ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS DE MÁXIMA ENTROPÍA Y SU RESOLUCIÓN NUMÉRICA MEDIANTE MÉTODOS DE SEGUNDO ORDEN CON APLICACIÓN EN LA PREDICCIÓN DE PRESENCIA DE ESPECIES EN ÁREAS NATURALES

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

PROYECTO DE INVESTIGACIÓN

ESTEFANNI MISHELLE CARPIO AMANCHA
estefanni.carpio@epn.edu.ec

DIRECTOR: PEDRO MARTÍN MERINO ROSERO
pedro.merino@epn.edu.ec

Quito, marzo, 2020

DECLARACIÓN

Yo, ESTEFANNI MISHELLE CARPIO AMANCHA declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Estefanni Mishelle Carpio Amancha

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Estefanni Mishelle Carpio Amancha, bajo mi supervisión.

Pedro Martín Merino Rosero
DIRECTOR

DEDICATORIA

A Yahveh, mi familia, John Cruz y amigos.

Agradecimientos

Agradezco a Dios, a Pedro Merino, mi madre Marlene Amancha, familia, John Cruz, amigos y al Centro de Modelización Matemática (MODEMAT).

Índice general

Índice de cuadros

Índice de figuras

Resumen

En este proyecto de investigación calculamos simulaciones precisas del modelo de máxima entropía, utilizando algoritmos de segundo orden aplicados en la predicción de la distribución de especies en áreas naturales. En particular usamos el método OESOM, que utiliza direcciones ortantes e información de segundo orden de la norma ℓ_1. Se considera una base de datos recolectada para una especie de oso perezoso (*Bradypus variegatus*) con la que se realizan comparaciones para el método usado por el software MaxEnt y el método OESOM. Este método de segundo orden puede realizar predicciones más precisas; además, reconoce eficientemente el conjunto de variables nulas. Se propone el estudio más profundo de los beneficios de los algoritmos de segundo orden aplicado a este tipo de problemas.

Abstract

In this project we calculate accurate simulations of the maximum entropy model, using second order algorithms applied in the prediction of the distribution of species in natural areas. In particular, we use the OESOM method, which uses orthant directions and second-order information of the ℓ_1 regularization. It is considered a database collected for a specie of sloth (*Bradypus variegatus*) with which comparisons are made for the method used by the MaxEnt software and the OESOM method. This second order method can make more accurate predictions. In addition, it efficiently recognizes the set of null variables. A deeper study of the benefits of second-order algorithms applied to these types of problems is proposed.

Capítulo 1

Introducción

El efecto de la acción humana y los cambios climáticos en el planeta durante los últimos años han cambiado las condiciones ambientales en el hábitat de las especies[14]. Por tanto, es de gran importancia el estudio de la conservación de los hábitats de las especies, tanto de plantas como de animales. En siglos anteriores, se han observado y registrado relaciones consistentes entre las distribuciones de especies y el entorno físico[35].

Los primeros escritos científicos de estas relaciones fueron en su mayoría cualitativos [19]. Los modelos matemáticos actuales se usan ampliamente para describir patrones y hacer predicciones. Estas técnicas matemáticas y computacionales permiten una gran diversidad de aplicaciones, con distintos rangos de éxito en sus resultados. Una aplicación es la de predicción de distribuciones de probabilidad, modelos de distribución.

Los modelos de distribución de especies (SDMs) pueden funcionar para la caracterización de las distribuciones naturales de las especies, en el caso que los datos recolectados estén bien diseñados y el modelo para analizar los predictores funcionalmente relevantes sea adecuado. Para esto, existen una gran gama de modelos desarrollados en las últimas décadas. Un modelo sobresaliente por su efectividad es el Modelo de Máxima Entropía (Maxent). En cuyas bases se ha desarrollado el software MaxEnt muy utilizado en el medio ecologista[36].

En lo que sigue del trabajo nos referiremos a Maxent como el método de Máxima Entropía, y a MaxEnt como el software.

Es importante tomar en cuenta que para modelar estadísticamente la distribución de especies son necesarios tres componentes[2]:

- Un modelo sobre la teoría ecológica
- Un modelo de recopilación de datos

- Un modelo sobre la teoría estadística

Con respecto al modelo ecológico existe un amplio conjunto de herramientas que son necesarias para el uso de los principios ecológicos que vinculan los datos espaciales con las respuestas fisiológicas y las limitaciones de las especies, lo cual es fundamental asignar al espacio geográfico para la inferencia en las restricciones del rango ambiental con el que una especie puede habitar[27]. Para las técnicas de recopilación de datos existen varios estudios útiles (ver [51]). En este trabajo el enfoque estará centrado en el modelo sobre la teoría estadística.

Esto, considerando la tolerancia de las especies a ciertas condiciones ambientales en la actualidad. Por el contrario, las aplicaciones que se ajustan a modelos para especies que no están sustancialmente en equilibrio con su entorno, y/o se extrapolan en el tiempo o el espacio, son mucho más desafiantes y los resultados son por lo general equívocos [13].

1.1. Modelos de Máxima Entropía

La teoría de la información nos proporciona un criterio constructivo para determinar distribuciones de probabilidad sobre datos obtenidos, y genera un tipo de inferencia estadística denominada Máxima Entropía[20]. El principio Maxent es un procedimiento rigurosamente probado que genera predicciones consistentes con el conocimiento previo[26]. Esta estimación es la menos sesgada posible con la información dada; es decir, es máximamente no comprometida con respecto a la información faltante.

La palabra "entropía" se refiere en el contexto ecológico a entropía de la información. La entropía de la información es una medida cuantitativa de la incertidumbre sobre el resultado de un evento de una distribución de probabilidad.[21]

La entropía de una distribución $p(n)$ discreta es

$$H(p) = -\sum_i p(i) \log p(i).$$

La idea principal en este método es que únicamente sean empleados los propios datos de entrenamiento (datos con los que se construye el modelo)[45]. Al usar un modelo de máxima entropía se asume la máxima incertidumbre, "si escogemos un modelo con menor entropía, estaremos añadiéndole restricciones que no están justificadas por la evidencia empírica"(Manning y SchÜtze, 1999)[30].

1.2. Predicción de especies en áreas naturales

Se han estudiado algunos métodos que modelan la distribución geográfica de las especies. Maxent es una de las mejores técnicas de modelado actualmente, por

su exactitud y algunas de sus propiedades. Las bases de datos generalmente no disponen datos de ausencia; en este trabajo se estudiará el método de máxima entropía que asume únicamente datos de presencia de la especie.

MaxEnt se basa en el Principio de Máxima Entropía, y de todos los modelos que se ajustan a nuestros datos de entrenamiento, selecciona el que tiene la mayor entropía, utilizando un método de regresión logística.

El método de regresión logística no siempre se puede usar directamente para estimar los parámetros de regresión. Para tratar el problema de la alta dimensionalidad y el "overfitting", una de las técnicas populares es el método de regularización. MaxEnt utiliza la penalización de red elástica

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1 \tag{1.1}$$

con la norma $||\beta||_2^2 = \sum_{j=1}^{p} |\beta_j|^2$ y la norma $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$, para $\beta \in \mathbb{R}^p$ y $\alpha \in [0,1]$.

El "overfitting" es un error que ocurre cuando el modelo predice muy bien con los datos que se han utilizado para entrenar el algoritmo, pero no predice bien al utilizar otros datos distintos a los que se han empleado para modelar. Generalmente, esto ocurre cuando el modelo es muy complejo y pierde generalización. La regularización en norma ℓ_1, al promover dispersión (anular coeficientes de la solución), disminuye el número de parámetros del modelo y ayuda a reducir el "overfitting"[5, 33].

1.3. Optimización numérica de modelos asociados a Maxent

Además de los problemas de la dimensionalidad y "overfitting", el uso de la norma ℓ_1 del vector de coeficientes es uno de los mecanismos más comunes para mejorar la dispersión empleado en la función de costo, por lo que la no diferenciabilidad de la norma ℓ_1 se convierte en un desafío en el diseño de algoritmos. La mayoría de los algoritmos consideran sólo información de primer orden. El paquete MaxEnt utiliza un método de Descenso Coordinado(CD)[52], el cual es un método de primer orden[24].

En este proyecto estudiamos la resolución de los problemas de regresión asociados al principio de máxima entropía, usando el método de segundo orden: OESOM[7].

El método OESOM, está diseñado para resolver problemas del tipo:

$$\min_x f(x) + \beta||x||_1$$

y se basa en la utilización de direcciones ortantes para construir direcciones de descenso y además utiliza información de segundo orden, tanto del término regular, como de la norma ℓ_1(en sentido débil). Esta última se realiza a través de una regularización de Huber[7].

Una de las principales características de este método consiste en una identificación rápida del conjunto activo ($x_i = 0$). Adicionalmente, para asegurar que las iteraciones pertenezcan al descenso debido, se utiliza una proyección adicional sobre los conjuntos activos. Una versión reducida de este método es equivalente a un algoritmo de Newton semi suave, bajo ciertos parámetros específicos. Es un método eficiente en comparación a otros algoritmos de última generación como se reporta en [7].

Algunos autores han investigado el uso de información de segundo orden relativa a la norma ℓ_1 (ver [15]). Sin embargo, una desventaja usual de los métodos de segundo orden es que cada iteración puede ser muy costosa computacionalmente[15].

Al usar información de segundo orden de la parte regular se ha obtenido algoritmos más rápidos en cuanto al número de iteraciones a un costo computacional más elevado; sin embargo, estos métodos no toman en cuenta información de segundo orden (generalizada) de los términos no diferenciables.

La contribución de este trabajo consiste en la aplicación del método de segundo orden OESOM específicamente para los problemas de máxima entropía, por lo que, la resolución del problema de optimización planteado por MaxEnt se beneficiaría del uso del método de segundo orden OESOM.

Capítulo 2

Máxima Entropía

Si se conoce algo sobre las probabilidades, se puede tomar en cuenta en la distribución de probabilidad que se usa para hacer predicciones sobre lo que va a suceder. Pero si no se sabe nada en absoluto, una distribución de probabilidad particularmente útil en este caso es aquella que maximiza la entropía, la de máxima incertidumbre, la de máxima ignorancia[26].

La idea fundamental de "maximizar la entropía" es elegir la distribución de probabilidad que resultaría de sumar todas las posibilidades permitidas por las leyes de naturaleza; como no se sabe nada, no se puede dejar de lado ninguna posibilidad.

Un ejemplo sencillo es el siguiente. Supongamos que se quiere predecir la calificación promedio para una clase de 100 estudiantes. Todo lo que se sabe son las reglas naturales: todos tienen una calificación, y es: A, B, C, D o F, no hay más. No se sabe nada de la capacidad de los estudiantes o la dificultad de la clase. Aplicando máxima entropía se asume que las calificaciones se pueden distribuir de todas las formas posibles: todas las combinaciones posibles igualmente probables. Por ejemplo, una posible distribución es 100 estudiantes que obtengan A y ninguna otra calificación. Otra sería que 50 obtengan C, 20 B, 20 D, 5 A y 5 F. Todas las combinaciones suman un conjunto de probabilidades que constituyen la distribución de probabilidad correspondiente a la máxima ignorancia sobre la clase, los estudiantes y sus calificaciones[41].

La entropía de una distribución $p(n)$ discreta (ver [47]) es

$$H(p) = -\sum_i P(i) \log P(i).$$

Ahora, supongamos que se tiene una distribución de probabilidad discreta $P(i|I)$, donde i representa alguna proposición (por ejemplo, una de las caras de

un dado) e I representa la información en la que se basa la distribución de probabilidad, luego

$$H = -\sum_{i}^{N} P(i|I) \log P(i|I)$$

es una medida de la cantidad de ignorancia en la distribución de probabilidad [26, pag 56].

En la teoría de información, el principio de máxima entropía (Maxent) establece que la distribución de probabilidad más apropiada entre todas aquellas que satisfacen las restricciones de nuestro conocimiento previo para modelar un conjunto dado de datos es la que tiene la mayor entropía. Así, establece que si uno tiene alguna información probada I, se puede asignar una distribución de probabilidad a una proposición i tal que $P(i|I)$ contiene solo información de I. Esta asignación se realiza maximizando H sujeto a las restricciones representadas por la información I.

Para demostrar su uso, supongamos que se tiene un dado de "honestidad" dudosa y se desea asignar una probabilidad a cada una de las seis caras. Si no se sabe nada sobre el dado, excepto que las probabilidades deben sumar 1, entonces

$$1 - \sum_{j=1}^{6} P(j|I) = 0$$

debe ser satisfecho. Si se cumple esta restricción, entonces se puede multiplicar por una constante λ (llamado multiplicador de Lagrange) y debido a que la restricción es cero se puede agregar a la entropía sin cambiar el valor de H:

$$H = -\sum_{j=1}^{6} P(j|I) \log P(j|I) + \lambda \left[1 - \sum_{j=1}^{6} P(j|I) \right]. \tag{2.1}$$

Se tiene que las probabilidades y λ son desconocidas. Para asignar las probabilidades, H está obligado a ser un máximo con respecto a las variaciones en las incógnitas: las probabilidades y λ. Puesto que H mide la cantidad de ignorancia en la distribución de probabilidad, exigir que H alcance su máximo es equivalente a hallar la distribución de probabilidad usando la menor cantidad de información. Y la distribución de probabilidad que mejor se ajusta a esta circunstancia es la que maximiza la entropía [26].

Este máximo se lo encuentra diferenciando H con respecto a las probabilidades y λ, e igualando a cero, ya que debe cumplir con la condición de punto crítico. En el ejemplo del dado, hay seis probabilidades desconocidas y un multiplicador de

Lagrange desconocido, pero cuando se toman las derivadas, habrá siete ecuaciones. Así, todas las incógnitas están completamente determinadas. Este sistema de ecuaciones se resuelve para obtener los valores de $P(j|I)$ y λ.

Cuando no se sabe nada, excepto que la distribución de probabilidad debe normalizarse (en el sentido que las distribuciones de probabilidad se ajusten a los valores), el principio de máxima entropía se reduce a maximizar la ecuación (2.1). Este es el principio de razón insuficente de Laplace[43].

Sin embargo, el principio de máxima entropía es mucho más general, pues permite asignar probabilidades que no son informativas al máximo, sin dejar de incorporar la información conocida. Si hubiera información disponible que indicara que el dado no era honesto, entonces esta información podría usarse en un cálculo de máxima entropía para obtener una distribución de probabilidad. El principio de máxima entropía representa una forma de asignar probabilidades basada únicamente en la información que uno realmente posee. En el ejemplo anterior esa información era que las probabilidades deberían sumar uno [26].

MaxEnt es el software más utilizado para la modelización de distribuciones de especies y se basa en el Principio de Máxima Entropía. De todos los modelos que se ajustan a nuestros datos de entrenamiento, selecciona el que tiene la mayor entropía.

2.1. Formulación

2.1.1. Glosario

- L: conjunto discreto de puntos de una grilla que cubre la región de interés. Por ejemplo, pixeles.
- L_1: subconjunto de L donde la especie está presente.
- y: variable dicotómica que representa la presencia de una especie, definida de la siguiente forma:

$$y = \begin{cases} 1 & \text{si la especie está presente} \\ 0 & \text{si la especie no está presente} \end{cases}$$

- z: vector de covariables de condiciones ambientales.
- $f(z)$: función de densidad de las covariables en L.
- $f_1(z)$: función de densidad de las covariables donde la especie está presente.
- $s(z)$: sesgo de la selección de la muestra.

2.1.2. Descripción del modelo

Sea L un conjunto discreto de puntos de una grilla que cubre la región de interés. Se tomará $L_1 \subset L$ un subconjunto donde la especie está presente. La distribución de las covariables se dará por una muestra finita, una colección de puntos de L con covariables asociadas. Como el terreno es una muestra aleatoria, la muestra de registros de presencia también es una muestra aleatoria de L_1[12].

Estamos interesados en predecir la presencia de cierta especie bajo ciertas condiciones ambientales, $\Pr(y = 1|z)$, con y la variable que indica si la especie está presente ($y = 1$) o si no ($y = 0$), y z el vector de covariables ambientales, las cuales veremos más adelante.

Puesto que no se encuentran con facilidad datos de ausencia, y si los hay no siempre se consideran fiables, MaxEnt utiliza datos sólo de presencia.

Con estas consideraciones, usando la fórmula de Bayes tendremos que la probabilidad de presencia de la especie bajo ciertas condiciones ambientales será:

$$Pr(y = 1|z) = \frac{f_1(z)Pr(y = 1)}{f(z)}, \tag{2.2}$$

donde $f_1(z)$ es la función de densidad de las covariables donde la especie se encuentra presente y $f(z)$ la función de densidad de las covariables en L. La única

cantidad que faltaría conocer es el término $Pr(y = 1)$, es decir, la prevalencia de la especie (proporción de sitios ocupados) en el terreno.

Una limitación fundamental de los datos de solo presencia es el sesgo en la selección de la muestra (por el cual algunas áreas en el terreno se muestrean más intensamente que otras), que tiene un efecto mucho más fuerte en modelos de sólo presencia que en modelos de presencia-ausencia [37].

Imaginemos $f_1(z)$ está contaminado por un sesgo de selección de muestra $s(z)$. Este sesgo ocurrirá más comúnmente en el espacio geográfico (por ejemplo, territorios cercanos a carreteras), pero también podría estar basado en el medio ambiente (por ejemplo, en barrancos mojados), sin embargo, independientemente se asignarán al espacio covariables.

Bajo el sesgo de muestreo, un modelo de solo presencia da una estimación de $f_1(z)s(z)$ en lugar de $f_1(z)$. Es decir, obtenemos un modelo que combina la distribución de especies con la distribución del esfuerzo de muestreo [42]. Por el contrario, para modelos de presencia-ausencia, el sesgo de selección de muestra afecta a ambos registros de presencia y ausencia, y en este tipo de modelos el efecto del sesgo se cancela [51].

MaxEnt se basa en una muestra insesgada, por lo que los esfuerzos en la recopilación de conjuntos complejos de registros de presencia (limpios de duplicados y errores) y lidiar con el sesgo es crítico [34]. Hay métodos implementados para la eliminación del sesgo en los datos, ver por ejemplo [37], [9] y [11]. La principal alternativa es proporcionar los datos de fondo con un sesgo similar a los registros de presencia; por ejemplo, usar las áreas encuestadas para otras especies en el mismo grupo biológico o usar grilla sesgada que indique el sesgo en los datos encuestados.

Covariables y características:

En el ámbito ecológico se suele llamar covariables a las variables independientes o "inputs" del modelo. Estas contienen factores relevantes acerca del medio ambiente en el lugar de estudio. Como, por ejemplo: estimaciones del clima, topografía, plantas, temperatura, etc. Puesto que, los resultados de la especie con respecto a estos suelen ser complejos, se usan funciones no lineales[2].

Las características (o "features") son transformaciones de las covariables. MaxEnt utiliza las siguientes clases de características: lineal, producto, cuadrática, bisagra, umbral y categórica.

1. Una variable continua f es en sí una "característica lineal"(L). Esta impone la restricción de que el valor medio de la covariable en el lugar donde se pronostica coincide aproximadamente con el valor medio donde se ha observado la ocurrencia.

2. El cuadrado de una variable continua f es una "característica cuadrática"(Q). Cuando se usa con la correspondiente característica lineal, impone la restricción de que la varianza de la covariable ambiental donde se pronostica debe estar cerca de su valor observado. Modela la tolerancia de la especie para variaciones en sus condiciones óptimas.

3. El producto de dos variables ambientales continuas f y g es una "característica producto"(P). Junto con las características lineales para f y g, impone la restricción de que la covarianza de esas dos variables debería estar cerca de su valor observado. Las características del producto por lo tanto incorporan interacciones entre variables predictoras.

4. Para una variable ambiental continua f, una "característica umbral"(T) está definida de la siguiente forma:

$$\text{threshold}_{f,h}(x) = \begin{cases} 0 & \text{si } f(x) < h \\ 1 & \text{caso contrario.} \end{cases}$$

   Impone la restricción de que la proporción de la covariable que tiene valores para f por encima del umbral h debe estar cerca de la proporción observada.

5. Para una variable ambiental categórica que toma valores $v_1, \cdots, v_k$, utilizamos k "características binarias"(C), donde la característica i-ésima es 1 donde la variable es igual a v_i y 0 en caso contrario. Al igual que con las funciones de umbral, estas características binarias limitan la proporción de las covariables en cada categoría para estar cerca de la proporción observada.

6. Una característica reciente que fue incorporada al modelo es la "característica de bisagra"(H). Es similar a la característica de umbral y está definida de la siguiente manera:

$$\text{hinge}_{f,h}(x) = \begin{cases} 0 & \text{si } f(x) < h \\ \frac{f(x)-h}{\text{máx}(f)-h} & \text{caso contrario.} \end{cases}$$

   En la terminología de splines, las características de umbral son funciones básicas de splines constantes por partes, mientras que las funciones de bisagra son funciones de splines lineales por partes. Las características de umbral y bisagra permiten a MaxEnt modelar una respuesta arbitraria de la especie en un entorno de variables de la que se deriva[38].

Generalmente, habrá más características que covariables. Las combinaciones comúnmente utilizadas son LC, LQC, HC, HQC, TC y HQPTC[36], [32] y [38]. Un subconjunto de estas características puede ser usado para simplificar la solución.

Por default, el programa restringe el modelo a características simples si se tiene pocas muestras disponibles, pues para cualquier método de modelado, una muestra pequeña proporciona información limitada para determinar las relaciones entre la especie y su entorno[3].

Las características lineales siempre se pueden usar, las cuadrática con al menos 10 muestras, las de bisagra con al menos 15, las características umbral y producto con al menos 80 muestras. En tales casos, es recomendable reducir el conjunto de candidatos a predictor usando conocimientos de biología de especies[28]. La característica de bisagra tiende a hacer que las características lineales y umbral sean redundantes. Existen varias recomendaciones a cerca de cómo usar estas características para formar modelos que sean sencillos de interpretar sin perder su capacidad de modelar interacciones complejas[11].

Anteriormente, se describe a MaxEnt como una estimación de la distribución en el espacio geográfico ($p(x) = P(x|y = 1)$, la distribución de probabilidad en el punto x de la grilla)[38]; sin embargo, ahora se analiza una caracterización que se centra en estimar la probabilidad de densidad en el espacio de las covariables ($Z(x)$ vector de covariables en la grilla), como se menciona en los resultados de la tesis doctoral de Gill Ward[48].

El estudio se centra en el espacio ambiental (Covariables). Así, maximizar la entropía de la distribución "en bruto", $p(x) = P(x|y = 1)$ distribuida en el espacio geográfico (grilla)[36] es equivalente a minimizar la entropía relativa de $f_1(z)$ con respecto a $f(z)$ [Apéndice A.1].

Llamaremos $h(z)$ al vector de las características y β al vector de coeficientes. Ahora, se tiene que la distribución de Gibbs q_β

$$q_\beta(z) = \frac{e^{\beta h(z)}}{Z_\beta}$$

(con Z_β una constante de normalización para que la suma de los $q_\beta(z)$ de 1), maximiza la entropía[38]. Por los resultados de [36] y [8], esto es equivalente a minimizar la entropía relativa resultante de la distribución de Gibbs, que es la familia exponencial:

$$f_1(z) = f(z)e^{\eta(z)}, \tag{2.3}$$

donde $\eta(z) = \alpha + \beta \cdot h(z)$, con α la constante de normalización que garantiza que la integral de $f_1(z)$ sea igual a 1 (ver [36]). Para esto, el objetivo del modelo de MaxEnt es $e^{\eta(z)}$. De (2.3) se aprecia que $e^{\eta(z)}$ estima el radio $f_1(z)/f(z)$ (ver [12]).

Tenemos de la ecuación (2.2) que, si conocemos la densidad condicional de las covariables en los sitios de presencia $f_1(z)$ y la densidad marginal de covariables en el área de estudio $f(z)$, entonces sólo necesitamos conocer la prevalencia $Pr(y = 1)$, para calcular la probabilidad condicional de ocurrencia. MaxEnt estima la

relación $f_1(z)/f(z)$, la cual se denomina salida "en bruto". Este es el indicador de MaxEnt para determinar qué características son relevantes y cuáles no. Además de la estimación relativa de si un lugar es más adecuado que otro para la presencia.

Puesto que la información requerida sobre la prevalencia no está disponible para calcular la probabilidad condicional de ocurrencia, una alternativa ha sido la implementación de la salida *logit*, definida de la siguiente forma[48]:

$$\text{logit}(Pr(y=1|z)) = \eta(z) \Longrightarrow Pr(y=1|z) = \frac{e^{\eta(z)}}{1+e^{\eta(z)}} \tag{2.4}$$

Así, se tomará $\eta(z) = log(f_1(z)/f(z))$ como un indicador *logit*, y se calibra el intercepto (α de la ecuación (2.3)) para que la probabilidad implícita de presencia en sitios con condiciones típicas para la especie sea un parámetro τ; MaxEnt toma $\tau = 0,5$.

τ mide cuan rara es una especie. Así, si una especie es invasiva tendrá un valor de τ mayor a 0,5 y si es una especie rara tendrá un valor menor a 0,5, lo cual puede ser regulado con mero conocimiento ecológico[12].

MaxEnt utiliza los datos de las covariables de los registros de ocurrencia y la muestra de datos de las covariables para estimar la relación $f_1(z)/f(z)$. Esto se puede hacer mediante una estimación de $f_1(z)$ que sea consistente con los datos de ocurrencia. MaxEnt estima la que está más cercana a $f(z)$, minimizando la distancia a $f(z)$. Puesto que $f(z)$ es un modelo *nulo* para $f_1(z)$ (la presencia no está determinada por las covariables, ver [18, 12]). Si no tenemos datos de ocurrencia, no se tendría razones para esperar que la especie prefiera un ambiente con condiciones particulares, por lo que no se podría predecir mejor que la especie ocupando lugares con condiciones ambientales proporcionales a su disponibilidad en el espacio (distribución uniforme sobre el terreno).

La distancia entre $f_1(z)$ y $f(z)$ es la entropía relativa de $f_1(z)$ con respecto a $f(z)$. El uso de datos de las covariables nos brinda información de la densidad de las covariables $f(z)$, y proporciona la base para la comparación con la densidad de las covariables donde la especie está presente $f_1(z)$.

Se imponen restricciones para que la solución refleje la información de presencia. Por ejemplo, si una covariable es la lluvia de verano, entonces hay restricciones que aseguran que la media de la precipitación de verano estará cerca de la media en los lugares que se ha observado presencia de la especie. La distribución de la especie se estima entonces minimizando la distancia entre $f_1(z)$ y $f(z)$, sujeto a restringir la precipitación media de verano estimada por $f_1(z)$(y las medias de otras covariables) para estar cerca de la media en los lugares de presencia.

Como hemos hablado anteriormente, las características son transformaciones de las covariables. Las restricciones se generalizan de ser restricciones de las medias de las covariables, a ser restricciones de las medias de las características.

Para encontrar una solución, MaxEnt necesita encontrar los coeficientes (β) que cumplan con las restricciones del modelo, cuidando que no haya sobreajuste y consecuentemente pérdida de generalización.

MaxEnt maneja el problema de encontrar esta solución con un límite para el error, o para la desviación máxima permitida para las medias de las características de la muestra. Primeramente, se reescalan las características para que estén en el rango $0-1$. Luego, la cota del error (λ_j de la ecuación (2.5)) se calcula para cada característica.

Este procedimiento reflejará la variación en los valores de la muestra para esa característica, ajustada por los cambios en los parámetros (preestablecidos) para cada clase de característica[38]. MaxEnt podría estimar la función error sólo de los datos, por ejemplo, usando validación cruzada, pero para simplificar el ajuste del modelo y puesto que los datos a menudo están sesgados, se usan parámetros específicos para cada clase de característica basado en un gran conjunto de datos globales[38].

Sin embargo, es posible que el ajuste pueda no funcionar bien para conjuntos de datos muy distintos, como por ejemplo si hubiese el conjunto de predictores es muy grande.

Este ajuste de los parámetros puede ser modificado. El ajuste previo también incluye restricciones para el conjunto de las clases de características para muestras pequeñas. En [12] se sugiere que los parámetros de regularización para las características h_j sean tomados como

$$\lambda_j = \lambda\sqrt{\frac{s^2[h_j]}{m}}. \tag{2.5}$$

La varianza de las características es $s^2[h_j]$ sobre los m sitios de presencia, y su clase de característica tiene un parámetro de ajuste λ. Conceptualmente, λ_j corresponde al ancho del intervalo de confianza, por lo tanto este toma la forma del error estándar (la expresi la raíz cuadrada) multiplicado por el parámetro λ de acuerdo con el nivel de confianza que se desee.

Los λ_j de la ecuación (2.5) inducen la regularización de la función de costo, es decir, promueven convexidad en el costo, por lo que es más regular. Esta regularización es una forma específica llamada ℓ_1-regularización [46]; que permite obtener soluciones dispersas (tienen muchos ceros, es decir, muchas características eliminadas). Se puede considerar como una forma de reducción de los coeficientes (β). La penalización dará valores que equilibren el ajuste y la complejidad, para que la predicción sea más exacta y generalizada.

Si el modelo es muy complejo tendrá una alta log-verosimilitud, pero no podrá generalizar bien. El objetivo de la regularización es hallar el equilibrio entre el ajuste del modelo y la complejidad del modelo (ecuación (2.6) lado izquierdo y

derecho). Maximizar el logaritmo de verosimilitud penalizado es equivalente a minimizar la entropía relativa, sujeto a las restricciones del límite del error[12].

$$\begin{cases} \text{máx}_{\alpha,\beta} \frac{1}{m} \sum_{i=1}^{m} \log(f(z_i)e^{\eta(z_i)}) - \sum_{j=1}^{n} \lambda_j |\beta_j| \\ \text{s.a. } \int_L f(z)e^{\eta(z)}dz = 1 \end{cases} \tag{2.6}$$

donde z es el vector de características para los puntos de ocurrencia i de los m sitios y para las $j = 1, \cdots, n$ características.

Cabe aclarar que al normalizar los datos de las covariables se reemplaza la restricción y se formúla el problema de optimización sin restricciones.

2.1.3. Penalización

En los últimos años se han desarrollado varios estudios a cerca de las regularizaciones de la función de costo. Uno de los más populares es la regularización ℓ_1 como podemos ver en la ecuación (2.6) ($||\beta||_1 = \sum_{j=1}^{n} |\beta_j|$), la cual es deseable cuando se espera que muchos coeficientes sean nulos, y un subconjunto pequeño sea importante y distinto de cero. Sin embargo, es indiferente a la correlación entre los predictores y tenderá a elegir uno y descartar el resto.

A diferencia de esta, la penalización en norma ℓ_2 (ridge), también muy utilizada, reduce los coeficientes de los predictores correlacionados con respecto a los otros, permitiéndoles tomar la fuerza prestada del uno al otro. Esto hace que la penalización con norma ℓ_2 sea ideal si hay muchos predictores. MaxEnt ha incorporado la siguiente regularización [22, pag 125],

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1 \tag{2.7}$$

si $\alpha = 0$ tenemos la penalización con norma ℓ_2 y si $\alpha = 1$ es la penalización con norma ℓ_1. Además de esto, tenemos la penalización "red elástica"que consiste en una combinación lineal de la penalización con norma ℓ_1 y ℓ_2, como se muestra en la ecuación (2.7). Tomando $\alpha = 1 - \varepsilon$ para algún $\varepsilon > 0$ pequeño.

Funciona de manera muy similar a la norma ℓ_1, pero elimina cualquier comportamiento inestable causado por correlaciones extremas. P_α crea una relación útil entre la penalización ℓ_1 y ℓ_2. A medida que α aumenta desde 0 a 1, para un λ dado, la escasez de la solución al problema de optimización (es decir, el número de coeficientes igual a cero) aumenta monotónicamente de 0 a la solución con la regularización de norma ℓ_1[16, 25].

2.2. Métodos numéricos

Para la resolución del problema planteado en la ecuación (2.6) el software MaxEnt utiliza una función llamada `maxnet`, la cual a su vez, utiliza un paquete llamado `glmnet`.

Glmnet ajusta un modelo lineal generalizado (GLM) a través de la máxima probabilidad penalizada. La regularización se calcula por la norma ℓ_1 o la red elástica de la penalización en una cuadrícula de valores para el parámetro de regularización λ. Glmnet resuelve el problema por defecto para 100 valores de λ.

Glmnet tiene la opción de trabajar con varias familias de redes elásticas. Nosotros nos centraremos en la que resolvería la ecuación (2.6). Para esto utilizaremos la regresión logística cuando la respuesta es categórica. Como hay dos resultados posibles, usaremos la distribución binomial, de lo contrario se usaría la multinomial. Denotemos "G" la predicción, tomando valores en $G \in \{1, 2\}$. El modelo de regresión logística representa las probabilidades condicionales de pertenecer a una clase a través de una función de predictores

$$\begin{aligned} Pr(G = 1|x) &= \frac{1}{1 + e^{-(\beta_o + \beta^T x)}}, \\ Pr(G = 2|x) &= \frac{1}{1 + e^{+(\beta_o + \beta^T x)}} \\ &= 1 - Pr(G = 1|x). \end{aligned} \tag{2.8}$$

Esto implica

$$\log \frac{Pr(G = 1|x)}{Pr(G = 2|x)} = \beta_0 + x^T \beta. \tag{2.9}$$

Se encuentra este modelo por máxima verosimilitud regularizada (binomial). Sea $p(x_i) = Pr(G = 1|x_i)$ la probabilidad de (2.8) para la observación i en un determinado valor de los parámetros (β_0, β), entonces la función de verosimilitud está dada por

$$\mathcal{L}(g_1, \cdots, g_N; \beta_0, \beta) = \prod_{i=1}^{N} p(x_i)^{I(g_i=1)} (1 - p(x_i))^{I(g_i=2)}, \tag{2.10}$$

con lo que tenemos la función de log verosimilitud

$$\ell(\beta_0, \beta) = \sum_{i=1}^{N} I(g_i = 1) \log p(x_i) + \sum_{i=1}^{N} I(g_i = 2)(1 - \log p(x_i)). \tag{2.11}$$

Ahora denotando $y_i = I(g_i = 1)$ en la ecuación (2.11), y de las ecuaciones (2.8) y (2.9), la función de verosimilitud penalizada quedaría de la siguiente forma

$$\ell(\beta_0, \beta) = \frac{1}{N} \sum_{i=1}^{N} \Big(y_i \cdot (\beta_0 + x_i^T \beta) - \log(1 + e^{(\beta_0 + x_i^T \beta)}) \Big) - \lambda P_\alpha(\beta) \tag{2.12}$$

con la penalización con parámetro α: $P_\alpha(\beta) = (1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1$ [16].

La ecuación (2.12) describe la función objetivo, la cual se desea maximizar, o lo que es equivalente

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} - \left[\frac{1}{N} \sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T \beta) - \log(1 + e^{(\beta_0 + x_i^T \beta)}) \right] + \lambda \left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1 \right] \tag{2.13}$$

La regresión logística no tiene un buen funcionamiento cuando $p > N$, es decir cuando el número de variables supera ampliamente el número de observaciones y muestra un comportamiento inestable incluso cuando N está cerca de p; la penalización de red elástica alivia parcialmente estos problemas[23].

El algoritmo que usa el paquete `glmnet` para resolver (2.13) utiliza una aproximación cuadrática a la función de log-verosimilitud, y luego coordina el descenso en el problema de mínimos cuadrados ponderados penalizados resultante. La función `glmnet` devuelve una secuencia de modelos para que los usuarios puedan elegir. Usualmente se selecciona uno de ellos. La validación cruzada es el método más simple y más utilizado para esta tarea (cv.glmnet)[23]. Para realizar la predicción se usa la función `predict`, utilizando el valor de λ que minimiza el error medio de validación cruzada.

Como Glmnet emplea el método de descenso coordinado cíclico, el cual utiliza información solo de primer orden de la regularización ℓ_1[52]. Nos interesa aplicar métodos de segundo orden para resolver (2.13)[4], para lo cual utilizaremos el algoritmo OESOM [7] que aprovecha la información de segundo orden valiosa del término con norma ℓ_1.

2.3. Método de segundo orden con direcciones ortantes

Orthantwise Enriched Second Order Method (OESOM) es un algoritmo de segundo orden, basado en direcciones ortantes, para resolver problemas de optimización relacionados con mejorar la dispersión con la norma ℓ_1.

La idea principal de este método consiste en incorporar información de segundo orden "oculta" en la norma ℓ_1. A pesar que la norma ℓ_1 no es diferenciable en un sentido clásico, tiene dos "derivadas" en un sentido generalizado. Utilizando

una regularización local de Huber, se puede extraer esa información para la actualización de la matriz de segundo orden, manteniendo el mismo tipo de direcciones ortantes de descenso.

En la siguiente imagen se ilustra geométricamente una comparación de la regularización de Huber con regularizaciones comunes (imágen tomada de [50]).

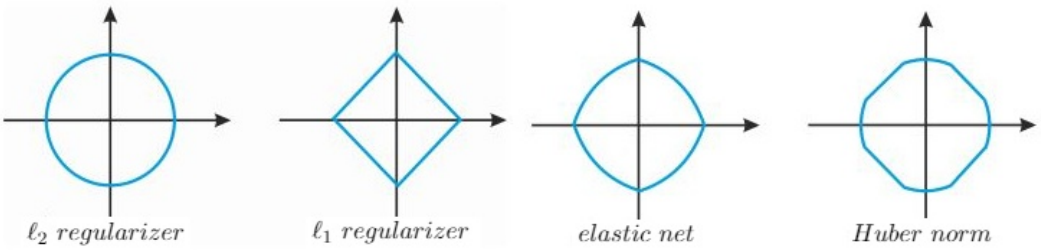


Figura 2.1: *Regularizaciones Comunes*

El algoritmo resultante permite una identificación rápida del conjunto activo y, por lo tanto, una rápida disminución de la función de costo[7].

Bajo una elección específica de los parámetros del algoritmo, se puede demostrar que este método es equivalente al método Newton semi suave. Por tanto, los resultados de convergencia importantes se heredan del marco de generalización de Newton, en particular convergencia superlineal local (ver [7]).

Aplicación del algoritmo OESOM al problema de máxima entropía

Sea $f : \mathbb{R}^m \to \mathbb{R}$ una función diferenciable y λ un número real positivo. Se está interesado en la solución del problema de optimización sin restricciones:

$$\min_{x \in \mathbb{R}^m} \varphi(x) := f(x) + \lambda \|x\|_1 \tag{2.14}$$

donde $\|\cdot\|_1$ corresponde a la norma ℓ_1 en $\mathbb{R}^m$: $\|x\|_1 := \sum_{j=1}^{m} |x_j|$. Este algoritmo también se puede utilizar en general para una función f regular, sin requerir convexidad. Se necesitan las siguientes condiciones:

- $f : \mathbb{R}^m \to \mathbb{R}$ es continuamente diferenciable con gradiente ∇f localmente Lipschitz continua.
- $\varphi = f + \lambda\|\cdot\|_1$ es coerciva.

Vamos a ver que la ecuación (2.13) cumple con las condiciones necesarias:

- Tenemos que para la ecuación (2.14)

$$f(\beta_0,\beta) = -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1+e^{(\beta_0+x_i^T\beta)})\right] + \lambda(1-\alpha)\|\beta\|_2^2/2$$

y nuestro término no diferenciable es $\lambda\alpha\|\beta\|_1$. Ahora, el gradiente de f es

$$\nabla f = \begin{bmatrix} -\left(\frac{1}{N}\sum_{i=1}^{N} y_i - \frac{e^{(\beta_0+x_i^T\beta)}}{1+e^{(\beta_0+x_i^T\beta)}}\right) \\ -\left(\frac{1}{N}\sum_{i=1}^{N} y_i - \frac{e^{(\beta_0+x_i^T\beta)}}{1+e^{(\beta_0+x_i^T\beta)}}\right) x_i^T + \lambda(1-\alpha)\beta_i^T \end{bmatrix}$$

y es continua. Además, tenemos que las segundas derivadas parciales son

$$\frac{\partial^2 f}{\partial\beta_0^2}(\beta_0,\beta) = \sum_{i=1}^{N} \frac{e^{(\beta_0+x_i^T\beta)}}{(1+e^{(\beta_0+x_i^T\beta)})^2},$$

$$\frac{\partial^2 f}{\partial\beta^2}(\beta_0,\beta) = \sum_{i=1}^{N} \frac{x_i x_i^T e^{(\beta_0+x_i^T\beta)}}{(1+e^{(\beta_0+x_i^T\beta)})^2} + \lambda(1-\alpha)I_p,$$

$$\frac{\partial^2 f}{\partial\beta_0\partial\beta}(\beta_0,\beta) = \sum_{i=1}^{N} \frac{x_i^T e^{(\beta_0+x_i^T\beta)}}{(1+e^{(\beta_0+x_i^T\beta)})^2},$$

y

$$\frac{\partial^2 f}{\partial\beta\partial\beta_0}(\beta_0,\beta) = \sum_{i=1}^{N} \frac{x_i^T e^{(\beta_0+x_i^T\beta)}}{(1+e^{(\beta_0+x_i^T\beta)})^2},$$

las cuales son continuas. Por tanto, ∇f es Lipschitz continua localmente [31].

- Para la coercividad,

$$\begin{aligned}\varphi(\beta_0,\beta) &= -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1+e^{(\beta_0+x_i^T\beta)})\right] + \\ &\quad + \lambda\left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right] \\ &\geq -\frac{1}{N}\sum_{i:y_i\neq 0}^{N} \cancel{y_i}^{1} \cdot (\beta_0 + x_i^T\beta) + \lambda(1-\alpha)||\beta||_2^2/2 + \underbrace{\lambda\alpha||\beta||_1}_{\geq 0}\end{aligned}$$

por Cauchy-Schwarz

$$\begin{aligned}&\geq -\frac{1}{N}\left(\sum_{i:y_i\neq 0}^{N} ||x_i||\right)||\beta||_2 + \lambda(1-\alpha)||\beta||_2^2/2 + c, \\ &\geq a_0||\beta||_2^2 - b_0||\beta||_2 \qquad \textit{para } \ a_0 > 0, \ \ b > 0.\end{aligned}$$

De donde $\varphi(\beta_0, \beta) \to \infty$ si $||\beta||_2 \to \infty$.

Por tanto, cumplimos con las hipótesis y podemos utilizar este algoritmo para la resolución del problema (2.13).

Se sabe que, dependiendo del tamaño del parámetro λ, la solución $\bar{x}$ tiende a tener más o menos entradas nulas. Usando resultados clásicos de análisis convexo las condiciones de optimalidad de primer orden para el problema (2.14) se pueden tomar como:

$$0 \in \nabla f(\bar{x}) + \lambda \partial \| \cdot \|_1(\bar{x}),$$

donde $\partial\phi(x)$ denota el subdiferencial de la función ϕ en x. Así podemos ver la última expresión de la siguiente manera:

$$\begin{array}{lll} 0 = \nabla_i f(\bar{x}) + \lambda & para & i \in \bar{\mathcal{P}} \\ 0 = \nabla_i f(\bar{x}) - \lambda & para & i \in \bar{\mathcal{N}} \\ 0 \in [\nabla_i f(\bar{x}) - \lambda, \nabla_i f(\bar{x}) + \lambda] & para & i \in \bar{\mathcal{A}} \end{array} \tag{2.15}$$

donde $\bar{\mathcal{P}}$, $\bar{\mathcal{N}}$ y $\bar{\mathcal{A}}$ están definidos como:

$$\bar{\mathcal{P}} = \{i : \bar{x}_i > 0\}, \quad \bar{\mathcal{N}} = \{i : \bar{x}_i < 0\}, \quad y \quad \bar{\mathcal{A}} = \{i : \bar{x}_i = 0\}. \tag{2.16}$$

Ahora, las direcciones ortantes asociadas a un vector $x \in \mathbb{R}^m$ se define como:

$$z_i(x) = \begin{cases} \text{sign}(x_i) & \text{si } x_i \neq 0, \\ 1 & \text{si } x_i = 0 \text{ y } \nabla_i f(x) < -\lambda, \\ -1 & \text{si } x_i = 0 \text{ y } \nabla_i f(x) > \lambda, \\ 0 & \text{caso contrario,} \end{cases} \tag{2.17}$$

para $i \in I := \{1, \cdots, m\}$, el conjunto de índices, y donde

$$\text{sign}(x_i) := \begin{cases} 1 & \text{si } x_i > 0, \\ 0 & \text{si } x_i = 0, \\ -1 & \text{si } x_i < 0. \end{cases}$$

Estas direcciones corresponden en realidad al elemento de subgradiente de norma mínima[44, pag 322]. Una caracterización de la definición del ortante de está dada por

$$\Omega := \{d : \text{sign}(d) = \text{sign}(z(x))\}. \tag{2.18}$$

De (2.17) podemos considerar el siguiente gradiente modificado de φ

$$\widetilde{\nabla}_i \varphi(x) = \begin{cases} \nabla_i f(x) + \lambda \text{sign}(x_i) & \text{si } x_i \neq 0, \\ \nabla_i f(x) + \lambda & \text{si } x_i = 0 \text{ y } \nabla_i f(x) < -\lambda, \\ \nabla_i f(x) - \lambda & \text{si } x_i = 0 \text{ y } \nabla_i f(x) > \lambda, \\ 0 & \text{caso contrario,} \end{cases} \tag{2.19}$$

que lo llamaremos pseudo gradiente. Cabe notar que el pseudo gradiente $\widetilde{\nabla}\varphi(x)$ pertenece a $\nabla f(x) + \lambda\partial\|\cdot\|_1(x)$.

Para asegurar que las soluciones aproximadas estén dentro de Ω, se requiere de un paso de proyección adicional. La proyección $\mathcal{P}$ correspondiente se define como:

$$\mathcal{P}(y)_i = \begin{cases} y_i & \text{si } \operatorname{sign}(y_i) = \operatorname{sign}(z_i(x)), \\ 0 & \text{caso contrario.} \end{cases} \tag{2.20}$$

Por otro lado, aunque la norma ℓ_1 no es diferenciable en un sentido clásico, es dos veces diferenciable en un sentido generalizado [6, pag 319]. La segunda derivada distributiva viene dada por la función delta de Dirac:

$$\delta(x) = \begin{cases} +\infty & \text{si } x = 0, \\ 0 & \text{caso contrario.} \end{cases}$$

Dado que, esta derivada débil es distinta de 0 en un solo punto, generalmente se descarta. En nuestro caso, sin embargo, estamos precisamente interesados en obtener una gran cantidad de ceros en el vector de solución y, por lo tanto, esta información puede ser valiosa. Tomando esta información de segundo orden, se considera una regularización de Huber de la norma ℓ_1 dada por $\|x\|_{1,\gamma} := \sum_{i=1}^{m} h_\gamma(x_i)$, donde

$$h_\gamma(x_i) = \begin{cases} \gamma\frac{x_i^2}{2} & \text{si } |x_i| \leq \frac{1}{\gamma}, \\ |x_i| - \frac{1}{2\gamma} & \text{si } |x_i| > \frac{1}{\gamma}, \end{cases} \tag{2.21}$$

para $\gamma > 0$. La primera derivada parcial está dada por

$$\nabla_i \|x\|_{1,\gamma} = \frac{\gamma x_i}{\max(1, \gamma|x_i|)}, \quad i = 1, \cdots, m, \tag{2.22}$$

es decir, el gradiente de la regularización de Huber es un vector cuyos componentes están dados por (2.22).

Dado que la primera derivada es una función semi suave [7, Section 5], también es posible calcular una segunda derivada generalizada. La Hessiana generalizada de la norma $\|\cdot\|_1$ es un matriz diagonal con entradas dadas por:

$$\Gamma_{ii} = \begin{cases} \gamma & \text{si } \gamma|x_i| \leq 1, \\ 0 & \text{otro caso.} \end{cases} \tag{2.23}$$

Al incluir esta matriz en el sistema de segundo orden, junto con las direcciones ortantes, obtenemos el siguiente sistema enriquecido:

$$\left(B^k + \lambda\Gamma^k\right) d^k = -\widetilde{\nabla}\varphi\left(x^k\right), \tag{2.24}$$

donde B^k representa la Hessiana de f (o una aproximación quasi-Newton simétrica definida positiva, como la matriz BFGS).

Los autores utilizan una variación de la matriz BFGS, sin embargo, esto no afecta a las principales propiedades de positividad de B^k[7], que se obtienen por construcción. Esta matriz está definida de la siguiente forma:

$$B^{k+1} = B^k - \frac{B^k \delta^k \delta^{k\top} B^k}{\delta^{k\top} B^k \delta^k} + \frac{y^k y^{k\top}}{y^{k\top} \delta^k}, \tag{2.25}$$

donde $y^k = \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right)$ y $\delta^k = x^{k+1} - x^k$. Además, se ha considerado la búsqueda lineal proyectada

$$\varphi\left[\mathcal{P}\left(x^k + s_k d^k\right)\right] \leq \varphi\left(x^k\right) + \widetilde{\nabla}\varphi\left(x^k\right)^T \left[\mathcal{P}\left(x^k + s_k d^k\right) - x^k\right], \tag{2.26}$$

utilizado un procedimiento de tipo "backtracking" para elegir s_k. También se asume que existe una constante $\hat{s} > 0$, tal que para la búsqueda lineal de los pasos s_k calculados con 2.26, para todo $k \in \mathbb{N}$, satisfacen

$$\hat{s} \leq s_k \leq 1. \tag{2.27}$$

El algoritmo está dado por los siguientes pasos:

Algorithm 1: Orthantwise Enriched Second Order Method (OESOM)

Se inicializa x^0 y B^0.
while *No se cumpla el criterio de parada* **do**
  Calcular la matriz Γ^k usando (2.23)
  Calcular el pseudo gradiente $\widetilde{\nabla}\varphi\left(x^k\right)$ usando (2.19)
  Calcular d^k resolviendo (2.24)
  Calcular

$$x^{k+1} = \mathcal{P}(x^k + s_k d^k),$$

  con el paso de búsqueda lineal calculado en (2.26).
  Actualizar la matriz B^k.
  $k \leftarrow k + 1$,
end
Result: x^k

La convergencia de este método se demuestra donde se puede consultar la evidencia numérica de la eficiencia del método [7].

Capítulo 3

Resolución Numérica

En esta sección vamos a evaluar el algoritmo OESOM aplicado al problema de máxima entropía. Se utilizará datos de una especie de oso perezoso *Bradypus*. Observaremos el desempeño del algoritmo de segundo orden y se comparará con el método usado por el paquete GLMNET. Para esto, utilizamos únicamente características lineales. Posteriormente, realizaremos predicciones y ejemplificaremos el uso de mapas de predicción.

3.1. Aplicación en la predicción de presencia de especies

La distribución que vamos a modelar es la del Oso Perezoso. La especie de muestra será *Bradypus variegatus*, Perezoso de tres dedos.

Figura 3.1: *Bradypus variegatus*

El conjunto de datos ambientales que será utilizado consiste en datos climáticos, de elevación, y una capa de vegetación potencial. Un conjunto que contiene datos ambientales en 116 puntos de ocurrencia de *Bradypus variegatus* (tomados de Anderson y Handley [1]), con 116 filas y 15 columnas. Las 15 columnas corresponden a las 14 covariables, que se tendrán en cuenta para la modelización, más los datos de presencia. Además, contiene 116 datos ambientales de América del Sur y Central, extraidos de una base de 1000 datos (tomados de `glmnet`). Hemos utilizados únicamente 116 datos ambientales pues estos daban una mejor predicción.

Se han tomado 2 datos para realizar una prueba de validación al modelo, el resto de datos para el entrenamiento del modelo. Vale la pena aclarar que esto no es una validación seria del modelo. Solo es una prueba, pues una validación adecuada debería realizarse con más datos. Nuestro interés es resolver el problema de optimización asociado al entrenamiento con la mayor cantidad de datos.

Las covariables bioclimáticas, topográficas y ecológicas que hemos utilizado son las que se muestran en el cuadro a continuación [36] y [29].

Covariable	**Descripción**
cld6190_ann	Nubosidad anual
dtr6190_ann	Rango de temperatura diurna anual
ecoreg	Vegetación potencial (Categórica)
frs6190_ann	Frecuencia de heladas anual
h_dem	Modelo digital de elevación(MDE)
pre6190_ann	Precipitación anual
pre6190_l1	Precipitación de Enero
pre6190_l4	Precipitación de Abril
pre6190_l7	Precipitación de Julio
pre6190_l10	Precipitación de Octubre
tmn6190_ann	Temperatura media anual
tmp6190_ann	Temperatura mínima anual
tmx6190_ann	Temperatura máxima anual
vap6190_ann	Presión anual de vapor

Cuadro 3.1: *Descripción de Covariables*

3.2. Evaluación numérica

En los siguientes experimentos numéricos evaluamos el desempeño del algoritmo OESOM para el entrenamiento del modelo de máxima entropía, usando los datos descritos en la sección anterior.

Por otro lado, tenemos 2 parámetros que hemos utilizado para el problema de entrenamiento: λ y α. Además, del parámetro γ utilizado en el algoritmo.

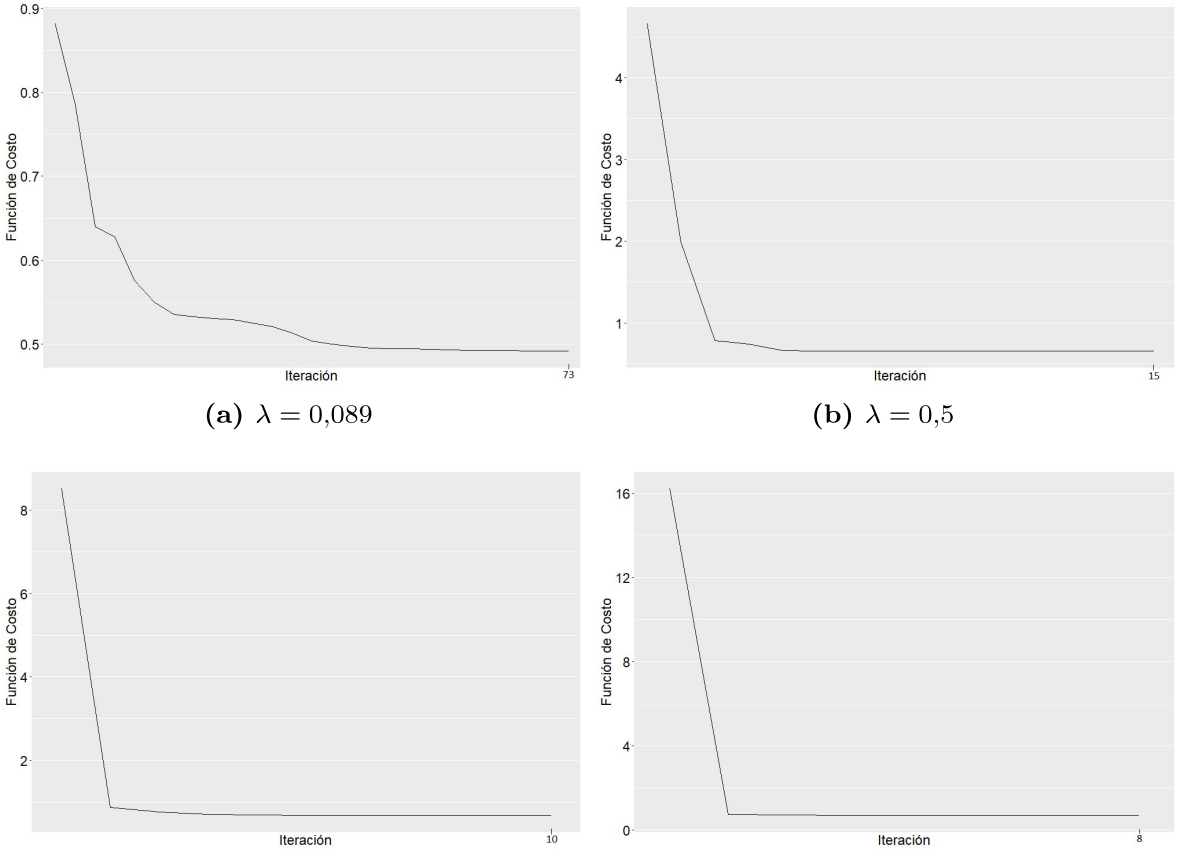
El valor de λ afecta directamente al peso que tendrá la penalización en el modelo, por lo que distintos valores de este parámetro nos darán distintos resultados en el número de variables con coeficiente igual a 0. En el Cuadro 3.2 podemos ver algunos resultados para ciertos valores de λ.

	Valores de λ ($\alpha = 0{,}1$ y $\gamma = 1500$)			
Covariable	0.089	0.5	1	2
β_0	0.774050	0.133656	0.146352	0.052635
β_1 : cld6190_ann	-0.055472	0.000000	0.000000	0.000000
β_2 : dtr6190_ann	-0.119796	0.000000	0.000000	0.000000
β_3 : ecoreg	-0.497554	0.000000	0.000000	0.000000
β_4 : frs6190_ann	-0.897440	-0.023705	0.000000	0.000000
β_5 : h_dem	-0.453303	-0.244501	-0.153064	-0.054872
β_6 : pre6190_ann	0.215895	0.000000	0.000000	0.000000
β_7 : pre6190_l1	-0.712446	0.000000	0.000000	0.000000
β_8 : pre6190_l10	1.389099	0.011780	0.000000	0.000000
β_9 : pre6190_l4	0.295433	0.000000	0.000000	0.000000
β_{10} : pre6190_l7	0.657763	0.000000	0.000000	0.000000
β_{11} : tmn6190_ann	0.767633	0.065740	0.283860	0.000000
β_{12} : tmp6190_ann	0.000000	0.000000	0.000000	0.000000
β_{13} : tmx6190_ann	-0.84872548	0.000000	0.000000	0.000000
β_{14} : vap6190_ann	0.332569	0.057209	0.102856	0.000000
# Iteraciones	73	15	10	8
$\|x^{k+1} - x^k\|$	4.7798e-05	3.4617e-05	8.1302e-05	1.2677e-05
Costo	0.491420	0.653410	0.675325	0.689493

Cuadro 3.2: *Comparación para distintos valores de λ*

Así, mientras el valor de λ aumenta, también lo hace el número de coeficientes nulos para las covariables. Además, a medida que crece el valor de λ el número de iteraciones necesarias, para que el criterio de parada ($\|x^{k+1} - x^k\| < 1e-04$) se cumpla, disminuye.

En la Figura 3.2 podemos ver el comportamiento del algoritmo en la función objetivo con los valores de λ del Cuadro 3.2.



(a) $\lambda = 0{,}089$ **(b)** $\lambda = 0{,}5$

(c) $\lambda = 1$ **(d)** $\lambda = 2$

Figura 3.2: *Comparación para distintos valores de* λ

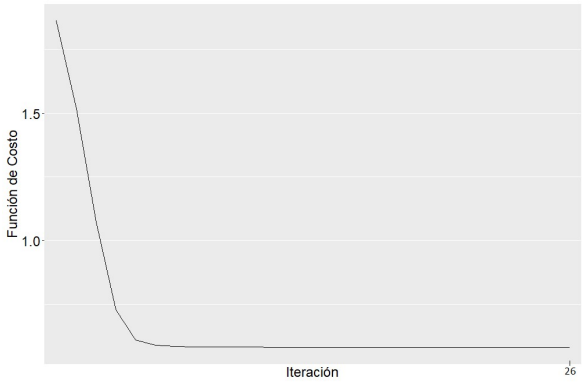
Tenemos, que cuando λ es cercano a 0 el algoritmo se demora más en encontrar la vecindad de solución, pues el OESOM en un método localmente convergente. Un valor muy pequeño de λ afecta a la coercividad del modelo, por lo que para valores muy cercanos a 0, el algoritmo podría no converger a una solución.

Para el parámetro α tenemos que mientras este sea pequeño (cercano a 0) la información obtenida por la norma ℓ_2 tiene mayor peso, y menor el peso de la información de la norma ℓ_1(red elástica); y viceversa para un valor de α grande (cercano a 1). En el cuadro siguiente, podemos ver algunos resultados para distintos valores de este parámetro.

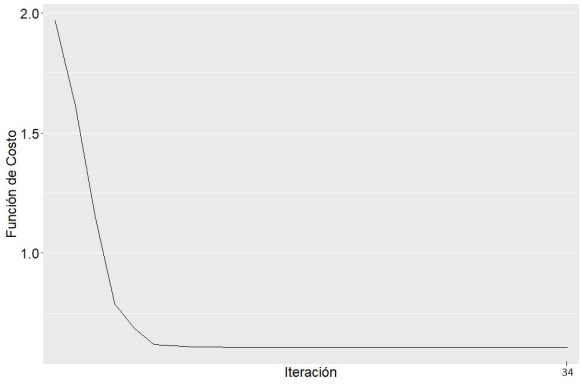
	Valores de α ($\lambda = 0{,}15$ y $\gamma = 1500$)			
Covariable	0.0005	0.1	0.25	0.5
β_0	0.043433	0.096419	0.220750	0.361124
β_1 : cld6190_ann	-0.077422	0.000000	0.000000	0.000000
β_2 : dtr6190_ann	-0.156399	-0.097733	0.000000	0.000000
β_3 : ecoreg	-0.098104	-0.018143	0.000000	0.000000
β_4 : frs6190_ann	-0.207875	-0.167391	-0.079889	0.000000
β_5 : h_dem	-0.335129	-0.338367	-0.358467	-0.382774
β_6 : pre6190_ann	0.092335	0.006209	0.000000	0.000000
β_7 : pre6190_l1	-0.145084	-0.063930	0.000000	0.000000
β_8 : pre6190_l10	0.270722	0.221797	0.117368	0.000000
β_9 : pre6190_l4	0.027778	0.000000	0.000000	0.000000
β_{10} : pre6190_l7	0.228540	0.174378	0.056456	0.000000
β_{11} : tmn6190_ann	0.252169	0.225381	0.164943	0.000000
β_{12} : tmp6190_ann	0.069340	0.000000	0.000000	0.000000
β_{13} : tmx6190_ann	-0.110114	-0.014564	0.000000	0.000000
β_{14} : vap6190_ann	0.189203	0.147785	0.065703	0.000000
# Iteraciones	26	34	28	13
$\|x^{k+1} - x^k\|$	9.8690e-05	4.6159e-05	3.8035e-05	1.1324e-05
Costo	0.580328	0.605285	0.628004	0.644760

Cuadro 3.3: *Comparación para distintos valores de α*

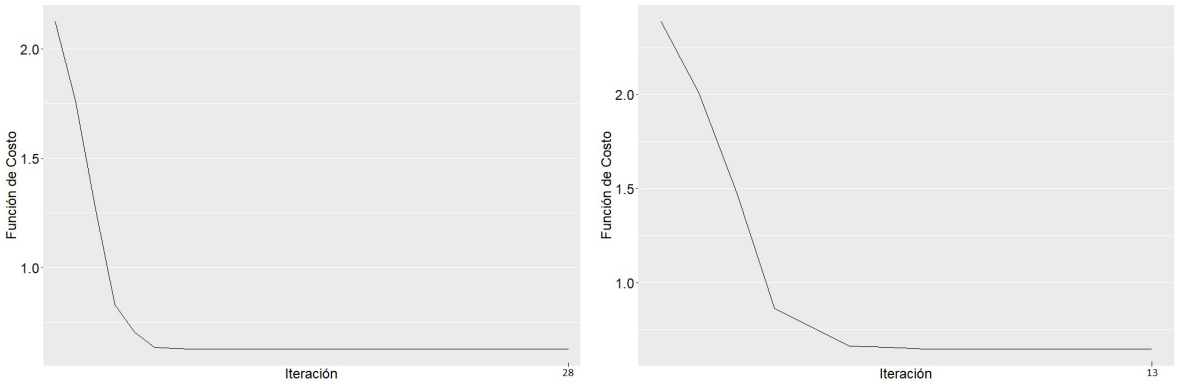
En la siguiente figura podemos observar el comportamiento de la función objetivo para valores de $\lambda = 0{,}15$, $\gamma = 1500$ y distintos valores de α del Cuadro 3.3.



(a) $\alpha = 0{,}0005$



(b) $\alpha = 0{,}1$



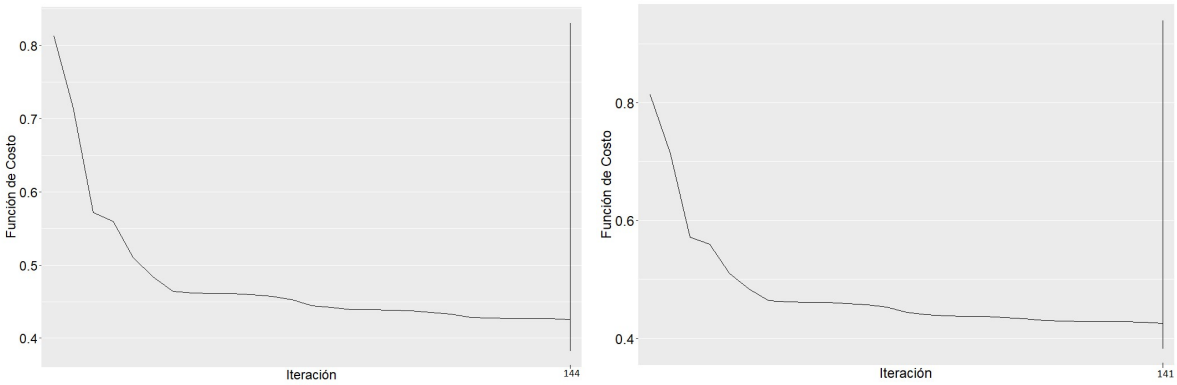
(c) $\alpha = 0{,}25$ **(d)** $\alpha = 0{,}5$

Figura 3.2: *Comparación para distintos valores de* α

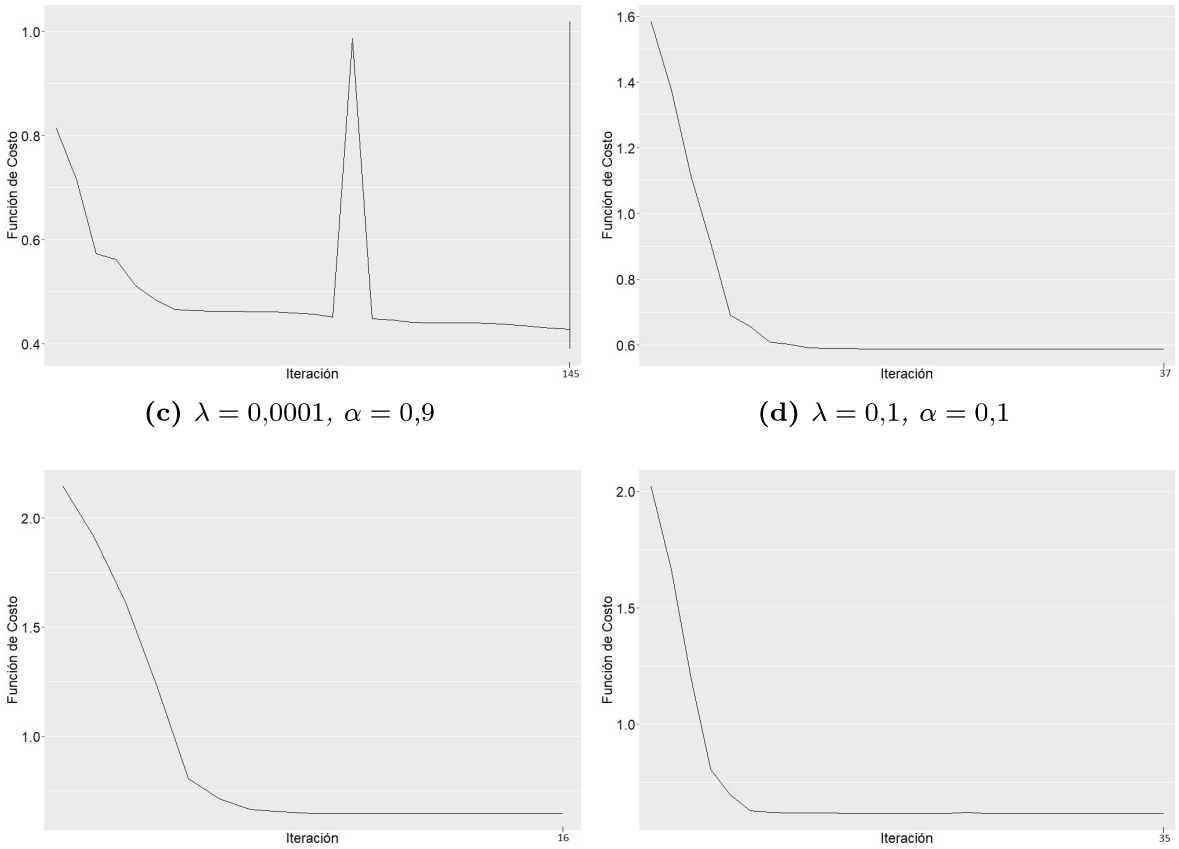
Mientras el valor de α aumenta también lo hace el número de coeficientes nulos.

Entonces, se puede "jugar" con estos parámetros para mantener la coercividad de la función y encontrar diferentes soluciones. Las cuáles serán evaluadas estadísticamente y seleccionadas de acuerdo a las que tienen mayor capacidad de predicción.

A continuación, en la Figura 3.3, se puede visualizar la función de costo para algunas combinaciones de los parámetros λ y α (con $\gamma = 1500$ fijo).



(a) $\lambda = 1e-05$, $\alpha = 1e-05$ **(b)** $\lambda = 1e-05$, $\alpha = 0{,}1$



(c) $\lambda = 0{,}0001$, $\alpha = 0{,}9$ **(d)** $\lambda = 0{,}1$, $\alpha = 0{,}1$

(e) $\lambda = 0{,}1$, $\alpha = 0{,}3$ **(f)** $\lambda = 0{,}15$, $\alpha = 0{,}15$

Figura 3.3: *Comparación para distintos valores de λ y α*

Como se puede ver en las primeras imágenes de la Figura 3.3, con λ muy cercano a 0, el valor de α grande no es suficiente para que se pueda encontrar una solución óptima en pocas iteraciones.

Consideremos ahora el parámetro γ, correspondiente a la información generalizada de segundo orden que viene de la regularización de Huber de la ecuación (2.21). Si γ es grande la aproximación de la información de segundo orden asociada a la norma ℓ_1 será más precisa.

En el siguiente cuadro se muestran los distintos resultados para varios valores del mismo.

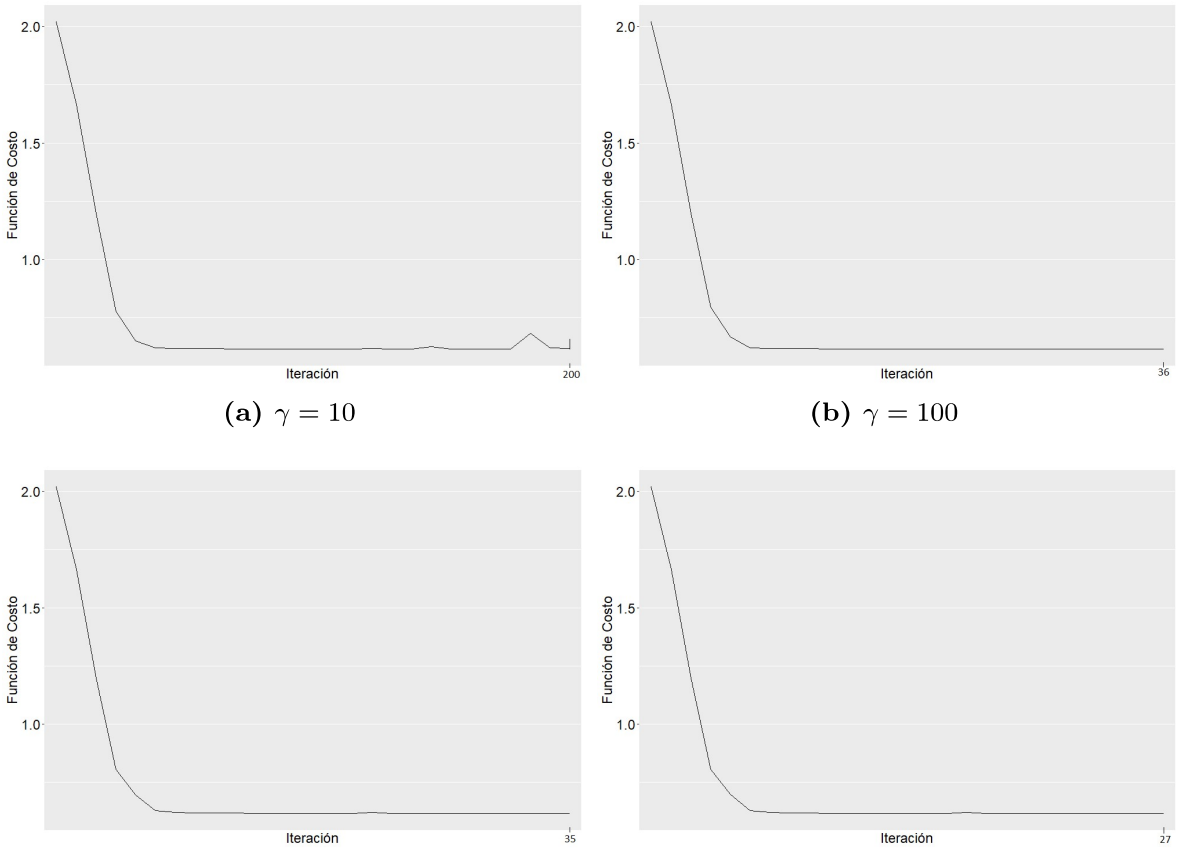
	Valores de γ ($\lambda = 0{,}15$ y $\alpha = 0{,}15$)			
Covariable	10	100	1500	1500000
β_0	0.857007	0.141486	0.141456	0.151329
β_1 : cld6190_ann	0.000000	0.000000	0.000000	0.000000
β_2 : dtr6190_ann	-0.043059	-0.059832	-0.059866	-0.059619
β_3 : ecoreg	0.000000	0.000000	0.000000	0.000000
β_4 : frs6190_ann	-0.105592	-0.140874	-0.140857	-0.140905
β_5 : h_dem	-0.364247	-0.341949	-0.341945	-0.342000
β_6 : pre6190_ann	0.000000	0.000000	0.000000	0.000000
β_7 : pre6190_l1	-0.003624	-0.018198	-0.018200	0.000000
β_8 : pre6190_l10	0.201245	0.191943	0.191967	0.192767
β_9 : pre6190_l4	0.000000	0.000000	0.000000	0.000000
β_{10} : pre6190_l7	0.150854	0.140833	0.140786	0.141649
β_{11} : tmn6190_ann	0.228609	0.207383	0.207371	0.207710
β_{12} : tmp6190_ann	0.000156	0.000000	0.000000	0.000000
β_{13} : tmx6190_ann	0.000000	0.000000	0.000000	0.000000
β_{14} : vap6190_ann	0.097336	0.122823	0.122847	0.122997
# Iteraciones	200*	36	35	27
$\|x^{k+1} - x^k\|$	0.981522	9.7398e-05	8.0529e-05	4.4684e-05
Costo	0.656793	0.614267	0.614267	0.614290

(200* no cumplió el criterio de parada hasta la iteración 200)

Cuadro 3.4: *Comparación para distintos valores de γ*

Nótese en el Cuadro 3.4 en el costo para $\gamma = 1500000$ que un γ muy grande puede afectar el desempeño numérico debido a errores de punto flotante. Este parámetro no influye demasiado en el número de variables que tendrán coeficientes nulos.

La información de segundo orden y las proyecciones no tiene un buen funcionamiento con un valor pequeño de γ pues la información de segundo orden para valores pequeños de γ puede confundir componentes de la solución con valores pequeños con soluciones nulas. Además, un γ pequeño no aproxima bien el segundo orden asociado a la norma ℓ_1. En la siguiente figura podemos observar, el comportamiento de la función objetivo, para los valores de γ del Cuadro 3.4, con $\lambda = 0{,}15$ y $\alpha = 0{,}15$.



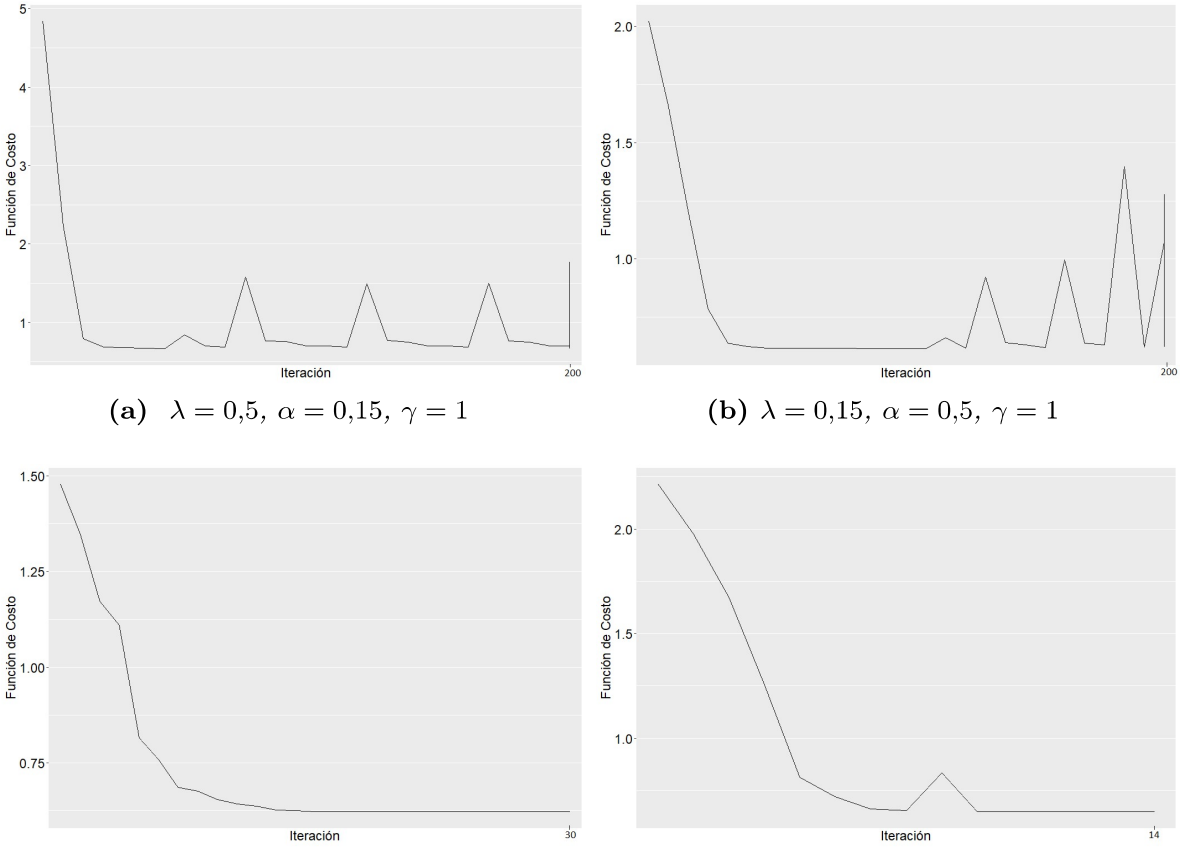
(a) $\gamma = 10$ **(b)** $\gamma = 100$

(c) $\gamma = 1500$ **(d)** $\gamma = 1500000$

Figura 3.4: *Comparación para distintos valores de* γ

En la Figura 3.4 **(a)** el criterio de parada no se cumple, pues un valor de γ pequeño afecta a la convergencia. Por lo que nos interesan valores de γ grandes.

A continuación, se muestra el comportamiento de la función para algunos valores de λ, α y γ.



(a) $\lambda = 0{,}5$, $\alpha = 0{,}15$, $\gamma = 1$

(b) $\lambda = 0{,}15$, $\alpha = 0{,}5$, $\gamma = 1$

(c) $\lambda = 0{,}05$, $\alpha = 0{,}9$, $\gamma = 100$

(d) $\lambda = 0{,}1$, $\alpha = 0{,}9999$, $\gamma = 1500000$

Figura 3.5: *Comparación para distintos valores de λ, α y γ*

3.3. Comparación

Por lo visto en la sección anterior tomaremos los parámetros $\lambda = 0{,}15$, $\alpha = 0{,}15$ y $\gamma = 1500$ para la comparación de los resultados de Glmnet y Oesom.

Así, se han obtenido los valores para los coeficientes de cada una de las covariables, que se muestran en el Cuadro 3.5.

Covariable	**Glmnet** ($\lambda = 0{,}15$, $\alpha = 0{,}15$)	**Oesom**	**Glmnet** ($\lambda = 0{,}1043$, $\alpha = 0{,}4$)	**Oesom**
β_0	0.1404327	0.1414562	0.3162623	0.3161963
β_1 : cld6190_ann	0.0000000	0.0000000	0.0000000	0.0000000
β_2 : dtr6190_ann	-0.0585139	-0.0598656	0.0000000	0.0000000
β_3 : ecoreg	0.0000000	0.0000000	0.0000000	0.0000000
β_4 : frs6190_ann	-0.1397514	-0.1408573	-0.0679589	-0.0674658
β_5 : h_dem	-0.3407171	-0.3419454	-0.3969864	-0.3968541
β_6 : pre6190_ann	0.0000000	0.0000000	0.0000000	0.0000000
β_7 : pre6190_l1	-0.0177409	-0.0181999	0.0000000	0.0000000
β_8 : pre6190_l10	0.1917424	0.1919667	0.1380047	0.1378326
β_9 : pre6190_l4	0.0000000	0.0000000	0.0000000	0.0000000
β_{10} : pre6190_l7	0.1410508	0.1407858	0.0310969	0.0310578
β_{11} : tmn6190_ann	0.2077601	0.2073716	0.1962950	0.1962985
β_{12} : tmp6190_ann	0.0000000	0.0000000	0.0000000	0.0000000
β_{13} : tmx6190_ann	0.0000000	0.0000000	0.0000000	0.0000000
β_{14} : vap6190_ann	0.1240615	0.1228466	0.0000075	0.0000000

(para $\gamma = 1500$)

Cuadro 3.5: *Coeficientes de las covariables Glmnet y Oesom*

Como se puede observar los valores para los coeficientes son bastante cercanos. También podemos ver que el Oesom reconoce mejor los coeficientes que son nulos. Con estos coeficientes para las covariables, tenemos que la función objetivo toma los siguientes valores:

	Glmnet	**Oesom**
Función Objetivo	0.6142671380	0.6142665132

($\lambda = 0{,}1$, $\alpha = 0{,}15$ y $\gamma = 1500$)

Cuadro 3.6: *Comparación de la función objetivo*

Si bien las dos soluciones son cercanas, la resolución por el método OESOM nos da una solución más precisa, y alcanza un menor valor del costo. La mayoría de coeficientes son nulos y se identifican las covariables que son más relevantes para el modelo de distribución de la especie.

En las siguientes figuras se muestran la solución del modelo y la trayectoria de la función objetivo a través de las iteraciones.

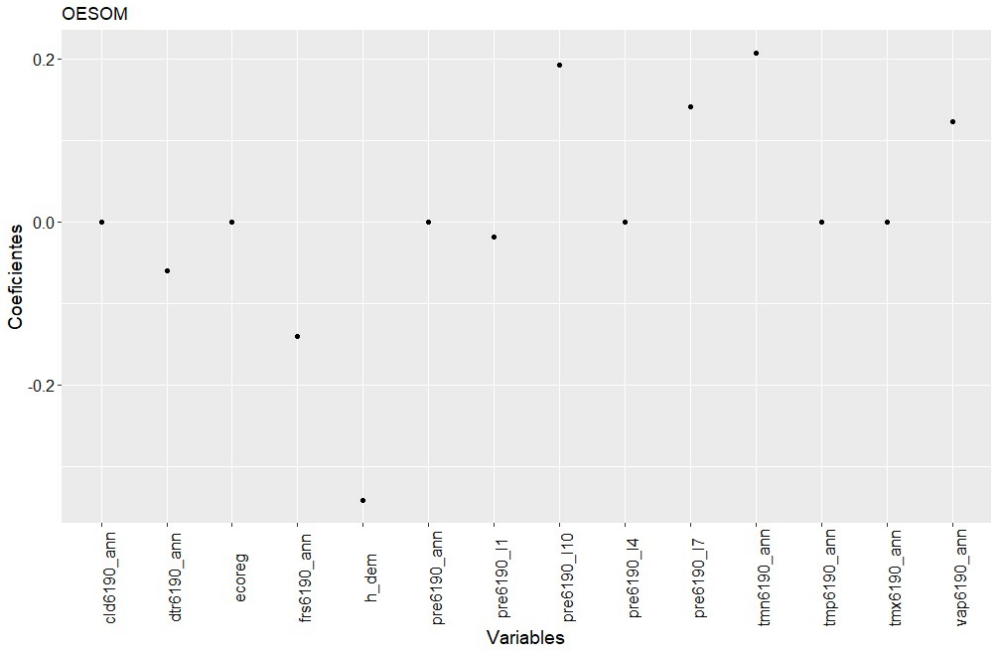


Figura 3.6: *Solución OESOM*

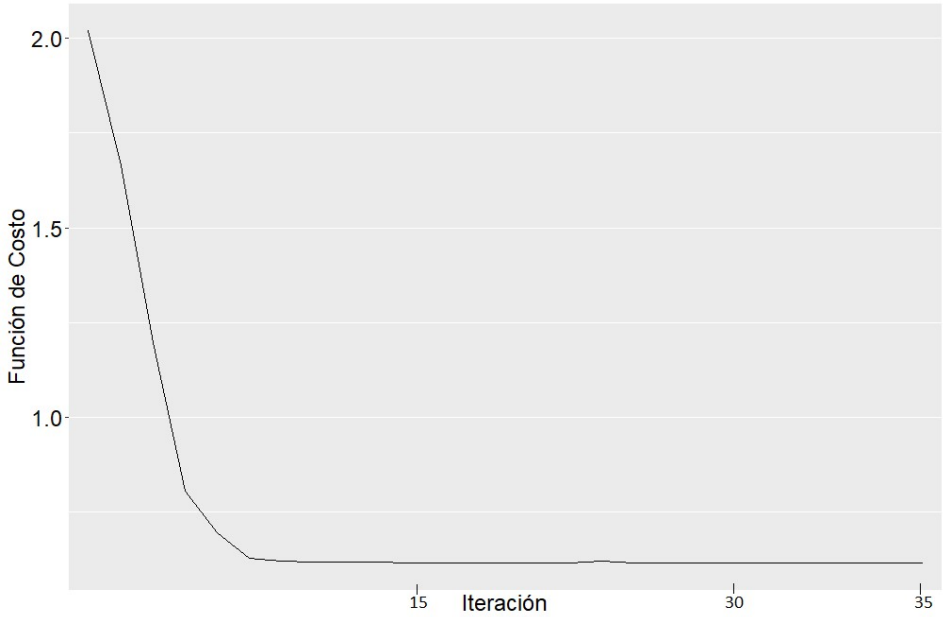


Figura 3.7: *Función Objetivo OESOM*

Ahora, `glmnet` utiliza por default 100 valores de λ, aunque el programa se detiene tempranamente si el porcentaje de desviación explicada no cambia lo su-

ficiente de λ a λ.

En la Figura 3.8 se visualiza la importancia que tienen las variables en el modelo, para $\alpha = 0{,}15$ y $\gamma = 1500$.

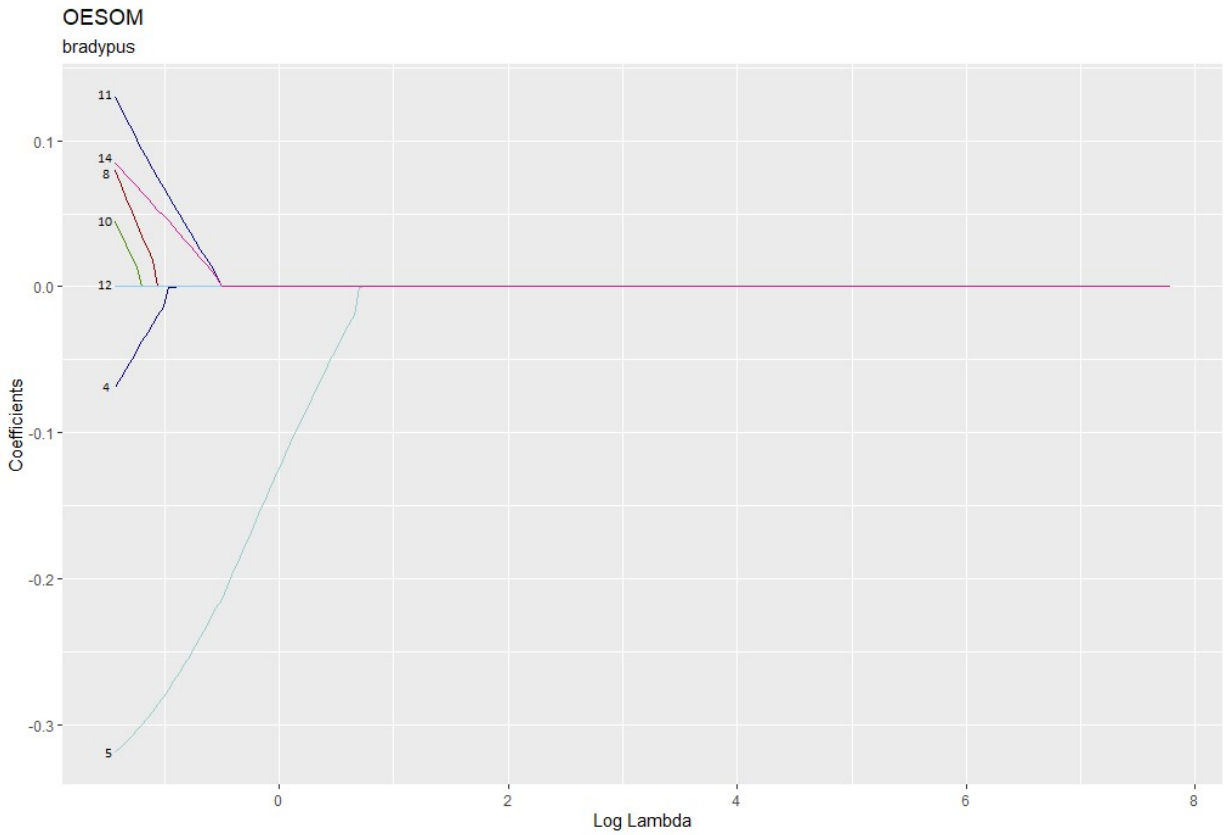


Figura 3.8: *Reducción de Variables Oesom*

Cada curva corresponde a una variable. Muestra el camino de sus coeficientes, a medida que λ varía crecientemente.

Así, podemos ver que las covariables con mayor relevancia para la modelización de la especie *Bradypus variegatus* son: Elevación (β_5), Temperatura media anual (β_{11}), Presión anual de vapor(β_{14}), Frecuencia de heladas anual(β_4), Precipitación de Octubre(β_8) y Precipitación de Julio (β_{10}).

Un ejemplo para otro valor de α es el que se presenta a continuación con $\alpha = 0{,}1$.

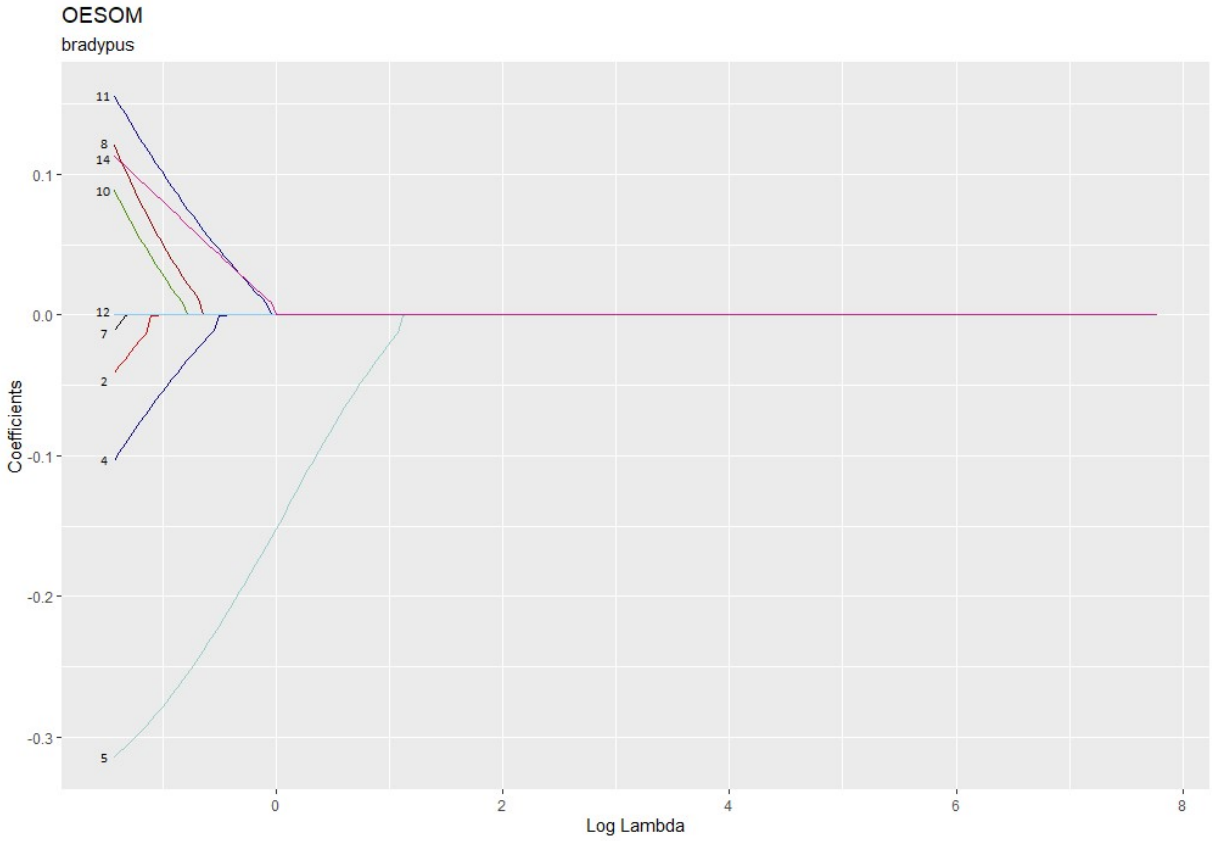


Figura 3.9: *Ejemplo con* $\alpha = 0{,}1$

Para este valor de α aumenta la importancia de las covariables: Rango de temperatura diurna anual (β_2) y Precipitación de Enero (β_7).

Entonces, para las predicciones los datos que se han tomado son los que se muestran en el Cuadro 3.7

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	y
82	53	96	10	1	242	81	53	106	66	88	187	246	309	256	1
134	44	147	12	142	4458	13	36	8	3	3	-73	57	171	25	0

Cuadro 3.7: *Datos de Validación*

Para lo que se ha obtenido la siguiente predicción:

	Predicción	Valor Real
82	0.9999986	1
134	0	0

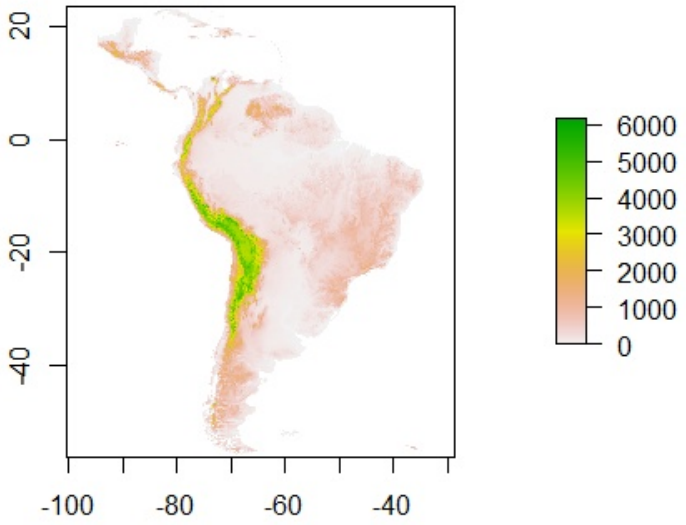
($\lambda = 0{,}15$, $\alpha = 0{,}15$ y $\gamma = 1500$)

Cuadro 3.8: *Predicciones*

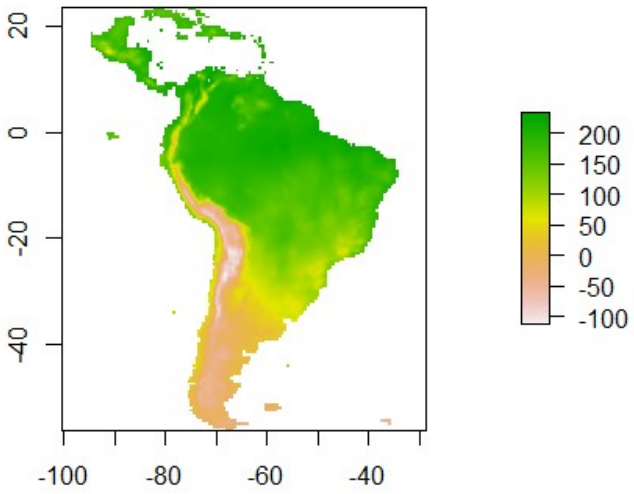
Así, se puede determinar que en el pixel con $longitud = -82{,}2500$ y $latitud = 9{,}1833$ la especie se encuentra presente. Para las predicciones se utilizó los resultados del Oesom. Aunque el beneficio no es muy notorio, razón por la cual en la siguiente sección se utilizará las soluciones del Glmnet por simplificar el uso del software.

3.4. Mapas de predicción

En esta sección se ejemplificará la aplicación en la generación de mapas de predicción de la distribución de la especie *Bradypus variegatus*. Para esto, usamos MaxEnt y sus herramientas de generación de mapas. Hemos tomado los 116 datos de presencia que se utilizaron para la sección anterior (datos de longitud y latitud) y 14 `layers` para la información del terreno correspondiente a cada una de las covariables, en formato `RasterStack`. Tomando 114 datos para el entrenamiento del modelo.



(a) *Elevación*

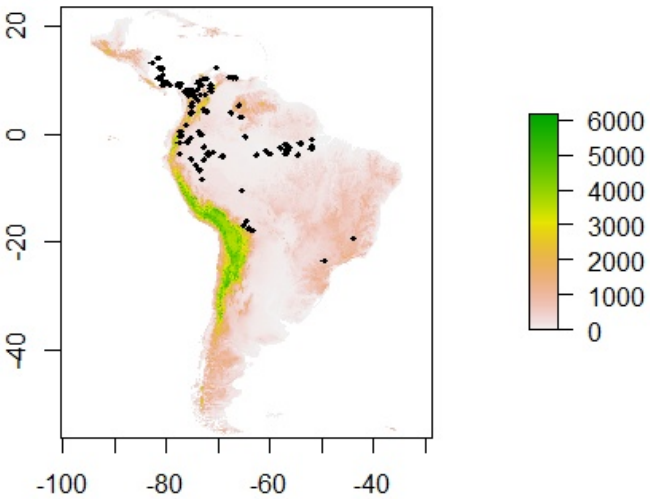


(b) *Temperatura media anual*

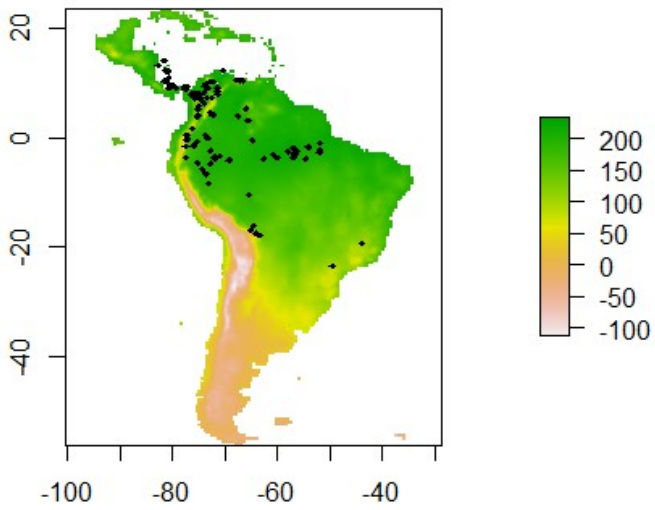
Figura 3.10: *Ejemplo de mapas de covariables*

En la figura anterior podemos visualizar las características que tendría América Centro y Sur con respecto a las covariables. Por ejemplo, las partes más elevadas y las que tendrían mayor temperatura media por año.

En la Figura 3.11 se presentan los ejemplos de la Figura 3.10 con la presencia de la especie (puntos en color negro).



(a) *Elevación*



(b) *Temperatura media anual*

Figura 3.11: *Ejemplo de mapas de covariables con presencia*

Utilizando el modelo entrenado y las herramientas de MaxEnt en R se han obtenido las predicciones siguientes.

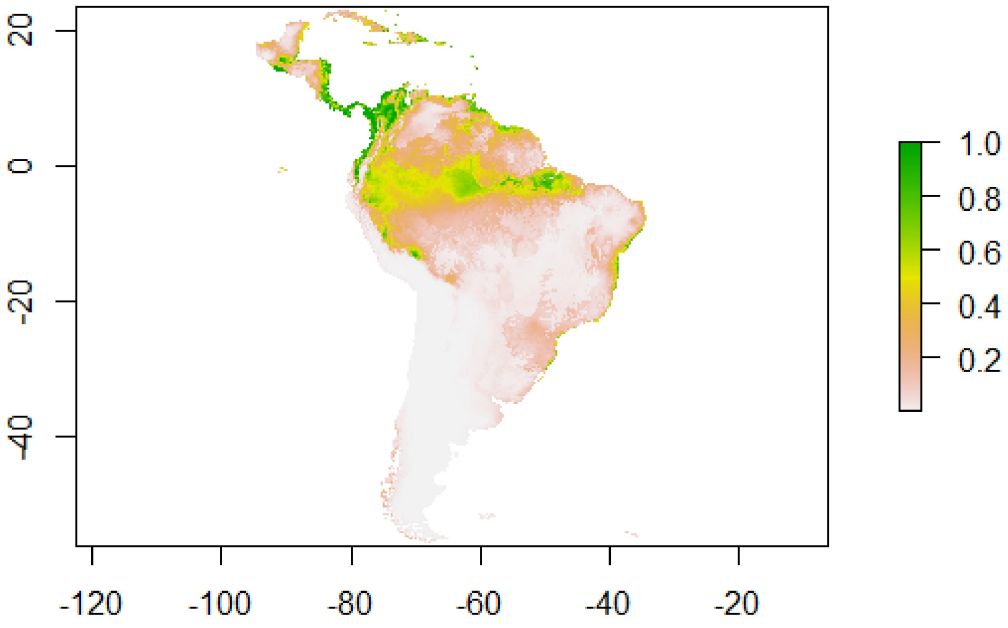


Figura 3.12: *Predicción*

El *AUC* (Área bajo la curva ROC) es una prueba estadística muy utilizada, que permite medir el rendimiento de varios métodos de clasificación. El cual se encuentra entre 0 (muy bajo rendimiento) y 1 (excelente rendimiento)[17]. MaxEnt calcula el valor de este estadístico. Para este ejemplo se obtuvo un valor $AUC = 0{,}8555$, lo que indica un alto rendimiento.

Todo esto se puede realizar desde la consola del software MaxEnt o desde R. Un ejemplo del uso de MaxEnt desde R se encuentra en (`https://github.com/jivelasquezt/courses/blob/master/Hangout_Maxent/maxent.demo.R`):

```
    #Instalación paquetes
#Estas lineas se deben correr si es la primera instalación de
Maxent
#install.packages("dismo") #Añadir maxent.jar en la carpeta
java de dismo
#install.packages("raster")
#install.packages("rgdal")
#install.packages("rJava")

#Configuración inicial
options(java.parameters = "-Xmx1g" )
library(raster)
library(dismo)
setwd("/Users/Marlene/Dropbox/Tesis/MaxEnt/")

#Cargar datos
occs <- read.csv("./samples/bradypus.csv") #Datos de presencia
View(occs)
layers <- stack(list.files("./layers","asc",full.names=TRUE))
plot(layers)
plot(layers[[11]])
layers #Ver en consola las caracteristicas del stack
layers[[1]] #Ver en consola las características de un stack
particular
points(occs[,2:3],pch=18,cex=0.6)

#Formato SWD (samples with data)
#Eliminacion de duplicados
pres.covs<-extract(layers, occs[,2:3],cellnumbers=T)
View(pres.covs)
pres.covs<-na.omit(pres.covs)
pres.covs<-unique(pres.covs)
pres.covs<-pres.covs[,-1] #Elimina columna de número de celda

bkg.covs<-sampleRandom(layers,10000,cells=T)
```

```
bkg.covs<-unique(bkg.covs)
bkg.covs<-bkg.covs[,-1] #Elimina columna de número de celda

#Condensar todo en una sola tabla
env.values<-data.frame(rbind(pres.covs,bkg.covs))
env.values$ecoreg<-as.factor(env.values$ecoreg)

#Etiquetar presencias (1) y background (0)
y <- c(rep(1,nrow(pres.covs)), rep(0,nrow(bkg.covs)))
#Run model
me <- maxent(env.values, y,
             args=c("addsamplestobackground=true"),
             path="./outputR")

#Visualizar los resultados
map <- predict(me, layers, progress="text")
plot(map)
```

Capítulo 4

Conclusiones

El objetivo fundamental de este trabajo consistió en calcular simulaciones precisas del modelo de máxima entropía en la predicción de la distribución de especies en áreas naturales mediante algoritmos de segundo orden. Por el trabajo desarrollado, se sabe que, las soluciones son más precisas. Aunque la diferencia en la solución no es mayormente significativa. Se pudo comprobar que se puede alcanzar soluciones más precisas. Esto, si bien es esperado, abre la posibilidad de estudios más profundos de los beneficios computacionales de los algoritmos de segundo orden en este tipo de problemas. Además, la perspectiva de usar otras funciones de costo no convexas para mejorar el problema de entrenamiento, hace especialmente relevante la consideración de algoritmos de segundo orden.

También se pudo evidenciar que el algoritmo de segundo orden aplicado al problema de máxima entropía reconoce eficientemente los coeficientes de las variables que son nulas.

Las conclusiones que se derivan del trabajo de investigación están relacionadas al funcionamiento de la métodología de máxima entropía y son las que se describen a continuación.

Para trabajar con métodos de solo presencia, se requiere un estudio minucioso a cerca del manejo del sesgo en los datos de muestra. Además de criterios ecológicos acerca de la especie en estudio. Por lo que, si se desea realizar un estudio de una especie, se debe trabajar de la mano con ecologistas. Cabe recalcar que en el presente trabajo se ha utilizado un conjunto de datos que pertenecen a ejemplos desarrollados anteriormente, que no constan de los problemas que se tendrían para una base de datos de alguna especie que no haya sido estudiada; por ejemplo para alguna especie de Ecuador que no haya sido modelada anteriormente.

Al trabajar con un software tan ampliamente desarrollado, el estudio de ciertos

parámetros y resultados se convierte en un problema, pues MaxEnt funciona como una "caja negra" y es difícil acceder a la información dentro del proceso de modelización de MaxEnt.

Apéndice A

Apéndice

A.1. Transición del enfoque geográfico al ambiental

Se demostrará la ecuación (2.3) del escrito principal.
En [36] tenemos que $Pr(x|y = 1)$ es la distribución, dado que la especie está presente, que se encuentre en el punto x de la grilla, y puede ser difícil la conceptualización.

MaxEnt realiza la estimación

$$\begin{aligned} \hat{p}(x|y = 1) &= q_\beta(x) \\ &= \frac{e^{\beta(x)}}{Q_b} \end{aligned}$$

La notación es consistente con el escrito principal en lo posible:

$Z(x)$ el vector de las covariables definida sobre la muestra de la grilla
β el vector de coeficientes
$q_\beta(x)$ la distribución de probabilidad sobre x
Q_β es la constante de normalización que asegura que q_β suma 1

MaxEnt trabaja en las transformaciones de $Z(x)$, es decir $h(Z(x))$, el conjunto de características, pero aquí por simplicidad utilizaremos $Z(x)$. Entonces, dada la

distribución de x $q_\beta(x)$

$$\begin{aligned} f_1(z) &= Pr(Z = z|y = 1) \\ &= \sum_{x \in L: Z(x)=z} q_\beta(x) \\ &= \sum_{x \in L: Z(x)=z} \frac{e^{\beta \cdot Z(x)}}{Q_\beta} \\ &= \frac{f(z)e^{\beta \cdot z}}{Q_\beta/|L|} \\ &= f(z)e^{\alpha+\beta \cdot z} \end{aligned}$$

con

$$f(z) = \frac{\#[x \in L : Z(x) = z]}{|L|}$$

es decir, el número de puntos x en L tales que $Z(x)$ es igual a z, dividido para el número de elementos (grilla) en el terreno L, describe la distribución de z en entorno discreto[12, Apéndice 2]. Para una constante de normalización α que asegura que la suma de $f_1(z)$ de 1.

Bibliografía

[1] Anderson, R. P. y Handley. Jr., C. O. A new species of three-toed sloth (mammalia: Xenarthra) from panama, with a review of the genus bradypus. *Proceedings of the Biological Society of Washington*, 144:1–33, 2001.

[2] Austin, M.P. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157:101–118, 2002.

[3] Barry, S.C. y Elith, J. Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43:413–423, 2006.

[4] Bazaraa, M.S., Sherali, H.D. y Shetty, C.M. *Nonlinear Programming, Theory and Algoritms.* A John Wiley Sons, INC., New Jersey, 3ra edition, 2006.

[5] Blanco, E. ¿Qué es overfitting y cómo evitarlo? Obtenido de `https://empresas.blogthinkbig.com/que-es-overfitting-y-como-evitarlo-html-2/`, 2019.

[6] Ciarlet, P.G. *Linear and Nonlinear Functional Analysis with Aplications.* Society for Industrial and Applied Mathematics, Philadelphia, 2013.

[7] De Los Reyes, J.C., Loayza, E. y Merino, P. Second-order orthant-based methods with enriched hessian information for sparse l_1-optimization. *to appear in Computational Optimization and Applications*, 67(2):225–258, 2017.

[8] Della Pietra, S., Della Pietra, V. y Lafferty, J. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell*, 19(4):1–13, 1997.

[9] Dudik, M., Phillips, S.J. y Schapirte, R.E. Performance guarantees for regularized maximum entropy density estimation. *Proceedings of the 17th Annual Conference on Computational Learning Theory, ACM Press, New York*, pages 655–662, 2004.

[10] Durbán, M. Modelos Aditivos Generalizados con P-splines. Obtenido de `https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf`, 2000.

[11] Elith, J., Kearney, M. y Phillips, S.J. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1:330–342, 2010.

[12] Elith, J., Phillips, S.J., Hastie, T., Dudik, M., Yung En Chee y Yates, C.J. A statistical explanation of maxent for ecologists. *Diversity Distrib.*, 17:43–57, 2011.

[13] Elith, J. y Leathwick, J.R. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, 40:677–697, 2009.

[14] European Comission. Causas del cambio climático. Obtenido de `https://ec.europa.eu/clima/change/causes_es`, 2000.

[15] Fountoulakis, K. y Gondzio, J. A second-order method for strongly convex ℓ_1-regularization problems. *Math. Program.*, 156(1-2):189–219, 2013.

[16] Friedman, J., Hastie, T. y Tibshirani, R. Regularization paths for generalized linear models via coordinate descen. *J Stat Softw*, 33:1–22, 2010.

[17] Google Developers. Clasificación: ROC y AUC. Obtenido de `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419`, 2020.

[18] Gotelli, N.J. y Graves, G.R. *Null models in ecology*. Smithsonian Institution Press, Washington, 1996.

[19] Grinnell, J. The origin and distribution of the chestnut-backed chickadee. *Auk*, 21:364–382, 1904.

[20] Harte, J. Maximum Entropy & Ecology. Obtenido de `http://www.di.fc.ul.pt/~jpn/r/maxent/maxent.html`, 2011.

[21] Harte, J y Newman, E.A. Maximum information entropy: a foundation for ecological theory. *Trends Ecol. Evol.*, 29:384–389, 2014.

[22] Hastie, T., Tibshirani, R. y Friedman, J.H. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York City, USA, 2da edition, 2009.

[23] Hastie, T. y Qian, J. Glmnet Vignette. Obtenido de `http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf`, 2016.

[24] Huang F.-L., Hsieh C.-J., Chang K.-W. y Lin C.-J. Iterative scaling and coordinate descent methods for maximum entropy. *J. Mach. Learn. Res.*, 11:815–848, 2010.

[25] Huang, H.-H., Liu, X.-Y. y Liang, Y. Feature selection and cancer classification via sparse logistic regression with the hybrid $l_{\frac{1}{2}+2}$ regularization. *Computer Science, Medicine, PloS one*, 11(5): e0149675, 2016.

[26] Jaynes, E.T. *Maximum Entropy and Bayesian Methods. Fundamental Theories of Physics.* Cambridge University Press, Cambrigde, 1era edition, 1988.

[27] Kearney, M. y Porter, W. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, 12:334–350, 2009.

[28] Leathwick, J.R., Moilanen, A., Ferrier, S. y Julia, K. Complementarity-based conservation prioritization using a community classification, and its application to riverine ecosystems. *Biol. Conserv*, 143:984–991, 2010.

[29] Loaiza, C.R. y Roque, J.R. Revalidación taxonómica y distribución potencial de armatocereus brevispinus madsen (cactaceae). *Rev. peru biol.*, 23:35–42, 2016.

[30] Manning, C.D. y Schüetze, H. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge Massachusetts, 1999.

[31] Maximenko E. Teorema del valor medio y funciones Lipschitz continuas. Obtenido de `http://esfm.egormaximenko.com/numerical/methods/Lipschitz_continuous_functions_es.pdf`, 2013.

[32] Merow, C., Smith, M.J. y Silander J.A.,Jr. A practical guide to maxent for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36:1058–1069, 2013.

[33] Nagpal, A. Overfitting and Regularization. Obtenido de `https://towardsdatascience.com/over-fitting-and-regularization-64d16100f45c`, 2017.

[34] Newbold, T. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, 34(1):3–22, 2010.

[35] Pant, G., Mansotae, D.K., Sharma S., Goswami, M. y Joshi, P.C. Role of species distribution models in biodiversity conservation. *Current Status of Researches in Biosciences*, pages 507–515, 2020.

[36] Phillips, S.J., Anderson, R.P. y Schapire, R.E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190:231–259, 2006.

[37] Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. y Ferrier, S. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data survey detection probabilities. *Ecological Applications*, 19:181–197, 2009.

[38] Phillips, S.J. y Dudík, M. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31:161–175, 2008.

[39] Pressé, S., Ghosh, K., Lee, J. y Dill, K.A. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85:1116–1139, 2013.

[40] Saha, A. y Tewari, A. On the nonasymptotic convergence of cyclic cordinate descent methods. *SIAM J. Optim.*, 23(1):576–601, 2013.

[41] Siegfried, T. *A Beautiful Math: John Nash, game theory, and the modern quest for a code of nature.* Joseph Henry Press, Washington, DC, 2006.

[42] Soberón, J. y Nakamura, M. Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences USA*, 106:9644–19650, 2009.

[43] Sociedad Andaluz de Eduación Matemática THALES. Criterio de Laplace. Obtenido de `https://thales.cica.es/rd/Recursos/rd99/ed99-0191-03/laplace.htm`, 2010.

[44] Sra, S., Nowozin, S. y Wright, S.J. Optimization for machine learning. *The MIT Press*, 2012.

[45] Suárez Cueto, A. *Resolución de la ambigüedad semántica de las palabras mediante modelos de probabilidad de máxima entropía.* info:eu-repo/semantics/article, Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos, Jun 2004.

[46] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

[47] Ueltschi, D. Shannon entropy. Obtenido de `http://www.ueltschi.org/teaching/chapShannon.pdf.`, 2010.

[48] Ward, G. Statistics in ecological modeling; presence-only data and boosted mars. *Statistics in ecological modeling; presence-only data and boosted mars, Palo Alto*, 2007.

[49] Wintle, B.A., McCarthy, M.A., Parris, K.M. y Burgman, M.A. Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications*, 14:703–712, 2004.

[50] Zadorozhnyi, O., Benecke, G., Mandt, S., Scheffer, T. y Kloft, M. Huber-norm regularization for linear prediction models. *In ECML-PKDD*, pages 714–730, 2016.

[51] Zadrozny, B. Learning and evaluating classifiers under sample selection bias. *In: Proceedings of the 21st International Conference on Machine Learning*, page 903–910, 2004.

[52] Zeng, J., Peng, Z., Lin, S. y Xu, Z. A cyclic coordinate descent algorithm for l_q regularization. *Mathematics, eprint arXiv:1408.0578*, pages 1–13, 2014.