos, entendiendo como evaluación el desempeño en el rendimiento predictivo de los modelos que hemos abordado en los anteriores dos capítulos. Las predicciones son del tipo categóricas (hábitat adecuado o inadecuado), ordinal (idoneidad de hábitat alto, medio o bajo), o del tipo probabilística (probabilidad de ocurrencia de la especie) y que son validados con datos categóricos (presencia/ausencia de la especie).

La segunda sección aborda someramente dos criterios utilizados para seleccionar modelos como lo son el Criterio de Información de Akaike (*AIC*) y el su contraparte bayesiana denominado Criterio de Información de la Devianza (*DIC*).

6.1. Datos para la evaluación de los modelos

En los MDEs, la validación resulta más adecuada si se utilizan datos nuevos e independientes, es decir, datos que no fueron utilizados para estimar los parámetros del modelo. Si se utilizan los mismos datos para calibrar y evaluar los MDEs, lo que se denomina “resustitución”, se tiende a sobrestimar el rendimiento previsto del modelo para nuevas observaciones. Sin embargo, en la mayoría de la veces, no es posible ni

6.1. Datos para la evaluación de los modelos

factible coleccionar nuevos datos. En estos casos, un enfoque bastante frecuente es dividir los datos en dos partes, una parte se utiliza para entrenamiento, y la otra se usa para validar las predicciones (Franklin, 2009).

Un enfoque bastante simple para modelos que trabajan con datos de presencias-ausencias fue propuesto por Huberty (1994), y consiste en calcular la proporción de datos usados para validación mediante $1/(1 + \sqrt{p - 1})$, donde p es el número de predictores. El dividir los datos en dos subconjuntos es un ejemplo de doble validación cruzada. En el caso de k validaciones cruzadas, los datos se dividen k veces, dando k estimaciones de la exactitud que pueden posteriormente promediarse; así por ejemplo, los datos pueden dividirse 10 veces, con 9/10 de las observaciones se usan para entrenamiento y el restante 1/10 se utiliza para medir el desempeño del modelo; lo anterior se repite 10 veces y el desempeño del modelo se promedia. Otro enfoque comúnmente utilizado es mediante la técnica conocida como *bootstrap* (muestreo con reemplazo) (Franklin, 2009).

6.1.1. ¿Cómo se miden los errores?

La mayoría de los modelos que se han discutido en el presente trabajo predicen la probabilidad (o algún resultado en una escala continua cuya relación con la verosimilitud es monótona) de que la especie esté presente en un sitio determinado. Estos resultados de tipo probabilístico se convierten a resultados de tipo categórico usando un valor umbral para la probabilidad que permite distinguir entre un evento de otro, es decir, distinguir entre presencia o ausencia de la especie. Para un resultado de tipo binario (presencia-ausencia) y una vez fijado el valor umbral para la probabilidad (que define qué predicciones se clasifican como presencias y cuales como ausencias), se construye una matriz de confusión donde se clasifican los verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN) (Véase Tabla 6.1) en función de los valores predichos y los observados.

La Tabla (6.1) presenta algunas medidas de precisión útiles para evaluar el modelo, los cuales se muestran en la Tabla (6.2). Las medidas de precisión para modelos con resultados del tipo presencia-ausencia frecuentemente son evaluados usando medidas como la *Sensibilidad* (proporción de presencias actuales predichas con exactitud) y la *Especificidad* (proporción de ausencias actuales predichas con exactitud). *Kappa* es

6.1. Datos para la evaluación de los modelos

Tabla 6.1: Matriz de confusión para dos clases, presencia-ausencia de la especie.

		Observados		
		Presente	Ausente	Suma
Predichos	Presente	VP (verdadero positivo)	FP (falso positivo)	Total de presencias predichas
	Ausente	FN (falso negativo)	VN (verdadero negativo)	Total de ausencias predichas
	Suma	Total de observaciones de presencia	Total de observaciones de ausencia	Número total de observaciones

otra medida de importancia que figura en muchos estudios en relación a MDEs; se ha usado para evaluar predicciones categóricas aunque se sugiere no utilizarse como medida de comparación entre modelos donde la prevalencia de especies se considera diferente. Una medida alternativa a *Kappa* es *True skill statistic* (TSE) que se define como $\{1 - \text{máximo}(\text{Sensibilidad} + \text{Especificidad})\}$ donde la Sensibilidad y la Especificidad se calculan en función del umbral de la probabilidad para la cual se maximiza su suma.

Tabla 6.2: Medidas de precisión en función del umbral para datos binarios.

Medida	Modo de cálculo
Sensibilidad	$VP/(VP + FN)$
Razón de falsos negativos	$1 - \text{Sensibilidad}$
Especificidad	$VN/(VN + FP)$
Razón de falsos positivos	$1 - \text{Especificidad}$
Porcentaje de clasificación correcta	$(VP + VN)/n$
Valor predictivo positivo	$VP/(VP + FP)$
Razón de momios	$(VP \times VN)/(FP \times FN)$
Kappa	$\frac{[(VP+VN) - (((VP+FN)(VP+FP) + (FP+VN)(FN+VN)))/n]}{[n - (((VP+FN)(VP+FP) + (FP+VN)(FN+VN)))/n]}$
True skill statistic	$1 - \text{máximo}(\text{Sensibilidad} + \text{Especificidad})$

6.1.2. La elección del umbral

La probabilidad umbral o criterio para clasificación de casos, es el valor de la probabilidad por encima del cual, un caso se prevé que sea positivo, es decir, que la especie se considera presente. Por convención, éste umbral es frecuentemente fijado en 0.5, por ejemplo en la regresión logística que está implementada en muchos paquetes estadísticos. Sin embargo, como lo señala Franklin (2009), la selección del valor óptimo para dicho umbral está asociada con el costo de los diferentes tipos o errores de clasificación, que a su vez depende del uso previsto del modelo. El efecto del umbral en las

6.1. Datos para la evaluación de los modelos

tasas de error de omisión y comisión, también depende de la prevalencia de positivos en la muestra. La probabilidad umbral en la cual las razones de falsos positivos y falsos negativos son iguales tiende a igualar la prevalencia de los eventos en la muestra para muchos tipos de modelos. En otras palabras, si la prevalencia de positivos en la muestra es cercana al 10 %, entonces el valor umbral para la probabilidad en la cual la *Sensibilidad* se iguala a la *Sensitividad* es más o menos 0.1

Según [Freeman y Moisen \(2008\)](#) el valor por default de 0.5 subestima la prevalencia en el caso de especies raras, los mismos autores señalan que se ha sugerido que es mejor elegir un valor de umbral igual a la prevalencia o a la media de las probabilidades predichas que el valor por default cuando se trabaja con especies raras con baja prevalencia, sin embargo, ellos encontraron que al hacer lo anterior no mejoró el valor de *kappa*. Se señala también como importante, que para especies con alto rendimiento predictivo (medidos por el AUC: área bajo la curva), y prevalencias cercanas al 50 %, entonces cualquier criterio de optimización, incluyendo el valor por default de 0.5 tienden a converger en los mismos resultados, dando mapas igualmente útiles.

Por otra parte, [Liu et al. \(2005\)](#) sugieren que el uso de (a) la prevalencia observada, (b) el promedio de la probabilidad predicha, (c) la suma de la *Sensibilidad* y la *Especificidad*, (d) $Sensibilidad = Especificidad$, o (e) el punto en el gráfico ROC más cercano a la esquina superior izquierda, como valores umbral, dan resultados muy similares.

La Figura (6.1a) ilustra un ejemplo de distintos criterios que pueden aplicarse a las predicciones de los MDEs. La prevalencia de la especie es 0.10, el valor umbral de la probabilidad donde la $Sensibilidad = Especificidad$ alcanza su máximo es 0.13, y el umbral donde la $Sensibilidad + Especificidad$ es máxima es 0.09. En otras palabras, los tres criterios anteriores dan aproximadamente el mismo umbral. Note que el valor donde *Kappa* se maximiza es mucho mayor, 0.41, y el umbral donde la proporción de presencias predichas es igual a las observadas es 0.34.

6.1.3. Área bajo la curva ROC (AUC)

Dado que los propósitos que se le pueden dar a un MDEs son amplios, generalmente el estadístico provee mapas de probabilidades en lugar de mapas de presencias-ausencias,

6.1. Datos para la evaluación de los modelos

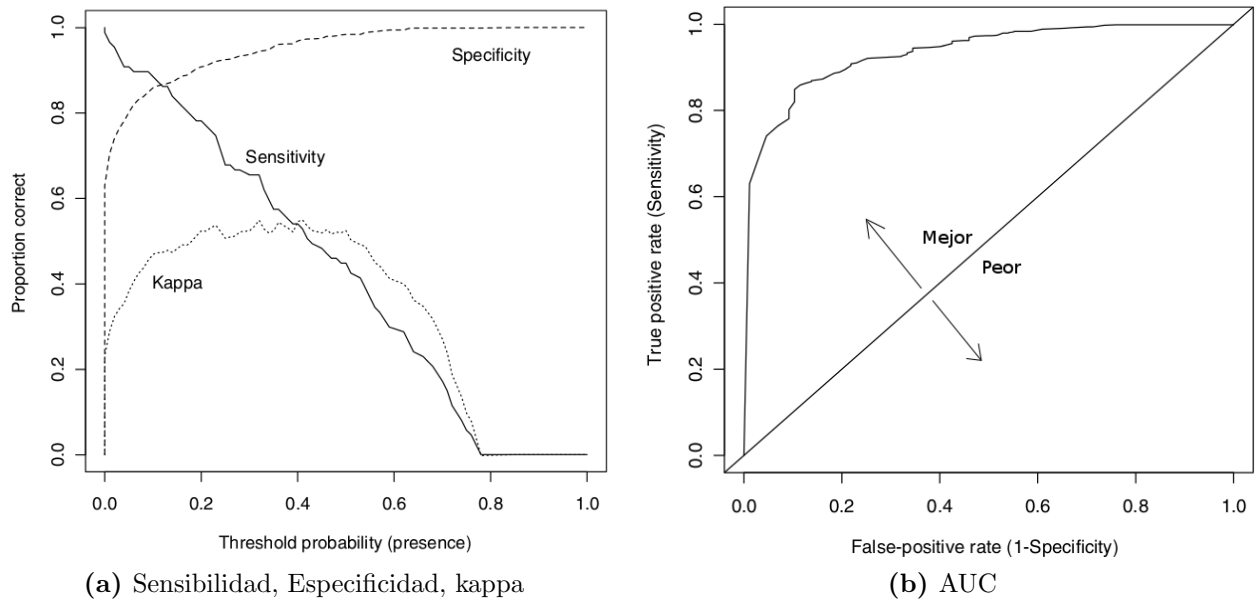


Figura 6.1: Representación gráfica de Sensibilidad, Especificidad, kappa y AUC

dejando a elección del investigador el valor umbral de acuerdo a los propósitos particulares de su estudio; por ello es mejor disponer de una medida que nos ayude a discriminar de un conjunto de MDEs al mejor modelo. Una medida independiente del umbral que se ha establecido para evaluar la bondad de ajuste en los MDEs es la denominada “área bajo la curva ROC ¹” AUC ².

El AUC se calcula sumando el área debajo de la curva ROC y toma valores que van desde 0.5 a 1.0, donde el valor 0.5 se interpreta como predicciones aleatorias y valores superiores indican un comportamiento del modelo mejor que el aleatorio (véase Figura (6.1b)). Esta estadística describe la capacidad global del modelo para discriminar entre los dos casos, es decir, la presencia de especies y la ausencia. Valores de AUC de 0.5 – 0.7 se consideran bajos (bajo rendimiento del modelo), 0.7 – 0.9 moderado, y > 0.9 alto. Los valores de AUC no se ven afectados por cambios en la prevalencia de la especie y por tanto es una estadística confiable en la comparación de modelos; algunos estudios han mostrado que el AUC no disminuye al aumentar la prevalencia de la especie (Franklin, 2009).

¹acrónimo de Receiver Operating Characteristic

²por sus siglas en Inglés

6.1.4. Evaluación en modelos de *solo presencias*

Como ya hemos abordado en el presente trabajo, en ocasiones no se disponen de registros de ausencias, por lo que existen modelos que solo trabajan con registros de presencia. En lugar de ausencias estrictas, en estos casos únicamente se disponen de datos de pseudo-ausencias (background), los cuales son registros que se toman como ausencias, más sin embargo, no en todas los casos resultan en verdaderas ausencias.

Selección del umbral

Para el caso de la elección del umbral óptimo, uno de los más recientes trabajos desarrollado por [Liu et al. \(2013\)](#) sugiere que el máximo de la suma de la sensibilidad y la especificidad (Max *SSS*) es un método prometedor para la selección del umbral cuando se trabajan con registros de *solo presencias*. Los autores probaron analíticamente y mediante simulación, que el método “max *SSS*” produce el mismo umbral tanto si se usa para datos de presencias-ausencias o cuando únicamente existen registros de presencias. En el mismo trabajo se señala que el umbral calculado mediante max *kappa* (el máximo valor de kappa) siempre se incrementa al incrementar las pseudo-ausencias por lo cual no se recomienda. Por otra parte, Max *SSS* es equivalente a maximizar la distancia vertical desde el punto sobre la curva ROC (curva de elevación) hasta la línea diagonal, de igual manera, Max *SSS* resulta en umbrales iguales a la que produce la medida de “True skill statistic (TSE)”.

AUC para solo presencias

Cuando se emplean datos de solo-presencias el uso de la curva ROC debe interpretarse como “ejemplos negativos” a todo el grid de celdas donde no se dispone de registros, aún cuando estos sitios proporcionen condiciones medioambientales propicias para la especie. El valor UAC_{PO} será por lo tanto siempre menor a 1 y mucho más pequeño a medida que la especie se distribuya en un área más amplia ([Phillips et al., 2004](#)). La interpretación por tanto, se hará en el sentido de que el AUC_{PO} describe la probabilidad de que el modelo clasifique un sitio de presencia (aleatorio) mayor que un sitio (aleatorio) del background ([Franklin, 2009](#)).

6.2. Selección de modelos

En esta sección abordaremos de manera superficial un tópico importante cuando se trabaja con MDEs, que es la selección de modelos, y en especial se hablará de dos medidas útiles para seleccionar modelos que son el Criterio de Información de Akaike (AIC) y su generalización en el contexto bayesiano conocido como Criterio de Información de la Devianza (DIC).

6.2.1. Criterio de Información de Akaike

El Criterio de Información de Akaike (*AIC*) fue propuesto originalmente por [Akaike \(1973\)](#) y se define como:

$$AIC = -2 \log l(\hat{\theta}) + 2k$$

donde $l(\hat{\theta})$ representa el valor máximo de la función de la log-verosimilitud y k corresponde al número de parámetros estimables en el modelo. El máximo de la función de log-verosimilitud corresponde a los valores de las estimaciones de máxima verosimilitud. Desde el punto de vista práctico se calcula el *AIC* para cada modelo dentro de un conjunto de posibles modelos y se selecciona aquel cuyo valor *AIC* sea el más pequeño.

Este criterio no asegura que se esté eligiendo al modelo verdadero, lo que si asegura es que dentro de un conjunto de modelos elige el mejor, por tanto, si el conjunto de modelos posibles todos tiene un pobre rendimiento en predicción, el *AIC* elegirá el menos malo. Ello conlleva a que el investigador de antemano tenga que considerar un conjunto de modelos consistentes con el problema que se esté estudiando.

Al examinar la ecuación de Criterio de Información de Akaike podemos notar que el primer término de la ecuación tiende a decrecer al adicionar más parámetros al modelo, mientras que el segundo término se mueve en dirección contraria a medida que ingresan más parámetros al modelo, este segundo término involucra por tanto, una penalización a medida que se ingresan más parámetros al modelo de tal forma que el modelo resultante sea más parsimonioso evitando el sobreajuste.

6.2.2. Criterio de información de la Devianza

El Criterio de Información de la Devianza (*DIC*) propuesto por Spiegelhalter *et al.* (2002) es una generalización del *AIC* para la selección de modelos en el paradigma bayesiano y se define como:

$$DIC = D(\bar{\theta}) + 2p_D = \bar{D} + p_D$$

donde $p_D = \mathbb{E}_{\theta|y}[D] - D(\mathbb{E}_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta})$, es decir, p_D es la media de la devianza *a posteriori* menos la devianza evaluada en la media *a posteriori* de los parámetros y en el caso de modelos no jerárquicos con poca información *a priori*, p_D se aproxima al verdadero número de parámetros. Dentro de un conjunto de posibles modelos, se elijé aquel cuyo valor del *DIC* sea el menor. Es importante señalar que el *DIC*, en el caso de que se utilice información *a priori* vaga (independiente de los datos) el *DIC* será mayor que en aquellos casos donde la información *a priori* es informativa.

Capítulo 7

Propuesta de modelos bayesianos para modelar la distribución de especies con registros de *solo presencias*

En éste capítulo se proponen dos metodologías en el marco de inferencia bayesiana que se han denominado *MaxBayes* e *IPPBayes*. Estas dos alternativas se han construido a partir del modelo conceptual de *Maxlike* y el modelo *IPP*, respectivamente. En el capítulo siguiente se muestra la implementación práctica de estos modelo.

7.1. Modelo *MaxBayes*

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

7.2. Modelo *IPPBayes*

para β_0 que represente dicho conocimiento.

Reescribiendo la ecuación (5.25) en términos del espacio medioambiental \mathbf{z} y suponiendo que β se distribuye *a priori* como una normal multivariada, es decir $\beta \sim NM(\beta_0, \mathbf{V}_0)$, donde β_0 y \mathbf{V}_0 corresponden a la media y la covarianza *a priori*, respectivamente; esto es, dichos hiperparámetros cuantifican el estado de conocimiento sobre los parámetros β antes de observar los datos. Aplicando el teorema de Bayes e ignorando los términos que no involucran a β , la distribución *a posteriori* queda expresada proporcionalmente como

$$p(\beta \mid y_i = 1, \mathbf{z}) \propto \prod_{i=1}^n \frac{\psi(y_i = 1 \mid \mathbf{z}; \beta)}{\sum_{x \in \mathcal{X}} \psi(y_i = x \mid \mathbf{z}; \beta)} \times \exp \left\{ -\frac{1}{2} (\beta - \beta_0)' \mathbf{V}_0^{-1} (\beta - \beta_0) \right\} \quad (7.1)$$

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

$$\text{logit}(\psi(y_i = 1 \mid \mathbf{z}; \beta)) = \log \frac{\psi(y_i = 1 \mid \mathbf{z}; \beta)}{1 - \psi(y_i = 1 \mid \mathbf{z}; \beta)} = \mathbf{z}'\beta.$$

Note que el cálculo de momentos y otras cantidades de interés a partir de (7.1) no puede realizarse analíticamente, por lo que es factible la implementación de algún algoritmo de simulación de cadenas de Markov Monte Carlo (MCMC, por sus siglas en Inglés), por ejemplo, el algoritmo de Metrópolis-Hastings.

7.2. Modelo *IPPBayes*

De manera análoga que en el modelo *MaxBayes*, puede construirse un enfoque bayesiano del modelo *IPP* para datos de solo presencias. El enfoque adoptado hace que D (área de interés) sea un espacio discreto compuesto por una rejilla de celdas muy pequeñas del espacio geográfico de interés. Resulta natural asignar a β una distribución *a priori* normal multivariada, es decir, $\beta \sim NM(\beta_0, \mathbf{V}_0)$, donde al igual que en el modelo *MaxBayes*, β_0 y \mathbf{V}_0 corresponden a la media y la covarianza *a priori*, respectivamente. Recuerde que α es el parámetro asociado al número de presencias

7.2. Modelo *IPPBayes*

en el área de estudio (abundancia) y que en muchas ocasiones el experto tiene una idea de en qué rango de valores se encuentra dicha cantidad. Este conocimiento puede reflejarse asignado una distribución *a priori* informativa para α .

Tomando como función de verosimilitud el antilogaritmo de (5.34) tal que $\boldsymbol{\beta} = (\alpha, \beta)'$ se tiene que

$$L(\boldsymbol{\beta} \mid \mathbf{y}; \mathbf{z}) \approx \frac{1}{n_1!} \exp\left(-\frac{|D|}{n_0} \sum_{i:y_i=0} e^{\mathbf{z}'\boldsymbol{\beta}}\right) \prod_{i:y_i=1} e^{\mathbf{z}'\boldsymbol{\beta}}. \quad (7.2)$$

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

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}; \mathbf{z}) &\propto \exp\left(-\frac{|D|}{n_0} \sum_{i:y_i=0} e^{\mathbf{z}'\boldsymbol{\beta}}\right) \prod_{i:y_i=1} e^{\mathbf{z}'\boldsymbol{\beta}} \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \mathbf{V}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \end{aligned} \quad (7.3)$$

De manera similar que en *MaxBayes*, para el cálculo de momentos y otras cantidades de interés a partir de (7.3) es factible la implementación de algún algoritmo MCMC.

Capítulo 8

Caso de estudio

Para ilustrar de forma práctica los métodos estadísticos propuestos en el capítulo 7, se implementaron ambos modelos en un conjunto de datos simulados, y en un conjunto de datos reales. En el ejemplo con datos reales, las distribución potencial del género *Dalea* obtenida mediante los modelos *MaxBayes* e *IPPBayes* se compararon con los de *Maxent*.

8.1. Simulación de datos

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

8.2. Datos de género *Dalea*

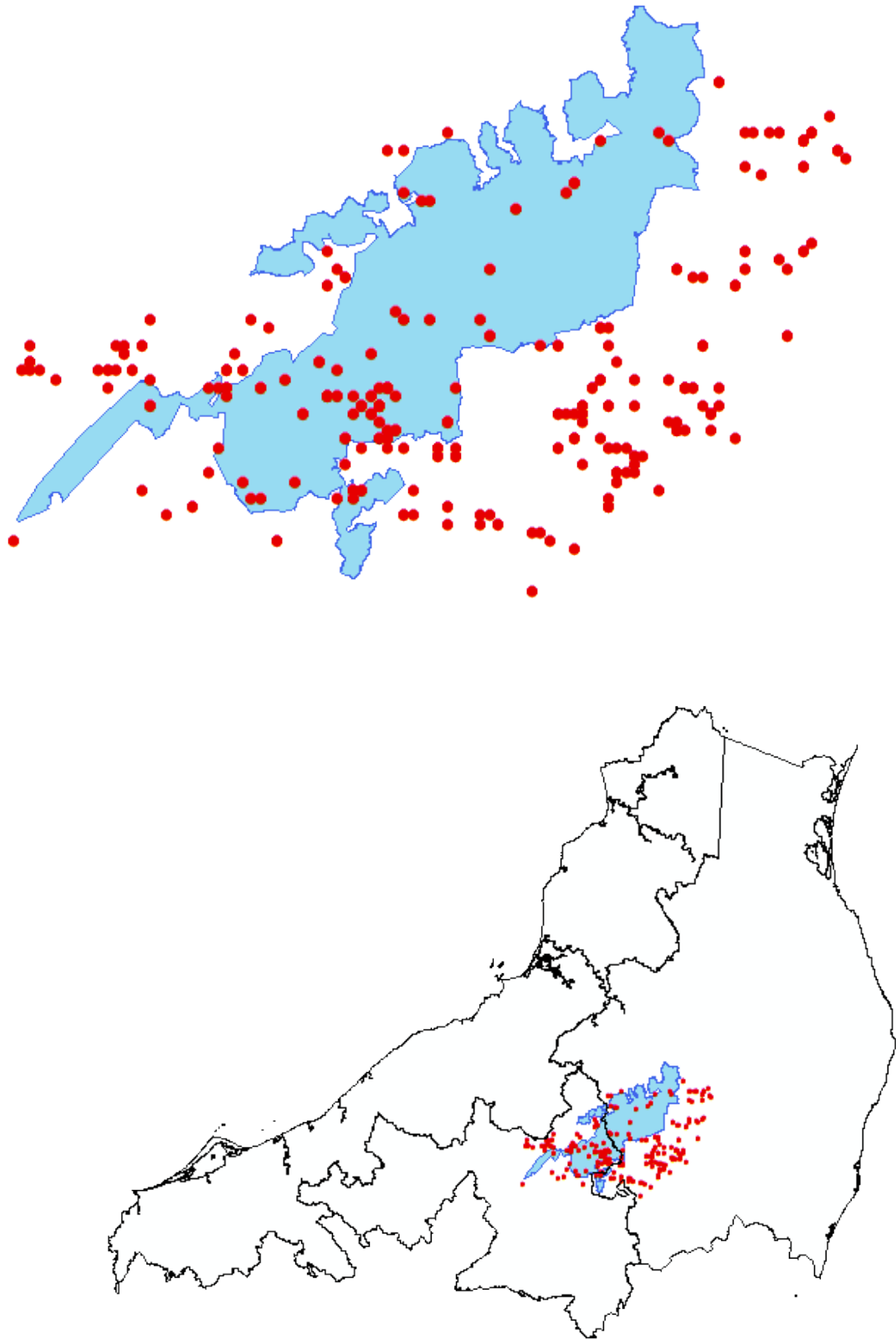
Para implementar el modelo *MaxBayes* se partió de la ecuación (7.1), asignando para β una distribución normal con media cero y varianza grande, para reflejar el desconocimiento *a priori*. Ésto es, $\beta \sim NM(\mathbf{0}, \mathbf{V}_0)$, donde $\beta = (\beta_0, \beta_1, \beta_2)'$ son los parámetros asociados al intercepto y a las covariables z_1 y z_2 simuladas, y

$$\mathbf{V}_0 = \begin{bmatrix} 10^5 & 0 & 0 \\ 0 & 10^5 & 0 \\ 0 & 0 & 10^5 \end{bmatrix}.$$

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

8.2. Datos de género *Dalea*

Como ejemplo con datos reales se eligió el género *Dalea* cuyos datos constan de 301 presencias (Ver figura 8.1b) provenientes de la Reserva de la Biosfera Tehuacán-Cuicatlán, disponibles en el portal de la CONABIO como parte del proyecto Q014 (<http://www.conabio.gob.mx/remib/doctos/remibnodosdb.html>). El género *Dalea* se considera endémico de la zona que comprende a la Provincia Florística del Valle de Tehuacán-Cuicatlán (Méndez *et al.*, 2004), cuyos límites abarcan los estados de Oaxaca y Puebla. La zona total de estudio (*landscape*) se consideró a los estados de Veracruz, Oaxaca y Puebla. La información de las covariables medioambientales fue descargado del portal <http://www.worldclim.org/bioclim> que corresponde a la base de datos global BIOCLIM. Las covariables medioambientales utilizadas fueron la precipitación anual (Pp), la altitud sobre el nivel del mar (Alt), la temperatura media anual (Tmedia), el rango de la temperatura anual (RangoT), además de la latitud (Lat) y la longitud (Lon). En la figura (8.2) se muestran los mapas correspondientes a las cuatro primeras covariables listadas anteriormente.



(b) Presencias *Dalea*.

(a) Zona de estudio.

Figura 8.1: Zona de estudio y Reserva Cuicatlán-Tehuacán.

8.2. Datos de género *Dalea*

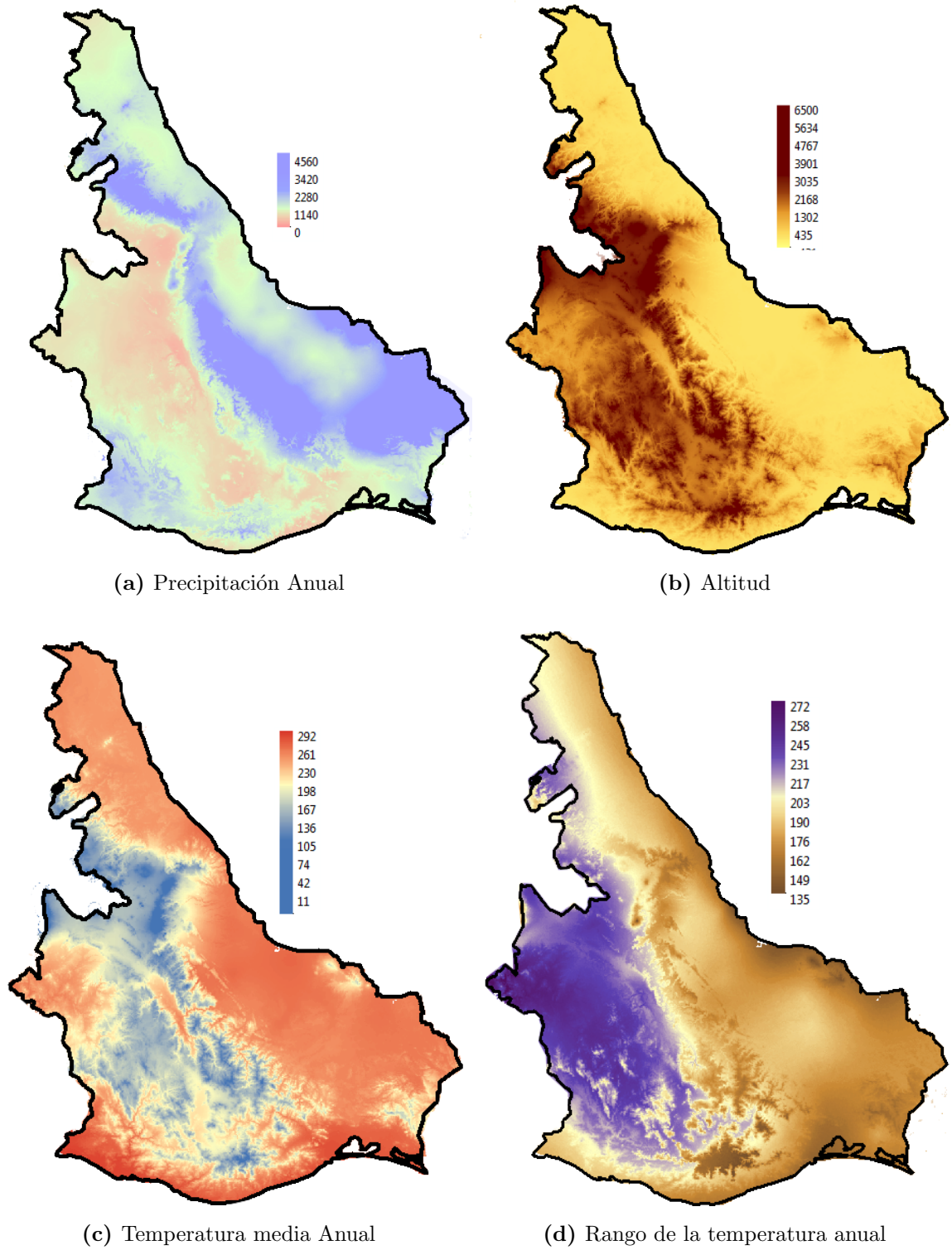


Figura 8.2: Covariables medioambientales

8.2. Datos de género *Dalea*

Tabla 8.1: Nombre de covariables medioambientales

Variable	Clave
Precipitación anual (mm)	Pp
Altitud (msnm)	Elev
Temperatura media Anual (°C)	Tmedia
Rango de la temperatura anual (°C)	RangoT
Latitud (°)	Lat
Longitud (°)	Lon

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

Para ajustar el modelo *MaxBayes* se partió de la ecuación (7.1), asignando para β una distribución normal con media cero y varianza grande, para reflejar el desconocimiento *a priori*. Esto es, $\beta \sim NM(\mathbf{0}, \mathbf{V}_0)$, donde $\beta = (\beta_0, \beta_1, \dots, \beta_6)'$ son los parámetros asociados al intercepto y a las covariables medioambientales, y

$$\mathbf{V}_0 = \begin{bmatrix} 10^5 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 10^5 \end{bmatrix}.$$

En el caso del modelo *IPPBayes*, se partió de la ecuación (7.3). Se especificó la misma distribución *a priori* para β utilizada en *MaxBayes*. El *intensidad de ocurrencia* se modeló como $\lambda(w_i) = \exp(\mathbf{z}'\beta)$ donde $\beta = (\alpha, \beta_1, \dots, \beta_6)$ es el vector de coeficientes que incluye al intercepto α .

8.3. Simulación de la distribución *a posteriori* mediante MCMC

Tabla 8.2: Covariables medioambientales

Celda	Presencias	Altitud	Tmedia	Pp	RangoT	Lon	Lat
1	0	6.0	27.7	699.0	16.2	-96.616	15.682
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
76186	1	2020.0	17.2	802.0	22.4	-96.985	17.254
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
110758	3	2313.0	14.9	676.0	23.8	-97.468	17.770
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
243070	0	3.0	24.4	857.0	19.0	-98.187	22.437

8.3. Simulación de la distribución *a posteriori* mediante MCMC

Las distribuciones *a posteriori* derivadas de (7.1) y (7.3) no pueden integrarse en forma analítica por lo que se utilizó el algoritmo Metrópolis-Hastings mediante el paquete *MHadaptive* (Chivers, 2012) de R (R Core Team, 2013). En ambos ejemplos de aplicación, tanto en *MaxBayes* como en *IPPBayes* se simularon 100,000 valores de la distribución *a posteriori*, tomando como *burnIn* a los primeras 50,000 iteraciones. A las restantes 50,000 valores de cada cadena se le aplicó un adelgazamiento (*thinning*), reteniendo únicamente valores a distancia 5. Posteriormente se calcularon los estimadores bayesianos bajo pérdida 0 – 1 para cada componente de β , esto es, la moda de los valores simulados después del periodo de calentamiento.

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

Todos los códigos empleados, se incluyen en el Apéndice de ésta tesis.

Capítulo 9

Resultados y discusión

9.1. Datos de simulación

En la Tabla (9.1) se presenta la información correspondiente al modelo *MaxBayes*. El número de registros n , para ajustar el modelo en este ejemplo de simulación fue de 2,000, 1,000 y 100. La tabla contiene las estimaciones de los parámetros β con sus respectivos intervalos de máxima probabilidad *a posteriori* (HPD). La figura (9.1) presenta la distribuciones *a posteriori* de cada componente de β , las líneas verticales delimitan los intervalos HPD respectivos. Note que la estimación de β_0 asociado al intercepto en el modelo *MaxBayes* es muy similar aún variando n , y cercana al valor real $\beta_0 = -1$, sin embargo, la incertidumbre asociada a la estimación del parámetro crece, lo que se refleja en intervalos HPD con mayor longitud. Un comportamiento similar lo presentan las estimaciones para β_1 y β_2 , donde las estimaciones correspondientes se acercan al valor real ($\beta_1 = 2$ y $\beta_2 = -2$), y los intervalos HPD tienen mayor amplitud a medida que n disminuye. Recuerde que este ejemplo, se han utilizado distribuciones *a priori* no informativas, por lo que la inferencia del modelo bayesiano se basa en los datos y que por tanto, hereda las propiedades asintóticas de los estimadores de máxima verosimilitud (EMV) y a medida que n crece, la incertidumbre asociada a los parámetros disminuye.

La prevalencia estimada por *MaxBayes* se resume en la tabla (9.2). Note que aún cuando el número de presencias es pequeño ($n = 100$), la prevalencia estimada es cercana a la real. Recuerde que a través de β_0 *MaxBayes* estima la prevalencia, por lo

9.1. Datos de simulación

Tabla 9.1: Resumen de *MaxBayes* para distintos n (simulación).

$n = 2000$					
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se
Intercepto	β_0	-0.88	-1.06	-0.70	0.09
z_1	β_1	1.87	1.73	2.16	0.11
z_2	β_2	-1.86	-2.19	-1.75	0.11
$n = 1000$					
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se
Intercepto	β_0	-0.96	-1.13	-0.62	0.13
z_1	β_1	1.84	1.66	2.26	0.15
z_2	β_2	-1.79	-2.29	-1.68	0.16
$n = 100$					
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se
Intercepto	β_0	-1.05	-1.61	0.52	0.57
z_1	β_1	1.97	1.35	4.60	0.89
z_2	β_2	-1.96	-3.69	-1.02	0.70

que en el caso de muestras pequeñas, cualquier información de expertos en relación a la especie de interés que ayude a identificar la prevalencia, debe hacerse a través de una distribución *a priori* informativa sobre todo para el parámetro β_0 . La figura (9.2) muestra el área bajo la curva ROC (AUC) correspondiente al modelo *MaxBayes*. Este estadístico describe la capacidad global del modelo para discriminar entre los dos casos, esto es, presencia y ausencia de la especie. Valores del AUC superiores a 0.9 se interpreta que el modelo tiene alto rendimiento predictivo, tal como en el presente ejemplo de estudio.

Tabla 9.2: Prevalencia estimada por *MaxBayes* bajo diferentes n (simulación).

prevalencia real = 0.38	
n	prevalencia estimada
100	0.37
1000	0.38
2000	0.39

Por otra parte, en la Tabla (9.3) se presenta la información del modelo *IPPBayes*. Note que la estimación del parámetro α es quien tiene mayor variación a medida que n_1 lo hace, mientras que las estimaciones para β_1 y β_2 son menores. Al disminuir el número de los registros, los intervalos HPD tienen mayor amplitud en virtud de que se han utilizado distribuciones *a priori* no informativas para α , β_1 y β_2 y la inferencia bajo las distribuciones *a posteriori* se basan casi por completo en la verosimilitud. Lo anterior puede verse también en la figura (9.3) donde se presentan las distribuciones *a posteriori* para cada parámetro del modelo; las líneas verticales delimitan los HPD

9.1. Datos de simulación

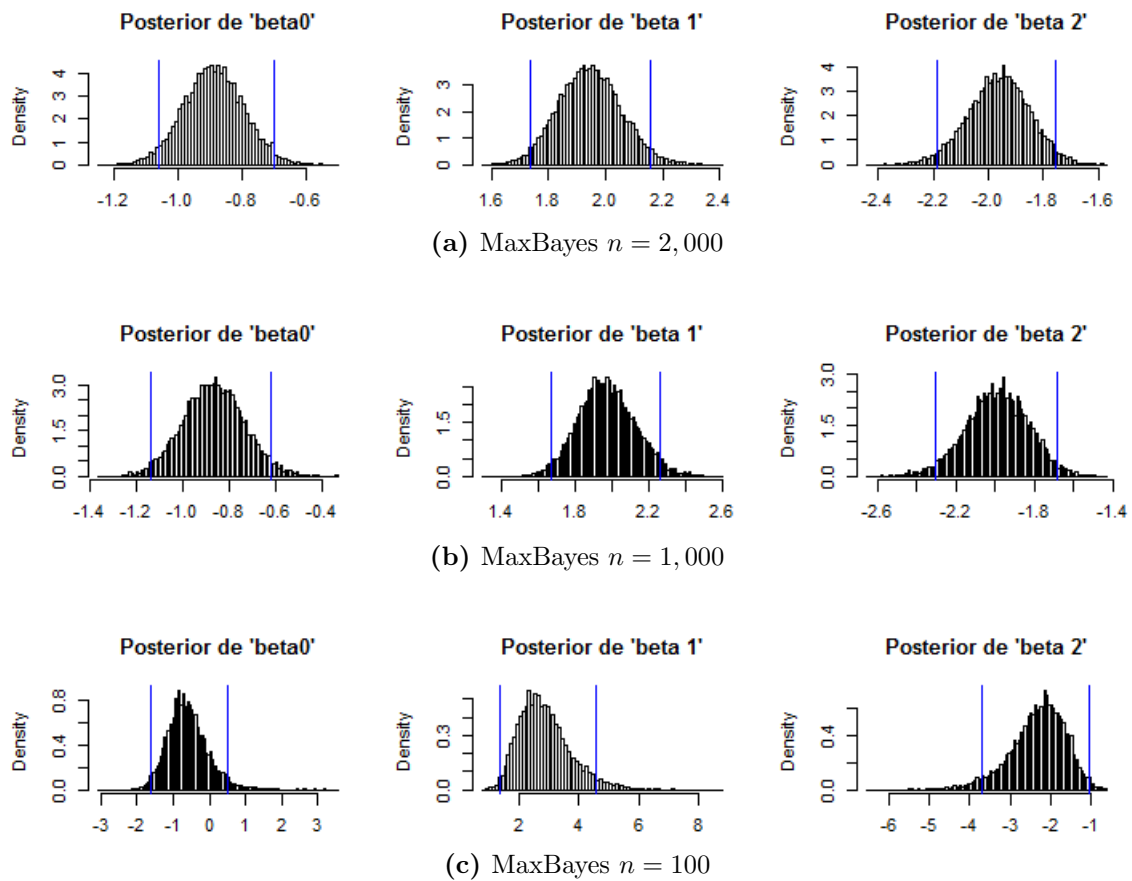


Figura 9.1: Distribuciones *a posteriori* de los parámetros de *MaxBayes* para diferentes n (simulación).

9.1. Datos de simulación

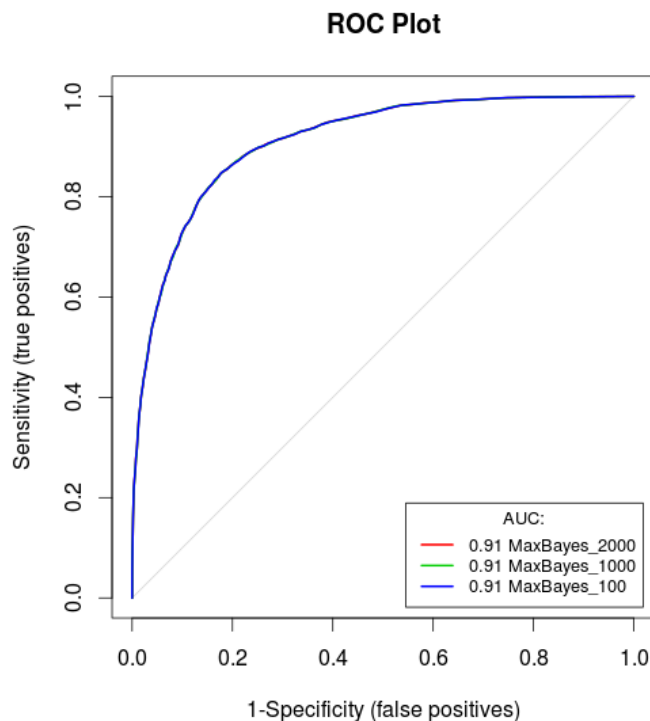


Figura 9.2: Área bajo la curva ROC (AUC) de *MaxBayes* (simulación).

para cada parámetro. Como ya se mencionó, en el modelo *IPPBayes*, α garantiza que $n_1 \approx \sum_D e^{z'\hat{\beta}}$, es decir, el número de registros utilizados para ajustar el modelo. Note que a medida que n_1 disminuye, la estimación de α es menor, ajustando en cada caso las intensidades de ocurrencia hacía abajo. En este sentido el modelo *IPPBayes* bajo distribuciones *a priori* no informativas proporciona intensidades de ocurrencia relativas al tamaño de registros utilizados para ajustar el modelo.

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

Una comparativa de los modelos *MaxBayes* e *IPPBayes* en términos del DIC se presenta en la Tabla (9.4). Note que en todos los casos, el valor del DIC para el modelo *IPPBayes* es menor, lo que sugiere que es un modelo más parsimonioso. Si el interés del investigador fuera únicamente conocer como se distribuye la especie en el

9.1. Datos de simulación

Tabla 9.3: Resumen de *IPPBayes* para distintos n_1 (simulación).

$n_1 = 2000$					
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se
Intercepto	α	-1.94	-2.00	-1.89	0.03
z_1	β_1	0.58	0.54	0.63	0.02
z_2	β_2	-0.57	-0.61	-0.53	0.02
$n_1 = 1000$					
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se
Intercepto	α	-2.62	-2.72	-2.56	0.04
z_1	β_1	0.59	0.52	0.64	0.03
z_2	β_2	-0.55	-0.62	-0.49	0.03
$n_1 = 100$					
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se
Intercepto	α	-4.98	-5.24	-4.73	0.13
z_1	β_1	0.68	0.48	0.87	0.10
z_2	β_2	-0.49	-0.69	-0.30	0.10

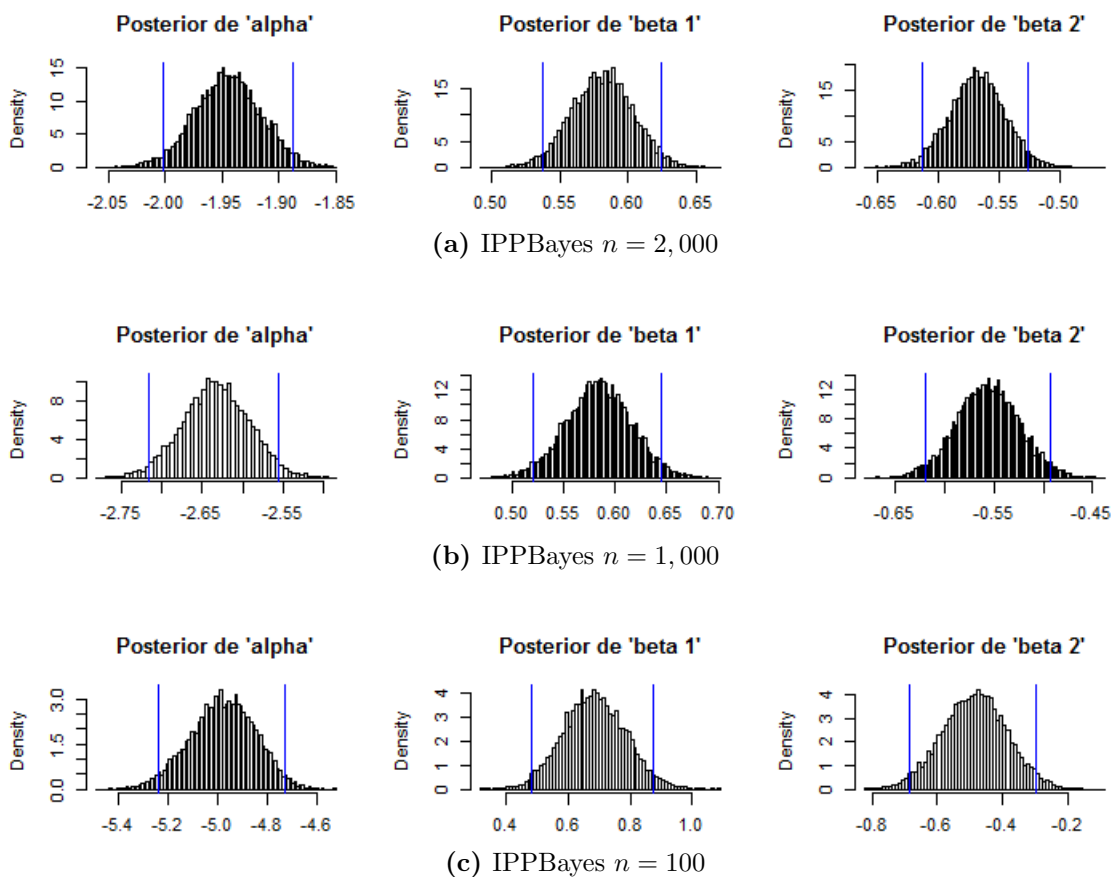


Figura 9.3: Distribuciones *a posteriori* de los parámetros de *IPPBayes* para diferentes n (simulación).

9.2. Género *Dalea*

espacio geográfico, el modelo *IPPBayes* resultaría el mejor.

Tabla 9.4: *MaxBayes* vs *IPPBayes* en términos del DIC (Simulación)

n	<i>MaxBayes</i>	<i>IPPBayes</i>
2000	87,813	22,841
1000	43,987	14,975
100	4,501	2,747

La prueba de convergencia de [Gelman y Rubin \(1992\)](#) que se aplicó a las muestras simuladas de la distribución *a posteriori*, tanto para *MaxBayes* como en el modelo *IPPBayes*, indicó que dichas cadenas convergieron a la distribución estacionaria. Los resultados de dichas pruebas se presentan en el Apéndice.

9.2. Género *Dalea*

En las figuras (9.4a-9.4b) se observan los mapas de probabilidad de presencia obtenidas con *MaxBayes* y la salida logística de *Maxent*, respectivamente. En la Tabla (9.5) se resumen las estimaciones de β tanto para el modelo *MaxBayes* como el modelo *Maxent*. Note que los signos asociados a cada estimador son iguales en ambos modelos, lo que nos da cuenta en qué sentido las covariables afectan la presencia del género estudiado. También se resumen los intervalos de máxima probabilidad *a posteriori* (HPD) asociadas a la estimación de cada parámetro en el modelo *MaxBayes*. En la figura (9.5) se presentan las distribuciones *a posteriori* para cada parámetro del modelo *MaxBayes*, donde la líneas verticales delimitan los intervalos HPD correspondientes.

Tabla 9.5: Resumen de *MaxBayes* y *Maxent* para distintos tamaños de muestra.

$n = 301$	<i>MaxBayes</i>					<i>Maxent</i>
Variable	Parámetro	Estimación	HPD Inf	HPD Sup	se	Estimación
Intercepto	β_0	-10.39	-12.26	-9.49	0.72	-
Altitud	β_1	-11.90	-13.17	-8.38	1.26	-19.45
Tmedia	β_2	-11.05	-12.22	-8.21	1.06	-27.38
Pp	β_3	-5.35	-6.57	-4.46	0.55	-24.95
RangoT	β_4	4.34	3.08	4.92	0.48	10.08
Lon	β_5	1.33	-0.27	2.48	0.72	0.74
Lat	β_6	-1.40	-2.26	-0.71	0.40	-5.49

Tal como se aprecia en la Figura (9.4), tanto *Maxent* como *MaxBayes* proporcionan el mismo patrón de predicción, aunque como señala [Royle et al. \(2012\)](#), *Maxent* tiende

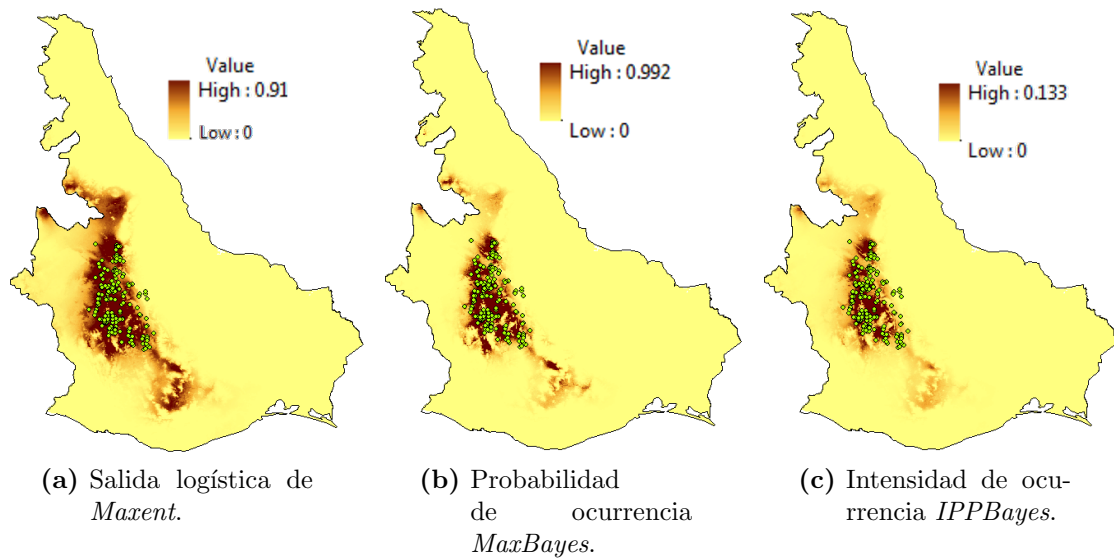


Figura 9.4: Distribución potencial del género *Dalea* obtenidos mediante los modelos *Maxent*, *MaxBayes* e *IPPBayes*.

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

Por otra parte, en la figura (9.4c) se ilustra el mapa de intensidades de ocurrencia del género *Dalea* mientras que en la Tabla (9.6) se resume las estimaciones de los parámetros del modelo *IPPBayes*, también se incluyen los intervalos de máxima probabilidad *a posteriori* respectivos. La figura (9.6) presenta las distribuciones *a posteriori* para cada parámetro del modelo, donde las líneas azules delimitan los intervalos HPD.

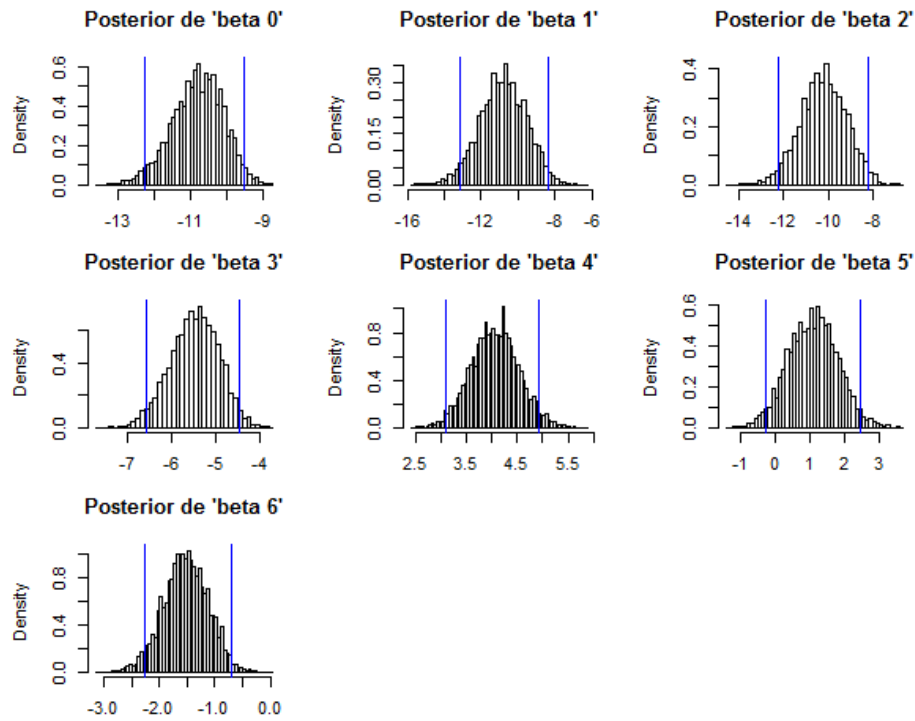


Figura 9.5: Distribuciones *a posteriori* de los parámetros del modelo *MaxBayes*.

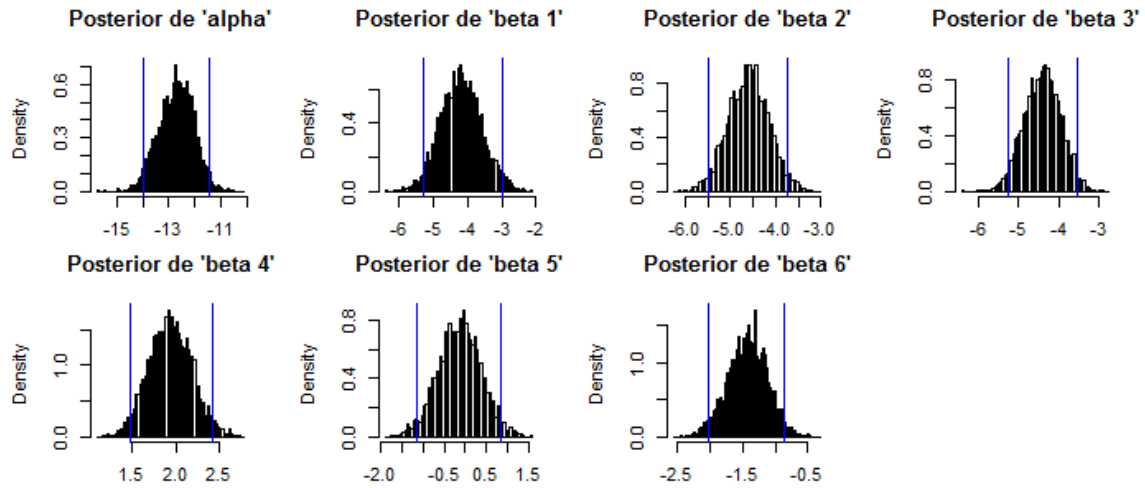


Figura 9.6: Distribuciones *a posteriori* de los parámetros del modelo *IPPBayer*.

Tabla 9.6: Resumen del modelo *IPPBayes*

<i>Variable</i>	Parámetro	Estimación	HPD inf	HPD sup	se
Intercepto	α	-12.99	-13.97	-11.43	0.64
Altitud	β_1	-3.95	-5.31	-2.97	0.59
Tmedia	β_2	-4.31	-5.48	-3.73	0.44
Pp	β_3	-4.73	-5.26	-3.55	0.44
RangoT	β_4	1.92	1.47	2.42	0.24
Lon	β_5	-0.26	-1.17	0.84	0.49
Lat	β_6	-1.37	-2.02	-0.86	0.29

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

Recuerde que en el modelo *IPPBayes* lo que se modela es la intensidad de ocurrencia de la especie por unidad de área. Note que el concepto de intensidad de ocurrencia está íntimamente ligada al concepto de probabilidad de presencia, dado que en aquellos sitios donde la intensidad de ocurrencia es mayor corresponde a las máximas probabilidades asignadas por *MaxBayes*, y como se observan en la figuras (9.4b-9.4c), el modelo *IPPBayes* proporciona la misma distribución potencial del género *Dalea* que el modelo *MaxBayes*. Es importante destacar que las intensidades de ocurrencia calculadas el modelo *IPPBayes* son intensidades de ocurrencia *relativas* ya que los registros utilizados en el modelo conforman una fracción de la población de la especie de interés, sin embargo, constituye una forma alternativa para modelar la distribución potencial de la especie.

La Tabla (9.7) presenta los valores del DIC para ambos modelos. Al igual que en ejemplo de simulación, el modelo *IPPBayes* resulta ser mejor que el modelo *MaxBayes* dado que el valor del DIC correspondiente es menor.

Tabla 9.7: *MaxBayes* vs *IPPBayes* en términos del DIC

n	<i>MaxBayes</i>	<i>IPPBayes</i>
301	15,449	8,515

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

9.2. Género *Dalea*

resultados de dichas pruebas se presentan en el Apéndice.

Capítulo 10

Conclusiones y recomendaciones

10.1. Conclusiones

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

En lo respecta a los ejemplos de aplicación se concluye lo siguiente:

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

10.2. Recomendaciones

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

10.2. Recomendaciones

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

Referencias

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory*, 267–281.
- Araujo, M., Pearson, R., Thuiller, W. y Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, 11, 1504–1513.
- Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Model*, 157, 101–118.
- Austin, M. (2006). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200, 1–19.
- Benediktsson, J., Swain, P. y Ersoy, O. (1993). Conjugate-gradient neural networks in classification of multisource and very-high-dimensional remote sensing data. *International Journal of Remote Sensing*, 14, 2883–2903.
- Benito Garzón, M., Blazek, R., Neteler, M., Sánchez de Dios, R., Ollero, H. y Furlanello, C. (2006). Predicting habitat suitability with machine learning models: the potential area of *Pinus Silvestris* L. in the iberia peninsula. *Ecological modelling*, 197, 383–393.
- Casella, G. y Robert, C. (2004). *Monte Carlo Statistical Methods*. Springer.
- Chivers, C. (2012). *MHadaptive: General Markov Chain Monte Carlo for Bayesian Inference using adaptive Metropolis-Hastings sampling*. R package version 1.1-8.
- Civco, D. (1993). Artificial neural networks for land-cover classification and mapping. *International Journal of Geographic Information Systems*, 7, 173–186.
- Elith, J., Phillips, S. J., Hastie, T., Dukiv, M., Chee, Y. E. y Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17, 43–57.
- Elith, J. *et al.* (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography*, 29, 129–151.

REFERENCIAS

- Fithian, W. y Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Annals of Applied Statistics*, 7, 1917–1939.
- Fitzgerald, R. y Lees, B. (1992). *The application of neural network to floristic classification of remote sensing and GIS data in complex terrain*. Proceedings of the XVII Congress ISPRS.
- Franklin, J. (2009). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press.
- Freeman, E. A. y Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in term of predicted prevalence and kappa. *Ecological Modeling*, 217, 48–58.
- Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38, 367–378.
- Gelman, A. y Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.
- Graham, C. H., Moritz, C. y Williams, S. E. (2006). Habitat history improves prediction of biodiversity in rainforest fauna. *The National Academy of Sciences of the USA*, 1, 632–636.
- Guisan, A., Edwards Jr, T. C. y Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157, 89–100.
- Hastie, T. J. y Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.
- Hastie, T. J., Tibshirani, R. J. y Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. springer.
- Huberty, C. (1994). *Applied Discriminant Analysis*. New York, USA: Wiley Inter-science.
- Hutchinson, G. (1957). *Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology*.
- Latimer, A. M., Wu, S., Gelfan, A. E. y Silander, J. A. J. (2006). Building statistical models to analyze species distributions. *Ecological Applications*, 16(1), 33–50.
- Leathwick, J., Elith, J. y Hastie, T. (2006b). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, 199, 188–196.

REFERENCIAS

- Leathwick, J., Rowe, D., Richardson, J., Elith, J. y Hastie, T. (2005). Using multivariate adaptive splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biol*, 50, 234–252.
- Liang, F., Liu, C. y Carroll, R. (2010). *Advanced Markov Chain Monte Carlo Methods*. Wiley.
- Liu, C., White, M. y Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, 40, 778–789.
- Liu, C. R., Berry, P. M., Dawson, T. P. y Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28, 385–393.
- Manly, B., McDonald, L., Thomas, D., McDonal, T. y Erickson, W. (2002). *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. New York, NY, USA: Kluwer Press.
- Méndez, L., Ortiz, E. y Villaseñor, J. (2004). Las Magnoliophyta endémicas de la porción xerofítica de la provincia florística del Valle de Tehuacán-Cuicatlán, México. *Anales del Instituto de Biología. UNAM. Serie Botánica*, 75(1), 87–104.
- Merow, C. y Silander, J. A. (2014). A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution*, 5, 215–225.
- Moisen, G. (2008). Classification and Regression Trees. *Encyclopedia of Ecology: Ecological Informatics*, 582–588.
- Moisen, G. y Frescino, T. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological modelling*, 157, 209–225.
- Muñoz, L. y Felicísimo, A. M. (2008). Comparison of statistical methods commonly used in predicting used in predictive modelling. *Journal of Vegetation Science*, 285–292.
- Nelder, J. y Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135, 370–384.
- Olden, J. D., Lawler, J. J. y Poff, N. L. (2008). Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology*, 83, 171–193.
- Pawitan, Y. (2001). *In all Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.
- Pearson, R. (2007). *Species Distribution Modeling for Conservation Educators and Practitioners. Synthesis*. American Museum of Natural History, Lesson in Conservation.

REFERENCIAS

- Phillips, S., Dudik, M. y Schapire, R. (2004). A Maximum Entropy Approach to Species Distribution Modeling. *Proceedings of the Twenty-Firts International Conference on Machine Learning*, 1–8.
- Phillips, S. *et al.* (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19, 181–197.
- Plummer, M., Best, N., Cowles, K. y Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6, 1, 7–11.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ridgeway, G. (1999). The stage of boosting. *Computing Science and Statistics*, 31, 172–181.
- Royle, J., Chandle, R. B., Yackulic, C. y Nichols, J. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling speciesdistributions. *Methods in Ecology and Evolution*, 3, 545–554.
- Spiegelhalter, D. *et al.* (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, 64(4), 583–639.
- Thomas, C. *et al.* (2004). Extinction risk from climate change. *The National Academy of Sciences of the USA*, 427, 145–148.
- Thuiller, W., Lavorel, S. y Sykes, M. (2006). Using niche-based modelling to asses the impact of climate change on tree functional diversity in Europe. *Diversity and distribution*, 12, 49–60.
- Thuiller, W. *et al.* (2005). Niche-based modelling as a tool for predicting the risk of alien plant invasion at a global scale. *Global Change Biology*, 11, 2234–2250.
- Van Neil, K., Laffan, S. y Less, B. (2004). Effect of error in the DEM on enviromental variables for predictive vegetation modelling. *International Journal of Vegetable Science*, 15, 747–756.
- Warton, D. y Shepherd, L. (2010). Poisson Point Process Models solve the “Pseudo-absence problem for presence-only data in ecology”. *The Annals of Applied Statistics*, 4, 1383–1402.
- Yee, T. W. y Mitchel, N. D. (1991). Generalized additive models in plant ecology. *Journal of vegetation sciencie*, 2, 587–602.

Apéndice

.1. Pruebas de convergencia para el modelo *MaxBayes* e *IPPBayes*

En esta sección se presentan las pruebas de convergencia del modelo *MaxBayes* e *IPPBayes* en ambos ejemplos de aplicación.

.1.1. Datos simulados

Las cadenas simuladas de las distribuciones *a posteriori* para los tamaños de muestra $n = 2000$, $n = 1000$, $n = 100$, para los modelos *MaxBayes* e *IPPBayes* se presentan en la figuras (.1) y (.2), respectivamente. Las cadenas no incluyen los valores de las cadenas en el periodo de calentamiento. Note que las cadenas oscilan como un proceso de ruido blanco lo cual implica que han convergido a la distribución estacionaria.

Los resultados de la prueba de [Gelman y Rubin \(1992\)](#) al correr tres cadenas con diferentes valores iniciales se resumen en las Tablas (.1) y (.2) para ambos modelos. De acuerdo a dichos resultados, el factor de \hat{R} (estimador de reducción potencial de escala) en los todos los casos es igual a 1 lo cual evidencia que las cadenas simuladas en cada uno de los modelos se han trasladado y que por tanto, pertenecen a la distribución estacionaria.

.1 Pruebas de convergencia para el modelo *MaxBayes* e *IPPBayes*

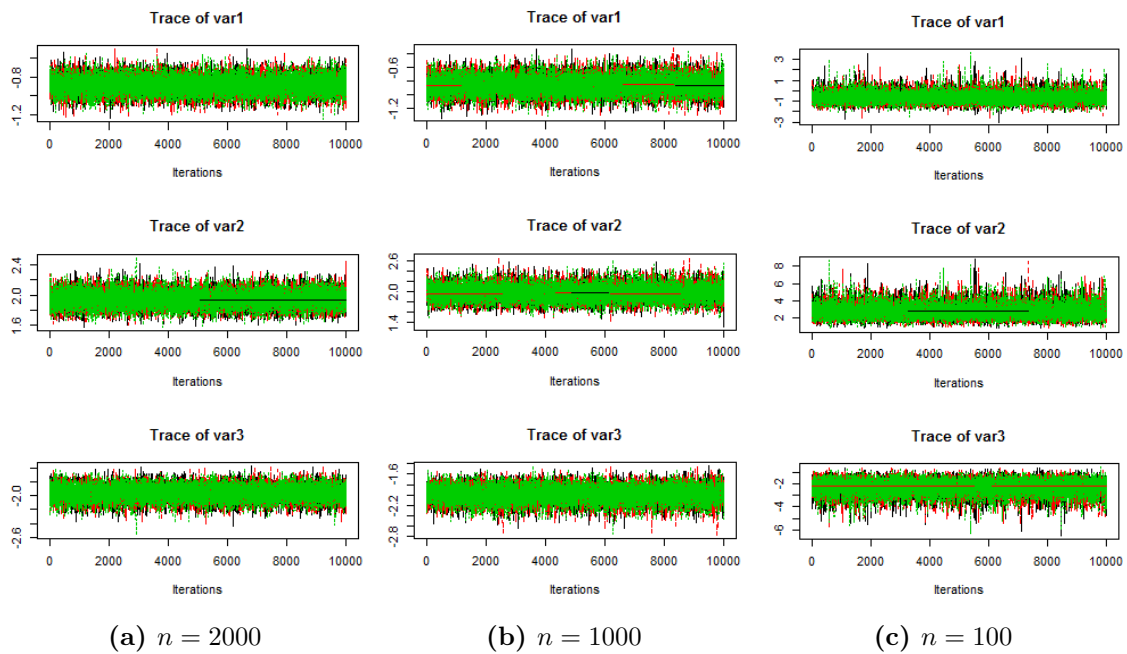


Figura .1: Cadenas simuladas del modelo *MaxBayes*(ejemplo simulación).

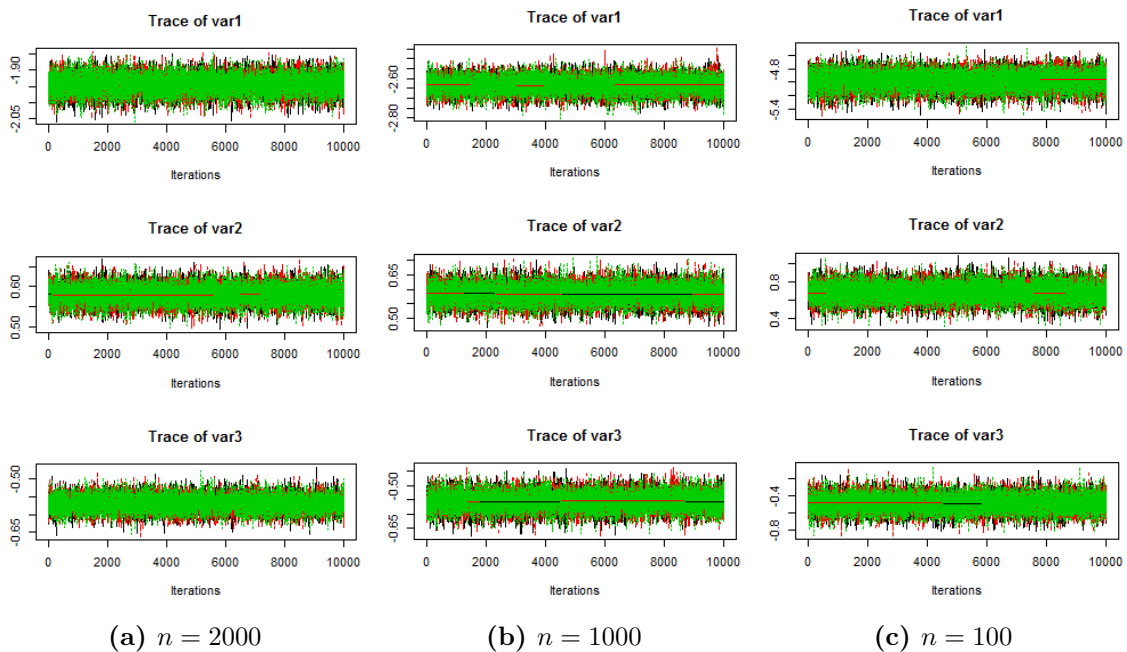


Figura .2: Cadenas simuladas del modelo *IPPBayes*(ejemplo simulación).

.1 Pruebas de convergencia para el modelo *MaxBayes* e *IPPBayes*

Tabla .1: Prueba de convergencia de Gelman y Rubin en el modelo *MaxBayes* (simulación).

Parámetro	<i>MaxBayes</i> $n = 2000$		<i>MaxBayes</i> $n = 1000$		<i>MaxBayes</i> $n = 100$	
	Estimador Puntual	Cuantil al 97.5 %	Estimador Puntual	Cuantil al 97.5 %	Estimador Puntual	Cuantil al 97.5 %
β_0	1.00	1.00	1.00	1.00	1.00	1.00
β_1	1.00	1.00	1.00	1.00	1.00	1.00
β_2	1.00	1.00	1.00	1.00	1.00	1.00
β_3	1.00	1.00	1.00	1.00	1.00	1.00
β_4	1.00	1.00	1.00	1.00	1.00	1.00
β_5	1.00	1.00	1.00	1.00	1.00	1.00
β_6	1.00	1.00	1.00	1.00	1.00	1.00
\hat{R}	1.00		1.00		1.00	

Tabla .2: Prueba de convergencia de Gelman y Rubin en el modelo *IPPBayes* (simulación).

Parámetro	<i>IPPBayes</i> $n = 2000$		<i>IPPBayes</i> $n = 1000$		<i>IPPBayes</i> $n = 100$	
	Estimador Puntual	Cuantil al 97.5 %	Estimador Puntual	Cuantil al 97.5 %	Estimador Puntual	Cuantil al 97.5 %
β_0	1.00	1.00	1.00	1.00	1.00	1.00
β_1	1.00	1.00	1.00	1.00	1.00	1.00
β_2	1.00	1.00	1.00	1.00	1.00	1.00
β_3	1.00	1.00	1.00	1.00	1.00	1.00
β_4	1.00	1.00	1.00	1.00	1.00	1.00
β_5	1.00	1.00	1.00	1.00	1.00	1.00
β_6	1.00	1.00	1.00	1.00	1.00	1.00
\hat{R}	1.00		1.00		1.00	

.1.2. Datos género *Dalea*

Las cadenas simuladas de la distribución *a posteriori* de los modelos *MaxBayes* e *IPPBayes* para el ejemplo con datos reales se muestran en las figuras (.3) y (.4). Ambas figuras muestran que la serie ha convergido a la distribución estacionaria. Por otra parte la Tabla (.3) resume los resultados de la prueba de Gelman y Rubin (1992) en ambos modelos. De acuerdo a dichos resultados, el factor de \hat{R} (estimador de reducción potencial de escala) en los todos los casos es igual a 1 lo cual evidencia que las cadenas simuladas en cada uno de los modelos se han traslapado y que por tanto, pertenecen a la distribución estacionaria.

.1 Pruebas de convergencia para el modelo *MaxBayes* e *IPPBayes*

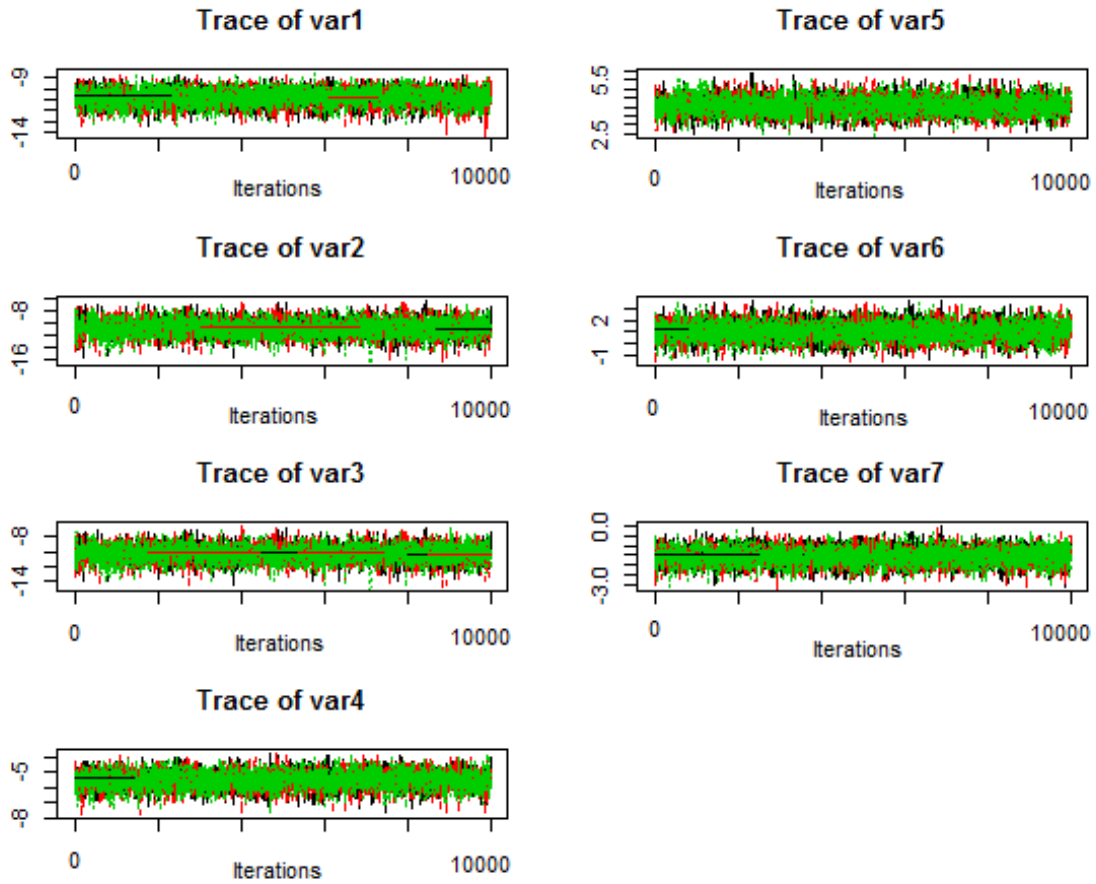


Figura .3: Cadenas simuladas de las distribuciones *a posteriori* del modelo *MaxBayes*.

.1 Pruebas de convergencia para el modelo *MaxBayes* e *IPPBayes*

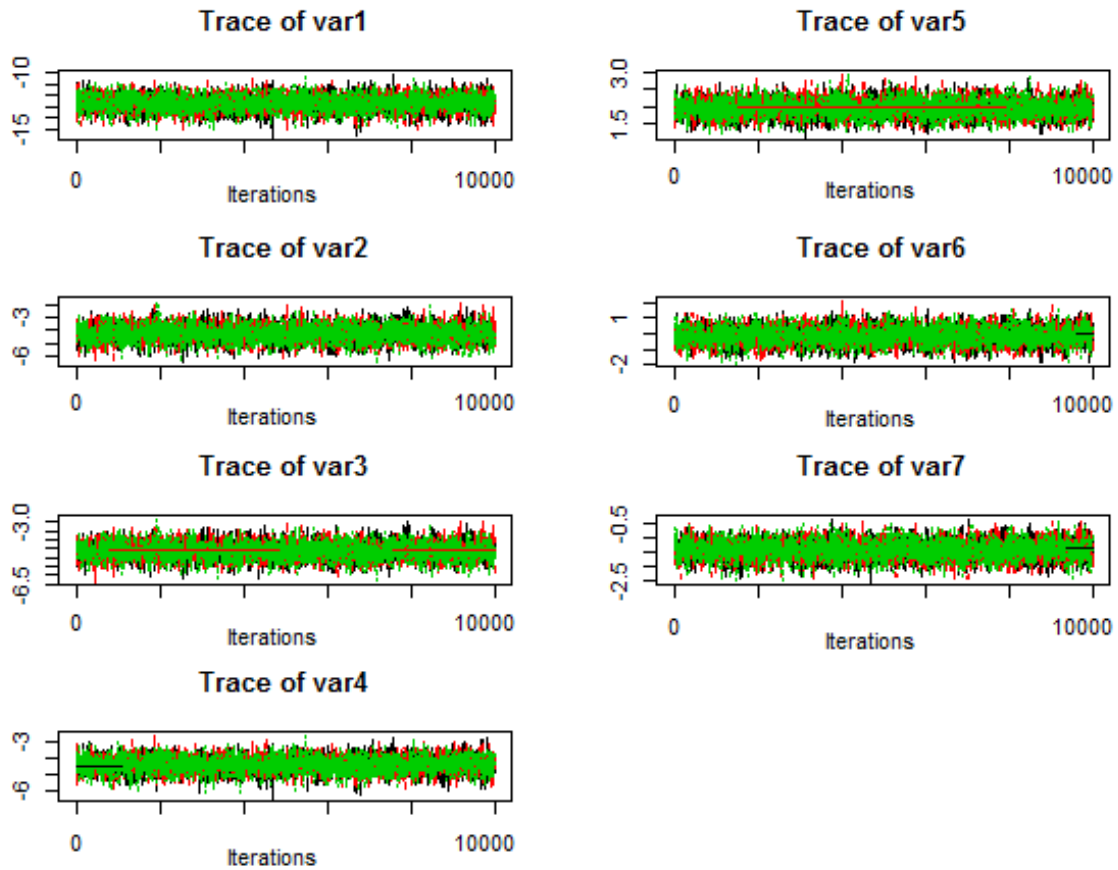


Figura .4: Cadenas simuladas de las distribuciones *a posteriori* del modelo *IPPBayes*.

Tabla .3: Prueba de convergencia de Gelman y Rubin en los modelos *MaxBayes* e *IPPBayes*.

Parámetro	<i>MaxBayes</i>		<i>IPPBayes</i>	
	Estimador Puntual	Cuantil al 97.5 %	Estimador Puntual	Cuantil al 97.5 %
β_0	1.00	1.00	1.00	1.00
β_1	1.00	1.00	1.00	1.00
β_2	1.00	1.00	1.00	1.00
β_3	1.00	1.00	1.00	1.00
β_4	1.00	1.00	1.00	1.00
β_5	1.00	1.00	1.00	1.00
β_6	1.00	1.00	1.00	1.00
\hat{R}	1.00		1.00	

.2. Código R

Las rutinas que se presentan en esta sección se desarrollaron en R-3.0.2 en el SO Ubuntu 12.04 LTS.

Código modelo *MaxBayes* - Simulación

```
#=====
# Código MaxBayes - Simulación
#=====
rm(list=ls(all=TRUE))
setwd("")
library(MHadaptive)
library(mvtnorm)
library(coda)
library(modeest)

set.seed(1984)
nobs <- 10000
trials <- rep(1,nobs)
X1 <- rnorm(n=nobs,0,1)
X2 <- rnorm(n=nobs,0,1)
X <- cbind(rep(1,nobs),X1,X2)

# Parámetros
beta.target <- matrix(c(-1,2,-2), ncol=1)
logit.psi <- X %*% beta.target
psi <- exp(logit.psi)/(1+exp(logit.psi))
hist(psi); mean(psi)

# Se simula la variable respuesta (1=0: Presencias-Ausencias)
Y <- rbinom(n=nobs,size=trials,prob=psi)

#== Dataset
Data <- data.frame(Y,trials,X1,X2)
str(Data)
par(mfrow=c(2,1))
```

.2 Código R

```
plot(Data$X1,logit.psi)
plot(Data$X2,logit.psi)
par(mfrow=c(1,1))

# Tamaño de muestra de presencias
#n.pres <- 100
#n.pres <- 1000
#n.pres <- 2000
data <- subset(Data, Data$Y!=0)
indices.m <- sample(1:nrow(data), size=n.pres)
muestra.p <- data[indices.m,]
str(muestra.p); str(data)

# MAXLIKE para comparación
# Matrices diseño
X <- model.matrix(~ X1+X2, muestra.p)
Z <- model.matrix(~ X1+X2, Data)

npars <- ncol(X)
parnames <- colnames(X)
starts <- rep(1, npars)
names(starts) <- parnames

#Funcion de Verosimilitud
log.likelihood <- function(pars){
  betas <- pars[1:npars]
  psix <- exp(X%*%betas)/(1+exp(X%*%betas))
  psiz <- exp(Z%*%betas)/(1+exp(Z%*%betas))
  return(sum(log(psix/sum(psiz))))
}

lik1 <- function(pars){
  betas <- pars[1:npars]
  psix <- exp(X%*%betas)/(1+exp(X%*%betas))
  psiz <- exp(Z%*%betas)/(1+exp(Z%*%betas))
  return(-1*sum(log(psix/sum(psiz))))
}
```

.2 Código R

```
resultado <- optim(par=starts, lik1, method=c("L-BFGS-B"),
                  lower=-25, upper=15, hessian=TRUE)

resultado$par                # Coeficientes estimados
resultado$value              # verosimilitud
resultado$convergence        # 0 indica convergencia
Fisher.matrix <- solve(resultado$hessian) # matriz de cov
se <- sqrt(diag(Fisher.matrix))          # errores estandar para coef.
#aic <- 2*npars+2*resultado$value; aic
#bic <- 2*log(nrow(X))+2*resultado$value

# Intervalos de confianza al 95%
IC.sup <- resultado$par+qnorm(.95,0,1)*se
IC.inf<- resultado$par-qnorm(.95,0,1)*se
ICs <- cbind(IC.inf,IC.sup); ICs

#=====
# Inferencia Bayesiana
#=====

prior <- function(pars) {
  betas<- pars[1:npars]
  abetas <- dnorm(betas, 0, 1000,log=TRUE)
  return(sum(abetas))
}
posterior <- function(pars){
  return (log.likelihood(pars)+prior(pars))
}

Semillas1 = resultado$par # Semillas
Sigma <- Fisher.matrix    # Semillas
MCMC.1 <- Metro_Hastings(li_func=posterior, pars=Semillas1, prop_sigma=Sigma,
                        par_names = c('beta0','beta1','beta2'), iterations = 100000,
                        burn_in = 50000, adapt_par = c(100, 20, 0.5, 0.75))

cadena.11 <- mcmc_thin(MCMC.1, thin=5)
cadena.1 <- mcmc(data=cadena.11$trace)
se.b <- sqrt(diag(cadena.11$prop_sigma))
```


.2 Código R

```
# Intervalos de Max Prob a posteriori al 95%
HPD.1 <- HPDinterval(cadena.1, prob = 0.95)

# Prediccion del MaxBayes
mode.beta0 <- mlv(cadena.1[,1], method = "mfv")$M
mode.beta1 <- mlv(cadena.1[,2], method = "mfv")$M
mode.beta2 <- mlv(cadena.1[,3], method = "mfv")$M
post.mode <- rbind(mode.beta0, mode.beta1, mode.beta2)

W <- model.matrix(~ X1+X2, Data)
prob.predichos <- plogis(W %*% post.mode); mean(prob.predichos)

# Gráficas
par(mfrow=c(3,3))
hist(cadena.1[,1], breaks=100, prob=TRUE, main="Posterior de 'beta0'",
     xlab="")
abline(v=HPD.1[1,], col="blue")
hist(cadena.1[,2], breaks=100, prob=TRUE, main="Posterior de 'beta 1'",
     xlab="")
abline(v=HPD.1[2,], col="blue")
hist(cadena.1[,3], breaks=100, prob=TRUE, main="Posterior de 'beta 2'",
     xlab="")
abline(v=HPD.1[3,], col="blue")
cadena1 <- as.matrix(cadena.1)
plot(cadena1[,1], type = "l", main = "Cadena de beta0")
plot(cadena1[,2], type = "l", main = "Cadena de beta 1")
plot(cadena1[,3], type = "l", main = "Cadena de beta 2")
hist(prob.predichos); hist(psi)
par(mfrow=c(1,1))

# Para pruebas de convergencia
Semillas2 <- rep(2,npars)
Semillas3 <- rep(-2,npars)
MCMC.2 <- Metro_Hastings(li_func=posterior, pars=Semillas2, prop_sigma=Sigma,
                        par_names = c('beta0','beta1','beta2'), iterations = 100000,
                        burn_in = 50000, adapt_par = c(100, 20, 0.5, 0.75))
```

.2 Código R

```
MCMC.3 <- Metro_Hastings(li_func=posterior, pars=Semillas3, prop_sigma=Sigma,
                        par_names = c('beta0','beta1','beta2'), iterations = 100000,
                        burn_in = 50000, adapt_par = c(100, 20, 0.5, 0.75))

cadena.22 <- mcmc_thin(MCMC.2, thin=5)
cadena.33 <- mcmc_thin(MCMC.3, thin=5)

cadena.2 <- mcmc(data=cadena.22$trace)
cadena.3 <- mcmc(data=cadena.33$trace)

# Prueba de Gelman and Rubin para convergencia de Cadenas
cadenas_mcmc <- mcmc.list(cadena.1,cadena.2,cadena.3)
plot(cadenas_mcmc)
gelman.diag(cadenas_mcmc)
gelman.plot(cadenas_mcmc)

#=====
# Fin del programa
#=====
```

Código modelo *IPPBayes* - Simulación

```
#=====
# Código IPPBayes - Simulación
#=====

rm(list=ls(all=TRUE))
setwd("")
library(MHadaptive)
library(mvtnorm)
library(coda)
library(modeest)
set.seed(1984)
nobs <- 10000
trials <- rep(1,nobs)
X1 <- rnorm(n=nobs,0,1)
```

.2 Código R

```
X2 <- rnorm(n=nobs,0,1)
X <- cbind(rep(1,nobs),X1,X2)

# Parámetros
beta.target <- matrix(c(-1,2,-2), ncol=1)
logit.psi <- X %*% beta.target
psi <- exp(logit.psi)/(1+exp(logit.psi))
hist(psi); mean(psi)

# Se simula la variable respuesta (1=0: Presencias-Ausencias)
Y <- rbinom(n=nobs,size=trials,prob=psi)

# Dataset
Data <- data.frame(Y,trials,X1,X2)
str(Data)
par(mfrow=c(2,1))
plot(Data$X1,logit.psi)
plot(Data$X2,logit.psi)
par(mfrow=c(1,1))

# Tamaño de muestra de presencias
#n.pres <- 100
#n.pres <- 1000
n.pres <- 2000
data <- subset(Data, Data$Y!=0)
indices.m <- sample(1:nrow(data), size=n.pres)
muestra.p <- data[indices.m,]
str(muestra.p); str(data)

X <- model.matrix(~ X1+X2, muestra.p)
Z <- model.matrix(~ X1+X2, Data)

npars <- ncol(X)
parnames <- colnames(X)
starts <- rep(1, npars)
names(starts) <- parnames
n <- length(X[,1])
n0 <- length(Z[,1])
```

.2 Código R

```
D <- length(Data[,1])

#Funcion de Verosimilitud
log.likelihood <- function(pars){
  betas <- pars[1:npars]
  verosimilitud <- sum(X%*%betas)-(D/n0)*sum(exp(Z%*%betas))
  return(verosimilitud)
}

#=====
# Inferencia Bayesiana
#=====

prior <- function(pars) {
  betas<- pars[1:npars]
  abetas <- dnorm(betas, 0, 1000,log=TRUE)
  return(sum(abetas))
}

posterior <- function(pars){
  return (log.likelihood(pars)+prior(pars))
}

Semillas <- rep(1,npars)
Sigma <- diag(c(0.1, 0.1, 0.1), npars)

MCMC.1 <- Metro_Hastings(li_func=posterior, pars=Semillas, prop_sigma=Sigma,
                        par_names = c('beta0','beta1','beta2'), iterations = 100000,
                        burn_in = 50000, adapt_par = c(100, 20, 0.5, 0.75))

cadena.11 <- mcmc_thin(MCMC.1, thin=5)
cadena.1 <- mcmc(data=cadena.11$trace)
se.b <- sqrt(diag(cadena.11$prop_sigma))
round(se.b, 2)

# Intervalos de Max Prob apoteriori al 95%
HPD.1 <- HPDinterval(cadena.1, prob = 0.95)
round(HPD.1, 2)
```

.2 Código R

```
# Prediccion del MaxBayes
mode.beta0 <-mlv(cadena.1[,1], method = "mfv")$M
mode.beta1 <-mlv(cadena.1[,2], method = "mfv")$M
mode.beta2 <-mlv(cadena.1[,3], method = "mfv")$M
post.mode <- rbind(mode.beta0, mode.beta1, mode.beta2)
round(post.mode,2)

# Prediccion del IPPBayes
W <- model.matrix(~ X1+X2, Data)
rate.ipp <- exp(W %*% post.mode); mean(rate.ipp)

# Gráficos
par(mfrow=c(3,3))
hist(cadena.1[,1], breaks=100, prob=TRUE, main="Posterior de 'alpha'",
      xlab="");
abline(v=HPD.1[1,], col="blue")
hist(cadena.1[,2], breaks=100, prob=TRUE, main="Posterior de 'beta 1'",
      xlab="");
abline(v=HPD.1[2,], col="blue")
hist(cadena.1[,3], breaks=100, prob=TRUE, main="Posterior de 'beta 2'",
      xlab="");
abline(v=HPD.1[3,], col="blue")
cadena1 <- as.matrix(cadena.1)
plot(cadena1[,1], type = "l",main = "Cadena de beta alpha")
abline(h = mode.beta0, col="red")
plot(cadena1[,2], type = "l",main = "Cadena de beta 1")
abline(h = mode.beta1, col="red")
plot(cadena1[,3], type = "l",main = "Cadena de beta 2")
abline(h = mode.beta2, col="red")
hist(rate.ipp); hist(psi)
par(mfrow=c(1,1))

# Para pruebas de convergencia
Semillas2 <- rep(2,npars)
Semillas3 <- rep(-2,npars)
MCMC.2 <- Metro_Hastings(li_func=posterior, pars=Semillas2, prop_sigma=Sigma,
                        par_names = c('beta0','beta1','beta2'), iterations = 100000,
                        burn_in = 50000, adapt_par = c(100, 20, 0.5, 0.75))
```

.2 Código R

```
MCMC.3 <- Metro_Hastings(li_func=posterior, pars=Semillas3, prop_sigma=Sigma,
                        par_names = c('beta0','beta1','beta2'), iterations = 100000,
                        burn_in = 50000, adapt_par = c(100, 20, 0.5, 0.75))
cadena.22 <- mcmc_thin(MCMC.2, thin=5)
cadena.33 <- mcmc_thin(MCMC.3, thin=5)
cadena.2 <- mcmc(data=cadena.22$trace)
cadena.3 <- mcmc(data=cadena.33$trace)

# Prueba de Gelman and Rubin para convergencia de Cadenas
cadenas_mcmc <- mcmc.list(cadena.1,cadena.2,cadena.3)
plot(cadenas_mcmc)
gelman.diag(cadenas_mcmc)
gelman.plot(cadenas_mcmc)

#=====
# Fin del programa
#=====
```

Código modelo *MaxBayes* - Ejemplo *Dalea*

```
#=====
# Código MaxBayes para datos Género Dalea
# Datos de Conabio para Reserva de la biosfera
# Cuicatlán-Teotitlán
#=====
rm(list=ls(all=TRUE))
library(MHadaptive)
library(mvtnorm)
library(coda)
library(modeest)
setwd("")
base.datos <- read.csv("DatosTesis_GeneroDalea.csv",header=TRUE)
data <- subset(base.datos[,1:9], base.datos$Altitud!=-9999)
datos <- cbind(data[,c(1:3)],scale(data[,4:9])) #estandarizamos
data.pres <- subset(datos, datos$y!=0) # datos de presencias
freq <- data.pres[,2]
data.pres.rep <- as.data.frame(data.pres[rep(1:nrow(data.pres), times=freq),])
```

.2 Código R

```
# Matrices diseño
X <- model.matrix(~ Altitud+Tmedia+Pp+RangoT+Lon+Lat, data.pres.rep)
Z <- model.matrix(~ Altitud+Tmedia+Pp+RangoT+Lon+Lat, datos)

npars <- ncol(X)
parnames <- colnames(X)
starts <- rep(1, npars)
names(starts) <- parnames

#Funcion de Verosimilitud
log.likelihood <- function(pars){
  betas <- pars[1:npars]
  psix <- exp(X%*%betas)/(1+exp(X%*%betas))
  psiz <- exp(Z%*%betas)/(1+exp(Z%*%betas))
  return(sum(log(psix/sum(psiz))))
}

lik1 <- function(pars){
  betas <- pars[1:npars]
  psix <- exp(X%*%betas)/(1+exp(X%*%betas))
  psiz <- exp(Z%*%betas)/(1+exp(Z%*%betas))
  return(-1*sum(log(psix/sum(psiz))))
}

resultado <- optim(par=starts, lik1, method=c("L-BFGS-B"),
                  lower=-25, upper=15, hessian=TRUE)

#resultado <- optim(par=starts, lik1, method=c("SANN"),
#                  hessian=TRUE)

resultado$par                # Coeficientes estimados
resultado$value              # verosimilitud
resultado$convergence        # 0 indica convergencia
Fisher.matrix <- solve(resultado$hessian) # matriz de cov
se <- sqrt(diag(Fisher.matrix))          # errores estandar para coef.
aic <- 2*npars+2*resultado$value; aic
#bic <- 2*log(nrow(X))+2*resultado$value
```

.2 Código R

```
#bic

# Intervalos de confianza al 95%
IC.sup <- resultado$par+qnorm(.95,0,1)*se
IC.inf<- resultado$par-qnorm(.95,0,1)*se
ICs <- cbind(IC.inf,IC.sup); ICs

#=====
# Inferencia Bayesiana
#=====
prior <- function(pars) {
  betas<- pars[1:npars]
  abetas <- dnorm(betas, 0, 1000,log=TRUE)
  return(sum(abetas))
}

# Funcion Posterior
posterior <- function(pars){
  return (log.likelihood(pars)+prior(pars))
}

# MetroHasting Adaptativo
Semillas.1 = resultado$par
Semillas.2 = rep(2, npars)
Semillas.3 = rep(-2, npars)
Sigma = diag(diag(Fisher.matrix))

MCMC.1 <- Metro_Hastings(li_func=posterior, pars=Semillas.1, prop_sigma=Sigma,
  par_names = c('beta1','beta2','beta3','beta4','beta5',
  'beta6','beta7'), iterations = 100000, burn_in = 50000,
  adapt_par = c(100, 20, 0.5, 0.75))
MCMC.2 <- Metro_Hastings(li_func=posterior, pars=Semillas.2, prop_sigma=Sigma,
  par_names = c('beta1','beta2','beta3','beta4','beta5',
  'beta6','beta7'), iterations = 100000, burn_in = 50000,
  adapt_par = c(100, 20, 0.5, 0.75))
MCMC.3 <- Metro_Hastings(li_func=posterior, pars=Semillas.3, prop_sigma=Sigma,
  par_names = c('beta1','beta2','beta3','beta4','beta5',
  'beta6','beta7'), iterations = 100000, burn_in = 50000,
```


.2 Código R

```
adapt_par = c(100, 20, 0.5, 0.75))

cadena.11 <- mcmc_thin(MCMC.1, thin=5)
cadena.1 <- mcmc(data=MCMC.1$trace)
se.b <- sqrt(diag(cadena.11$prop_sigma))
round(se.b,2)

cadena.22 <- mcmc_thin(MCMC.2, thin=5)
cadena.33 <- mcmc_thin(MCMC.3, thin=5)
cadena.2 <- mcmc(data=MCMC.2$trace)
cadena.3 <- mcmc(data=MCMC.3$trace)

# Highest Posterior Density intervals (Bayesiano)
HPD.1 <- HPDinterval(cadena.1, prob = 0.95)
round(HPD.1,2)

# Modas a posteriori
mode.beta0 <-mlv(cadena.1[,1], method = "mfv")$M
mode.beta1 <-mlv(cadena.1[,2], method = "mfv")$M
mode.beta2 <-mlv(cadena.1[,3], method = "mfv")$M
mode.beta3 <-mlv(cadena.1[,4], method = "mfv")$M
mode.beta4 <-mlv(cadena.1[,5], method = "mfv")$M
mode.beta5 <-mlv(cadena.1[,6], method = "mfv")$M
mode.beta6 <-mlv(cadena.1[,7], method = "mfv")$M
post.mode <- rbind(mode.beta0, mode.beta1, mode.beta2,
                  mode.beta3, mode.beta4, mode.beta5,
                  mode.beta6)

round(post.mode,2)

# Gráficos
par(mfrow = c(3,3))
hist(cadena.1[,1], breaks=50, prob=TRUE, main="Posterior de 'beta 0'",
     xlab=""); abline(v=HPD.1[1,], col="blue")
hist(cadena.1[,2], breaks=50, prob=TRUE, main="Posterior de 'beta 1'",
     xlab=""); abline(v=HPD.1[2,], col="blue")
hist(cadena.1[,3], breaks=50, prob=TRUE, main="Posterior de 'beta 2'",
     xlab=""); abline(v=HPD.1[3,], col="blue")
```

.2 Código R

```
hist(cadena.1[,4], breaks=50, prob=TRUE, main="Posterior de 'beta 3'",
      xlab=""); abline(v=HPD.1[4,], col="blue")
hist(cadena.1[,5], breaks=50, prob=TRUE, main="Posterior de 'beta 4'",
      xlab=""); abline(v=HPD.1[5,], col="blue")
hist(cadena.1[,6], breaks=50, prob=TRUE, main="Posterior de 'beta 5'",
      xlab=""); abline(v=HPD.1[6,], col="blue")
hist(cadena.1[,7], breaks=50, prob=TRUE, main="Posterior de 'beta 6'",
      xlab=""); abline(v=HPD.1[7,], col="blue")
par(mfrow = c(1,1))

# Cadenas simuladas
cadena1 <- as.matrix(cadena.1)
par(mfrow = c(3,3))
plot(cadena1[,1], type = "l",main = "Cadena de beta 0")
abline(h = mode.beta0, col="red")
plot(cadena1[,2], type = "l",main = "Cadena de beta 1")
abline(h = mode.beta1, col="red")
plot(cadena1[,3], type = "l",main = "Cadena de beta 2")
abline(h = mode.beta2, col="red")
plot(cadena1[,4], type = "l",main = "Cadena de beta 3")
abline(h = mode.beta3, col="red")
plot(cadena1[,5], type = "l",main = "Cadena de beta 4")
abline(h = mode.beta4, col="red")
plot(cadena1[,6], type = "l",main = "Cadena de beta 5")
abline(h = mode.beta5, col="red")
plot(cadena1[,7], type = "l",main = "Cadena de beta 6")
abline(h = mode.beta6, col="red")
par(mfrow=c(1,1))

# Prueba de Gelman and Rubin para convergencia de Cadenas
cadenas_mcmc <- mcmc.list(cadena.1,cadena.2,cadena.3)
plot(cadenas_mcmc)
gelman.diag(cadenas_mcmc)
gelman.plot(cadenas_mcmc)

# Prediccion
W <- model.matrix(~ Altitud+Tmedia+Pp+RangoT+Lon+Lat, datos)
prob.predichos <- plogis(W %*% post.mode)
```

.2 Código R

```
GIS <- as.data.frame(cbind(data[,8:9], data[,3],prob.predichos))
colnames(GIS) <- c('X','Y','y','Probabilidad')
write.table(GIS, file="Prediccion_Dalea_MaxBayes.csv",sep = ",")
```

```
#=====
# Fin del programa
#=====
```

Código modelo *IPPB*ayes - Ejemplo *Dalea*

```
#=====
# Código IPPBayes para datos Género Dalea
# Datos de Conabio para Reserva de la biosfera
# Cuicatlán-Teotitlán
#=====
rm(list=ls(all=TRUE))
library(MHadaptive)
library(mvtnorm)
library(coda)
library(modeest)
setwd("")
base.datos <- read.csv("DatosTesis_GeneroDalea.csv",header=TRUE)
data <- subset(base.datos[,1:9], base.datos$Altitud!=-9999)
datos <- cbind(data[,c(1:3)],scale(data[,4:9])) #estandarizamos
data.pres <- subset(datos, datos$y!=0) # datos de presencias
freq <- data.pres[,2]
data.pres.rep <- as.data.frame(data.pres[rep(1:nrow(data.pres), times=freq),])

# Matrices diseno
X <- model.matrix(~ Altitud+Tmedia+Pp+RangoT+Lon+Lat, data.pres.rep)
Z <- model.matrix(~ Altitud+Tmedia+Pp+RangoT+Lon+Lat, datos)

npars <- ncol(X)
parnames <- colnames(X)
starts <- rep(1, npars)
names(starts) <- parnames
n <- length(X[,1])
```

.2 Código R

```
n0 <- length(Z[,1])
D <- length(datos[,1])

#Funcion de Verosimilitud
log.likelihood <- function(pars){
  betas <- pars[1:npars]
  verosimilitud <- sum(X%*%betas)-(D/n0)*sum(exp(Z%*%betas))
  return(verosimilitud)
}

#=====
# Inferencia Bayesiana
#=====
prior <- function(pars) {
  betas<- pars[1:npars]
  abetas <- dnorm(betas, 0, 1000,log=TRUE)
  return(sum(abetas))
}

# Funcion Posterior
posterior <- function(pars){
  return (log.likelihood(pars)+prior(pars))
}

Semillas1 = rep(1, npars)
Semillas2 = rep(2, npars)
Semillas3 = rep(-2, npars)

Sigma <- diag(c(0.43, 0.34, 0.19, 0.19, 0.06, 0.26, 0.09), npars)

MCMC.1 <- Metro_Hastings(li_func=posterior, pars=Semillas1, prop_sigma=Sigma,
  par_names = c('beta0','beta1','beta2','beta3','beta4',
  'beta5','beta6'), iterations = 100000, burn_in = 50000,
  adapt_par = c(100, 20, 0.5, 0.75))

MCMC.2 <- Metro_Hastings(li_func=posterior, pars=Semillas2, prop_sigma=Sigma,
  par_names = c('beta0','beta1','beta2','beta3','beta4',
  'beta5','beta6'), iterations = 100000, burn_in = 50000,
```

.2 Código R

```
      adapt_par = c(100, 20, 0.5, 0.75))

MCMC.3 <- Metro_Hastings(li_func=posterior, pars=Semillas3, prop_sigma=Sigma,
      par_names = c('beta0','beta1','beta2','beta3','beta4',
      'beta5','beta6'), iterations = 100000, burn_in = 50000,
      adapt_par = c(100, 20, 0.5, 0.75))

cadena.11 <- mcmc_thin(MCMC.1, thin=5)
cadena.1 <- mcmc(data=MCMC.1$trace)
se.b <- sqrt(diag(cadena.11$prop_sigma))
round(se.b,2)
cadena.22 <- mcmc_thin(MCMC.2, thin=5)
cadena.33 <- mcmc_thin(MCMC.3, thin=5)
cadena.2 <- mcmc(data=MCMC.2$trace)
cadena.3 <- mcmc(data=MCMC.3$trace)

# Highest Posterior Density intervals (Bayesiano)
HPD.1 <- HPDinterval(cadena.1, prob = 0.95)

# Modas a posteriori
mode.beta0 <-mlv(cadena.1[,1], method = "mfv")$M
mode.beta1 <-mlv(cadena.1[,2], method = "mfv")$M
mode.beta2 <-mlv(cadena.1[,3], method = "mfv")$M
mode.beta3 <-mlv(cadena.1[,4], method = "mfv")$M
mode.beta4 <-mlv(cadena.1[,5], method = "mfv")$M
mode.beta5 <-mlv(cadena.1[,6], method = "mfv")$M
mode.beta6 <-mlv(cadena.1[,7], method = "mfv")$M
post.mode <- rbind(mode.beta0, mode.beta1, mode.beta2,
      mode.beta3, mode.beta4, mode.beta5,
      mode.beta6)

# Graficos
par(mfrow = c(2,4))
hist(cadena.1[,1], breaks=100, prob=TRUE, main="Posterior de 'alpha'",
      xlab=""); abline(v=HPD.1[1,], col="blue")
hist(cadena.1[,2], breaks=100, prob=TRUE, main="Posterior de 'beta 1'",
      xlab=""); abline(v=HPD.1[2,], col="blue")
```

.2 Código R

```
hist(cadena.1[,3], breaks=100, prob=TRUE, main="Posterior de 'beta 2'",
     xlab=""); abline(v=HPD.1[3,], col="blue")
hist(cadena.1[,4], breaks=100, prob=TRUE, main="Posterior de 'beta 3'",
     xlab=""); abline(v=HPD.1[4,], col="blue")
hist(cadena.1[,5], breaks=100, prob=TRUE, main="Posterior de 'beta 4'",
     xlab=""); abline(v=HPD.1[5,], col="blue")
hist(cadena.1[,6], breaks=100, prob=TRUE, main="Posterior de 'beta 5'",
     xlab=""); abline(v=HPD.1[6,], col="blue")
hist(cadena.1[,7], breaks=100, prob=TRUE, main="Posterior de 'beta 6'",
     xlab=""); abline(v=HPD.1[7,], col="blue")
par(mfrow = c(1,1))

# Cadenas simuladas
cadena1 <- as.matrix(cadena.1)
par(mfrow = c(2,4))
plot(cadena1[,1], type = "l", main = "Cadena de beta alpha")
abline(h = mean(cadena.1[,1]), col="red")
plot(cadena1[,2], type = "l", main = "Cadena de beta 1")
abline(h = mean(cadena.1[,2]), col="red")
plot(cadena1[,3], type = "l", main = "Cadena de beta 2")
abline(h = mean(cadena.1[,3]), col="red")
plot(cadena1[,4], type = "l", main = "Cadena de beta 3")
abline(h = mean(cadena.1[,4]), col="red")
plot(cadena1[,5], type = "l", main = "Cadena de beta 4")
abline(h = mean(cadena.1[,5]), col="red")
plot(cadena1[,6], type = "l", main = "Cadena de beta 5")
abline(h = mean(cadena.1[,6]), col="red")
plot(cadena1[,7], type = "l", main = "Cadena de beta 6")
abline(h = mean(cadena.1[,7]), col="red")
par(mfrow=c(1,1))

# Prueba de Gelman and Rubin para convergencia de Cadenas
cadenas_mcmc <- mcmc.list(cadena.1, cadena.2, cadena.3)
plot(cadenas_mcmc)
gelman.diag(cadenas_mcmc)
gelman.plot(cadenas_mcmc)

# Prediccion IPP
```

.2 Código R

```
W <- model.matrix(~ Altitud+Tmedia+Pp+RangoT+Lon+Lat, datos)
rate.ipp <- exp(W %*% post.mode)
GIS <- as.data.frame(cbind(data[,8:9],rate.ipp))
colnames(GIS) <- c('X','Y','rateOcurrence')
write.table(GIS, file="PrediccionIPPBayes_Dalea.csv",sep = ",")

#=====
# Fin del programa
#=====
```