ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELIZACIÓN GEOESTADÍSTICA DE LA DISTRIBUCIÓN POTENCIAL DE RANAS APOSEMÁTICAS EN EL ECUADOR CONTINENTAL

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

PROYECTO DE INVESTIGACIÓN

JOSELYN ISABEL ZAMBRANO CHILIQUINGA
joselyn.zambrano@epn.edu.ec

Directora: UQUILLAS ANDRADE ADRIANA, PH.D.
adriana.uquillas@epn.edu.ec

QUITO, OCTUBRE 2022

DECLARACIÓN

Yo, JOSELYN ISABEL ZAMBRANO CHILIQUINGA, declaro bajo juramento que el trabajo aquí descrito 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.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Joselyn Isabel Zambrano Chiliquinga

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por JOSELYN ISABEL ZAMBRANO CHILIQUINGA, bajo mi supervisión.

Uquillas Andrade Adriana, Ph.D.

Directora del Proyecto

AGRADECIMIENTOS

A mis padres Susana y José, por su amor incondicional, ustedes son mi guía y soporte en todos los proyectos de mi vida.

A mis hermanas Mayra y María, gracias por ser mis mejores amigas, las más divertidas y las más leales.

A mis amigos por todos los buenos momentos, especialmente a Cristian gracias por acompañarme y apoyarme durante esta etapa de mi vida.

A mi tutora Adriana gracias por la confianza, por el tiempo dedicado y principalmente por los conocimientos compartidos.

DEDICATORIA

A mis padres y hermanas, todo es posible porque ustedes están conmigo.

Índice

Índice de Figuras

Índice de Tablas

Resumen

El presente trabajo se centra en el uso de varias técnicas de clasificación y estadística geoespacial con la finalidad de estimar la distribución del hábitat potencial de las ranas de la familia Dendrobatidae (ranas aposemáticas) en el Ecuador continental. Para este estudio, se consideran observaciones de presencia de ranas aposemáticas en el Ecuador durante los años del 2012 al 2019, además se hace uso de datos mensuales de clima y ambiente como variables explicativas.

Debido a la falta de datos de ausencia de las ranas aposemáticas, se usa un método de creación de "pseudoausencias", es decir, se realiza una simulación de observaciones de ausencia de las ranas con el propósito de obtener una variable objetivo binaria. El método de generación de pseudoausencias propuesto en este trabajo hace uso de máquinas de soporte vectorial de una clase y algoritmo K-medias, lo que permite incluir ausencias confibles en la muestra.

Se modela la presencia/ausencia de ranas aposemáticas empleando tres técnicas de clasificación diferentes (GLM, GAM y MARS) que permiten considerar las distintas relaciones que existen entre las variables regresoras y la variable objetivo, y al final se escoge la que presente mejores resultados en términos de poder de predicción. Luego, se realiza análisis Kriging de los residuos del modelo de clasificación escogido con la finalidad de considerar también la relación que tiene la variable objetivo con el espacio.

Finalmente, se presenta el mapa de distribución potencial de ranas aposemáticas en el Ecuador continental con peridiocidad mensual.

Palabras Clave: Modelo de Distribución de Especies, Máquina de Soporte Vectorial, K-medias, GLM, GAM, MARS, Kriging.

Abstract

The present work focuses on the use of various classification techniques and geospatial statistics in order to estimate the potential habitat distribution of Dendrobatidae family frogs (aposematic frogs) in mainland Ecuador. For this study, presence observations of aposematic frogs in Ecuador during the years 2012 to 2019 were considered, in addtion, monthly climate and environment data are used as explanatory variables.

Due to the lack of absence data of aposematic frogs, a method of creating "pseudoabsences" will be used, that is, we will simulate absence observations of frogs in order to obtain a a binary objective variable. The pseudoabsence generation method proposed in this work makes use of one-class support vector machines and a K-means algorithm, which will allow us to include reliable absences in the sample.

The presence/absence of aposematic frogs will be modeled using three different classification techniques (GLM, GAM and MARS) that will allow considering the different relationships that exist between the regressor variables and the target variable, and in the end the one that presents the best results in terms of predictive power will be chosen. Then, a Kriging analysis of the residuals of the chosen classification model will be carried out in order to also consider the relationship that the target variable has with space.

Finally, the potential distribution map of aposematic frogs in mainland Ecuador with monthly periodicity will be presented.

Keywords: Species Distribution Model, Support Vector Machine, K-means, GLM, GAM, MARS, Kriging.

Capítulo 1

1.1 Introducción

La fauna de anfibios del Ecuador es la tercera más diversa en el mundo con un total de 637 especies, de la cuales 291 son especies endémicas, solo superado por Brasil y Colombia según Ron (2021). Ecuador posee aproximadamente 2440 especies por cada millón de km^2 lo cual lo convierte en la región con la concentración más variada de ranas y sapos del planeta.

Las ranas y sapos cumplen un rol indispensable en el funcionamiento de los diversos ecosistemas (páramos, bosques, pantanos, etc.), pues como consumidores y presas son parte fundamental de la cadena de flujo de energía y nutrientes. Además, el Ministerio del Ambiente (2017) resalta que los anfibios también tienen un enorme potencial para contribuir al bienestar humano como fuente de medicinas porque producen substancias con propiedades analgésicas y antibióticas cuyo desarrollo es sujeto de intensa investigación.

Desafortunadamente, la riqueza de los anfibios ecuatorianos es casi desconocida para la mayoría de las personas porque muchas de estas especies viven en áreas de difícil acceso, son relativamente pequeños y de hábitos nocturnos. Esto sumado al cambio climático, destrucción del hábitat, y la introducción de especies, han provocado que el 33% de los anfibios del Ecuador se encuentren en peligro de extinción, y de ellos posiblemente 18 especies se encuentran ya extintas, según el Ministerio del Ambiente (2017).

Es así, que conocer la distribución del hábitat idóneo[1] (o distribución potencial) de

[1]El hábitat idóneo se refiere a las condiciones físicas bajo las cuales una especie podría vivir. La existencia de hábitat idóneo no garantiza la presencia de la especie en ese hábitat, la ausencia de una especie puede ser consecuencia de interacciones bióticas (competencia con otras especies, presencia de predadores, etc.) o la incapacidad de la especie de dispersarse a través de barreras geográficas.

las ranas en el Ecuador permitiría identificar las localidades que presenten características ambientales favorables para la existencia de estas especies, crear indicadores del estado del ecosistema que habitan, alertar posibles fuentes de riesgo y de esta forma establecer políticas de conservación de las mismas. En este contexto, se propone modelar la distribución potencial de ranas en el Ecuador, incorporando técnicas de generación de "pseudo-ausencias", técnicas de regresión y técnicas de geoestadística como Kriging. Para ello se considerarán datos georeferenciados e históricos de presencia de especies de ranas en distintas regiones del Ecuador, así como información del clima y condiciones ambientales asociadas a estas locaciones.

Como el problema es conocer la distribución potencial de una especie, en este caso las ranas, diversos estudios se han planteado con ese objetivo, varios de ellos se basan en técnicas de perfil, que modelan la distribución espacial considerando únicamente la información de las locaciones donde se ha observado la presencia de la especie analizada; sin embargo, según Zaniewski et al. (2002), el principal problema de estas técnicas es que dado que los conjuntos de datos de solo presencias carecen de información sobre las ausencias y generalmente sufren de un sesgo asociado con el muestreo ad hoc o no estratificado, a menudo se asume que son problemáticos e inadecuados para la mayoría de los métodos de modelado estadístico.

Para sobrellevar este problema se han propuesto una serie de técnicas alternativas que modelan la distribución del nicho ecológico adecuado de especies a partir de modelos estadísticos que, además de los conjuntos de datos de solo presencias, incorporan la generación "pseudo ausencias", las cuales tienen el potencial de ser una alternativa conveniente y útil cuando los datos de ausencia recopilados sistemáticamente no están disponibles o son imposibles de obtener.

Así, dado que se cuenta con datos únicamente de presencia de ranas de la familia Dendrobatidae, conocidas por su condición de aposematismo como ranas aposemáticas, en el período del 2012 al 2019 obtenidas de la página web del proyecto "Anfibios del Ecuador"

se plantea la generación de "pseudo ausencias" a partir del método Tres pasos con selección k-media (TSKM) y de esta manera poder utilizar modelos estadísticos (GLM, GAM, MARS) de distribución de nicho ecológico de especies.

El método de tres pasos con selección k-medias (TSKM) propuesto por Iturbide et al. (2015), comienza con una etapa de Selección Aleatoria de Perfil Ambiental (RSEP) que usa Máquinas de Soporte Vectorial de una clase (OCSVM), Schölkopf et al. (2001), para definir las áreas ambientales inadecuadas para la supervivencia de la especie, estas regiones dependen de la elección del umbral de distancia respecto de las presencias. En el segundo paso se construyen subregiones de las áreas ambientales inadecuadas, usando clustering k-medias sobre las variables ambientales (donde k es igual al número de presencias) y se toman las coordenadas de los centroides de estos clústeres como pseudo-ausencias. Finalmente, la etapa tres consiste en repetir los pasos uno y dos variando los umbrales de distancia respecto de las presencias hasta obtener las áreas inadecuadas óptimas.

1.2 Descripción de los Datos

1.2.1 Variable Independiente

Para este estudio se considera una base con 1510 observaciones georeferenciadas de ranas aposemáticas en el lapso de 2012 a 2019. La base de datos se obtuvo de la página web (Ron, S. R., Merino-Viteri, A. Ortiz, D. A. 2019. Anfibios del Ecuador. Version 2019.0. Museo de Zoología, Pontificia Universidad Católica del Ecuador. https://bioweb.bio/faunaweb/amphibiaweb).

En la tabla 1.1 se pueden observar las 22 especies de ranas pertenecientes a la familia Dendrobatidae que aparecen en la base, estas ranas son conocidas como ranas venenosas de dardo o ranas punta de flecha, endémicas de Centroamérica y América del Sur, se caracterizan por su piel brillante y coloreada, es decir por poseer una coloración aposemática. Además,

se incluye la clasificación que estas especies de ranas tienen en la lista roja de la IUCN, es posible observar que en la base tan solo 7 especies no se encuentran amenazadas pues tienen clasificación ”Preocupación menor”, el resto de especies, presentes en la tabla, presentan un mayor riesgo de extinción, pues 6 especies se encuentran en la categoría ”Casi amenazada”, 5 especies se encuentran ”Vulnerables” y 4 especies están ”En peligro”.

Estas especies de ranas habitan ecosistemas muy diversos: bosques de nubes, selvas tropicales de tierras bajas, bosques andinos xerofíticos, etc., siendo su rango altitudinal entre los 300m a los 2000m según Grant et al. (2006).

Especie	**Cantidad**	**Lista Roja IUCN**
Ameerega bilinguis	192	Preocupación menor
Ameerega hahneli	29	Preocupación menor
Ameerega parvula	112	Preocupación menor
Epipedobates anthonyi	103	Casi amenazada
Epipedobates boulengeri	72	Preocupación menor
Epipedobates darwinwallacei	214	Vulnerable
Epipedobates espinosai	6	Vulnerable
Epipedobates machalilla	66	Preocupación menor
Epipedobates tricolor	141	En peligro
Hyloxalus awa	25	Casi amenazada
Hyloxalus elachyhistus	4	En peligro
Hyloxalus infraguttatus	194	Vulnerable
Hyloxalus maculosus	2	En peligro
Hyloxalus mystax	2	En peligro
Hyloxalus sauli	1	Casi amenazada
Hyloxalus toachi	8	Vulnerable
Hyloxalus vertebralis	8	Vulnerable

Especie	Cantidad	Lista Roja IUCN
Hyloxalus yasuni	2	Casi amenazada
Leucostethus fugax	1	Casi amenazada
Oophaga sylvatica	270	Casi amenazada
Ranitomeya variabilis	41	Preocupación menor
Ranitomeya ventrimaculata	17	Preocupación menor

Tabla. 1.1: Especies presentes en la base de presencia de ranas

En la figura 1.1 se pueden observar las ubicaciones de las observaciones de presencias de ranas de la familia Dendrobatidae, se observa que la mayor parte de estas observaciones se encuentran en las regiones más calientes y húmedas del Ecuador.

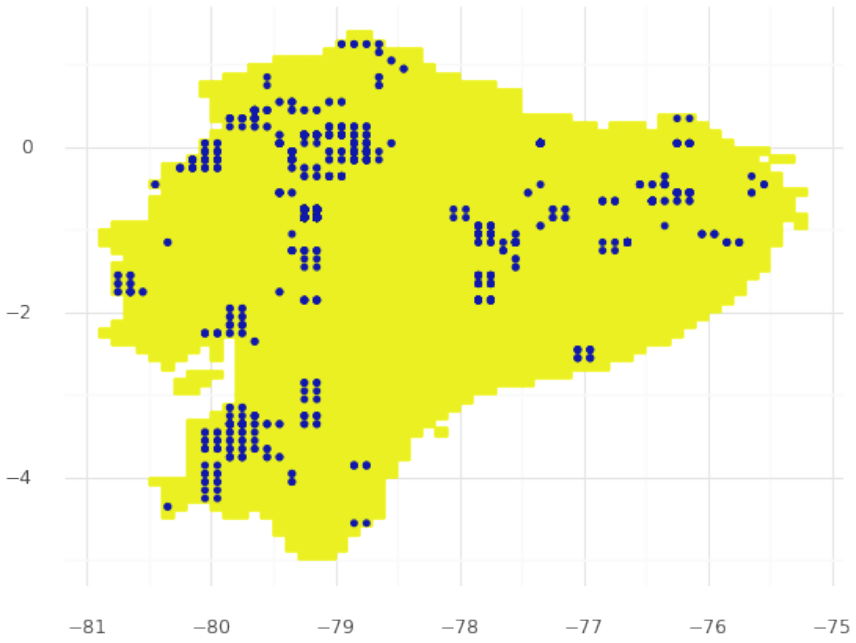


Figura. 1.1: Mapa de las observaciones de presencias de ranas

1.2.2 Variables explicativas

Para esta investigación se consideran tres tipos de variables: climáticas, geográficas y ambientales.

Variables Climáticas

Estas variables se obtuvieron de las bases de datos de la NASA y se descargaron desde el portal web del proyecto MODIS (https://modis.gsfc.nasa.gov/), estas variables fueron obtenidas con la herramienta MODIS (o espectrorradiómetro de imágenes de resolución moderada) la cual es un instrumento clave a bordo de los satélites Terra (originalmente conocido como EOS AM-1) y Aqua (originalmente conocido como EOS PM-1). Estos datos son recolectados para mejorar la compresión de la dinámica y los procesos globales que ocurren en la tierra, océanos y en la atmósfera inferior.

MODIS desempeña un papel vital en el desarrollo de modelos que expliquen el funcionamiento de los sistemas terrestres, capaces de predecir el cambio global con la suficiente precisión para mejorar la toma de decisiones acertadas con respecto a la protección de nuestro medio ambiente.

En este estudio se utilizaron los siguientes productos de la bases de datos satélitales MODIS.

Temperatura de la Superficie Terrestre

Esta variable proporciona valores mensuales de temperatura de la superficie terrestre (LST) en una cuadrícula de modelado climático de latitud/longitud de 0,05 grados (5600 metros).

La temperatura de la superficie terrestre es una variable importante dentro del sistema

climático de la Tierra, pues describe procesos como el intercambio de energía y agua entre la superficie terrestre y la atmósfera, y tiene gran influencia en la velocidad y el momento del crecimiento de las plantas. La comprensión precisa de la LST a nivel mundial y regional ayuda a evaluar los procesos de intercambio de la superficie terrestre con la atmósfera en modelos y, cuando se combina con otras propiedades físicas como la vegetación, la humedad del suelo, etc., proporciona una valiosa información sobre el estado de la superficie, esto según ESA Climate Office (2021).

Según la ESA Climate Office (2021), la temperatura superficial de la Tierra también proporciona datos de temperatura independientes para complementar las mediciones in situ y el reanálisis asociados con la temperatura del aire cerca de la superficie, un objetivo fundamental descrito en el acuerdo de París de la UNFCCC (United Nation Climate Change).

Como se puede observar en la figura 1.2 la temperatura de la tierra ha aumentado con el transcurso de los años, pues desde finales del siglo XIX, un cambio impulsado en gran medida por el aumento de las emisiones de dióxido de carbono a la atmósfera y otras actividades humanas ha hecho en los últimos 40 años se produzca una elevación de la temperatura, registrandose las temperaturas más altas en los años 2016 y 2017, según Susskind et al. (2019).

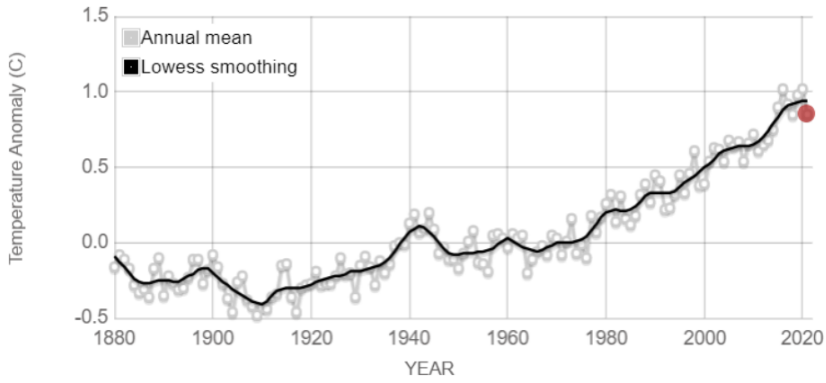


Figura. 1.2: Índice de la Temperatura Global de la Tierra **Fuente:** GISTEMP Team (2022)

Precipitación

Esta variable proporciona valores mensuales de la precipitación en una cuadrícula de modelado climático de latitud/longitud de 0,05 grados (5600 metros).

La precipitación es el resultado de la condensación del vapor de agua atmosférico que se deposita en la superficie de la Tierra. Este fenómeno ocurre cuando la atmósfera (que es una gran solución gaseosa) se satura con el vapor de agua, y el agua se condensa y cae (es decir, precipita). Las precipitaciones constituyen uno de los recursos naturales más preciados de la Tierra pues es un recurso básico del que dependen gran parte de los organismos, por lo que cualquier cambio en el mismo repercute sobre la naturaleza y la sociedad, según Intergovernmental Panel On Climate Change (IPOCC) (2007).

Evapotranspiración

Esta variable proporciona valores mensuales de la precipitación en una cuadrícula de modelado climático de latitud/longitud de 0,05 grados (5600 metros).

La evapotranspiración es la pérdida de humedad de la superficie terrestre por evaporación directa junto con la pérdida de agua por transpiración de la vegetación o de la superfecie del suelo. Es una variable importante debido a que enlaza el ciclo del agua, el ciclo de energía y el ciclo de carbono. Al ser cuantificada de manera precisa, puede aportar a un mejor manejo de los recursos hídricos y a mejorar las predicciones y la mitigación frente al cambio climático.

Variable Geográfica

La altitud o elevación del Ecuador medida en metros se obtuvo de las bases de datos de WordClim, Hijmans et al. (2005) con una resolución de 10 km, que a su vez deriva sus bases de datos de la Misión Topográfica Shuttle Radar (SRTM) cuya finalidad es obtener un modelo digital de elevación de la zona del globo terráqueo.

La altitud se define como la distancia vertical que existe entre cualquier punto de la Tierra en relación al nivel del mar, es por esto que para poder medirla se toma como referencia el nivel del mar.

La altitud suele utilizarse comunmente en los modelos de distribución de especie pues es un factor restrictivo para ciertas especies, las cuales no pueden sobrevivir a elevaciones muy altas, además este es un factor de cambios de temperatura, ya que esta disminuirá a mayores altitudes y aumentará a menores altidudes, según NASA Earthdata (2020)

Variable Ambiental

El Índice de Vegetación de Diferencia Normalizada (NDVI) se obtuvo de las bases de datos MODIS de la NASA, esta variable corresponde a valores mensuales del índice de vegetación de diferencia normalizada en una cuadrícula de modelado climático de latitud/longitud de 0,05 grados (5600 metros).

El NDVI corresponde a un índice de vegetación que se utiliza para estimar la cantidad, calidad y desarrollo de la vegetación con base a la medición de la intensidad de la radiación de ciertas bandas del espectro electromagnético que la vegetación emite o refleja. Para el cálculo de los índices de vegetación es necesaria la información que se encuentra en las bandas roja e infrarroja del espectro electromagnético. EOS (2020)

Este índice es muy utilizado para vigilar sequías, predecir la producción agrícola, ayudar a predecir zonas de incendios y áreas en proceso de desertización. EOS (2020)

La fórmula para el cálculo del Índice de Vegetación de Diferencia Normalizada es:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

donde,

NIR es la Espectroscopía de reflectancia en el infrarrojo cercano y RED es la Espectroscopía de reflectancia de la parte roja visible

Así, la densidad de la vegetación (NDVI) en un punto determinado de la imagen es igual a la diferencia en las intensidades de la luz reflejada en el rango rojo e infrarrojo dividido por la suma de estas intensidades.

Este índice define valores de -1.0 a 1.0, donde los valores negativos están formados principalmente por nubes, agua y nieve, y los valores negativos cercanos a cero están formados principalmente por rocas y suelo descubierto. Los valores muy pequeños (0,1 o menos) de la función NDVI corresponden a áreas sin rocas, arena o nieve. Los valores moderados (de 0,2 a 0,3) representan arbustos y praderas, mientras que los valores grandes (de 0,6 a 0,8) indican bosques templados y tropicales.

El NDVI es una medida del estado de la vegetación basada en la forma en que una planta refleja la luz en ciertas frecuencias (algunas ondas se absorben y otras se reflejan).

Por ejemplo, la clorofila (un indicador de salud de la vegetación) absorbe una gran cantidad de luz visible y la estructura celular de las hojas refleja intensamente la luz infrarroja cercana. Cuando una planta se deshidrata, enferma, o sufre alguna afectación, el tejido de las plantas se deteriora y la esta absorbe más luz infrarroja cercana, en lugar de reflejarla. De esta forma, la observación de cómo cambia la NIR en comparación con la luz roja proporciona una indicación precisa de la presencia de clorofila, que está fuertemente vinculada con la salud de las plantas. EOS (2020)

Capítulo 2

Marco teórico

En este capítulo se presentan definiciones y conceptos teóricos necesarios para comprender la metodología utilizada en la construcción del modelo de distribución de las ranas aposemáticas en el Ecuador continental, la cual consta de las siguientes etapas:

1. Análisis y limpieza de las variables climáticas y ambientales.
    a) Utilizar el método de IDW para realizar la imputación de las variables climáticas.
2. Creación de las pseudoausencias.
    a) Usar análisis de componentes principales en las variables regresoras para reducir la dimensionalidad de los datos con la finalidad de reducir la complejidad del algoritmo, lo que se traduce en menos tiempo computacional, además por Su et al. (2017) se conoce que usar SVM en conjunto con ACP mejora el poder de discriminación del modelo.
    b) Ajustar un modelo de máquina de soporte vectorial de una clase.
    c) Utilizar el modelo entrenado en el literal b) para realizar las predicciones sobre todo el espacio (espacio - temporal).
    d) En las locaciones donde se predijeron ausencias de las ranas Dendrobatidaes se utilza el algoritmo k-medias con la finalidad de ubicar las pseudoausencias de las ranas.
3. Una vez que se tiene la variable binaria conformada por presencias y pseudoaunsencias se crea un modelo de clasificación.

a) Como primera opción se ajusta un modelo de clasificación lineal logística.

b) De igual manera, se ajustará un modelo MARS logístico que se espera permita mayor flexibilidad que un modelo GLM.

c) Una tercera opción consiste en ajustar un modelo de clasificación GAM logístico puesto que, permite una mayor flexibilidad en la relación entre las variables regresoras y la variable dependiente.

4. Finalmente, se compararán los tres modelos obtenidos y se optará por aquel que produzca mejores resultados en términos de poder predictivo.

5. Una vez seleccionado el mejor modelo se modelarán los residuos usando Kriging.

Para el entendimiento de cada una de las etapas a continuación se detallan las nociones y base teórica de cada una de las técnicas mencionadas.

2.1 Interpolación con la Distancia Inversa Ponderada (IDW)

La interpolación mediante distancia inversa ponderada determina los valores de celda a través de una combinación ponderada linealmente de un conjunto de puntos de muestra. El interpolador de ponderación de distancia inversa asume que cada punto de entrada tiene una influencia local que disminuye con la distancia. Pondera más los puntos más cercanos a la celda de procesamiento que los más alejados. Se puede usar un número específico de puntos, o todos los puntos dentro de un radio específico, para determinar el valor de salida de cada ubicación. El uso de este método asume que la variable que se está mapeando disminuye en influencia con la distancia desde su ubicación muestreada.

La interpolación IDW de un punto $\hat{y}_j$ para una posición j es calculada como

$$\hat{y}_j = \sum_{i=1}^{n} w_{i,j} \cdot y_i$$

donde n es el número de puntos usados en la interpolación, y_i es el valor en el punto i-ésimo y $w_{i,j}$ es el peso asociado al dato i en el cálculo del nodo j. Los pesos $w_{i,j}$ para cada punto de la data está dado por:

$$w_{i,j} = \frac{d_{i,j}^{-\alpha}}{\sum_{k=1}^{n} d_{k,j}^{-\alpha}}$$

donde $d_{i,j}$ es la distancia euclidiana entre un punto de datos disponible en la ubicación i y los datos desconocidos en la ubicación j, n es el número de puntos disponibles en la base de datos, α es la potencia, y es un parámetro de control. Mientras más alto sea el valor de α, el peso de los puntos más cercanos será mayor. Para la optimización de este coeficiente se tiene que minimizar el error medio cuadrático (EMC) a través de una validación cruzada Brovelli et al. (2008).

El error medio cuadrático es calculado como

$$EMC = \sqrt{\frac{\sum_{j=1}^{n} (\hat{y}_j - y_j)^2}{n}}$$

En este trabajo, IDW se restringe a la ponderación de la distancia al cuadrado inverso pues $\alpha = 2$ es asumido, este es el valor más comúnmente adoptado.

2.2 Análisis de Componentes Principales (ACP)

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Sin embargo, hasta la aparición de los ordenadores no se empezaron a popularizar. Para estudiar las relaciones que se presentan entre p variables correlacionadas se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales.

Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

De modo ideal, se buscan $m < p$ variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

2.2.1 Cálculo de las Componentes Principales

Considerando una serie de variables $(x_1, x_2, ..., x_p)$, se busca calcular a partir de ellas un nuevo conjunto de variables $y_1, y_2, ..., y_p$, incorreladas entre sí, cuyas varianzas vayan decreciendo progresivamente.

Cada y_j donde $j = 1, ..., p$ es una combinación lineal de las variables originales $x_1, x_2, ..., x_p$, es decir,

$$\begin{aligned} y_j &= a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p \\ &= \mathbf{a_j'x} \end{aligned}$$

siendo $a_j' = (a_{1j}, a_{2j}, ..., a_{pj})$ un vector de constantes, y

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Obviamente, si lo que se quiere es maximizar la varianza, como se verá luego, una forma simple podría ser aumentar los coeficientes a_{ij}. Por ello, para mantener la ortogonalidad de la transformación se impone con el módulo del vector $\mathbf{a_j'} = (a_{1j}, a_{2j}, ..., a_{pj})$ sea 1, es decir,

$$\mathbf{a_j'a_j} = \sum_{k=1}^{p} a_{kj}^2 = 1$$

El primer componente se calcula eligiendo $\mathbf{a_1}$ de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción de que $\mathbf{a_j'a_j} = 1$. El segundo componente principal se calcula obteniendo $\mathbf{a_2}$ de modo que la variable obtenida, y_2 esté incorrelada con y_1.

Del mismo modo se eligen $y_1, y_2, ..., y_p$, incorreladas entre sí, de manera que las variables aleatorias obtenidas vayan teniendo cada vez menor varianza.

Proceso de extracción de factores:

Se busca elegir $\mathbf{a_1}$ de modo que se maximice la varianza de y_1 sujeto a la restricción de que $\mathbf{a_1'a_1} = 1$

$$Var(y_1) = Var(\mathbf{a_1'}x) = \mathbf{a_1'}\Sigma\mathbf{a_1}$$

El método habitual para maximizar una función de varias variables sujeta a restricciones el método de los multiplicadores de Lagrange.

El problema consiste en maximizar la función $\mathbf{a_1'}\Sigma\mathbf{a_1}$ sujeta a la restricción $\mathbf{a_1'a_1} = 1$. Se observa que la incógnita es precisamente $\mathbf{a_1}$ que es el vector desconocido que se tiene de la combinación lineal óptima.

Así, se construye la función L:

$$L(\mathbf{a_1}) = \mathbf{a_1'}\Sigma\mathbf{a_1} - \lambda(\mathbf{a_1'a_1} - 1)$$

y se busca el máximo derivando e igualando a 0:

$$\frac{\partial L}{\partial \mathbf{a_1}} = 2\Sigma\mathbf{a_1} - 2\lambda I\mathbf{a_1} = 0 \Rightarrow$$

$$(\Sigma - \lambda I)\mathbf{a_1} = 0$$

Esto es, en realidad, un sistema lineal de ecuaciones. Por el teorema de Roché-Frobenius, para que el sistema tenga una solución distinta de 0 la matriz $(\Sigma - \lambda I)$ tiene que ser singular. Esto implica que el determinante debe ser igual a cero:

$$|\Sigma - \lambda I| = 0$$

y de este modo, λ es un autovalor de Σ. La matriz de covarianzas Σ es de orden p y si además es definida positiva, tendrá p autovalores distintos $\lambda_1, \lambda_2, ..., \lambda_p$ tales que, por ejemplo, $\lambda_1 > \lambda_2 > ... > \lambda_p$.

Se tiene que, desarrollando la expresión anterior,

$$\begin{aligned}(\Sigma - \lambda I)\mathbf{a_1} &= 0 \\ \Sigma\mathbf{a_1} - \lambda I\mathbf{a_1} &= 0\Sigma\mathbf{a_1} \\ &= \lambda I\mathbf{a_1}\end{aligned}$$

entonces,

$$\begin{aligned}Var(y_1) = Var(\mathbf{a_1}'\mathbf{x}) &= \mathbf{a_1}'\Sigma\mathbf{a_1} \\ &= \mathbf{a_1}'\lambda I\mathbf{a_1} = \lambda\mathbf{a_1}'\mathbf{a_1} \\ &= \lambda \cdot 1 = \lambda\end{aligned}$$

Luego, para maximizar la varianza de y_1 se tiene que tomar el mayor autovalor, λ_1, y el correspondiente autovector $\mathbf{a_1}$.

En realidad, $\mathbf{a_1}$ es un vector que da la combinación de las variables originales que tienen mayor varianza, esto es, si $\mathbf{a_1}' = (a_{11}, a_{12}, ..., a_{1p})$, entonces

$$y_1 = \mathbf{a_1}'\mathbf{x} = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

El segundo componente principal, $y_2 = \mathbf{a_2}'\mathbf{x}$, se obtiene mediante un argumento parecido. Además, se quiere que y_2 esté incorrelado con el anterior componente y_1, es decir, $Cov(y_2, y_1) = 0$. Por lo tanto:

$$\begin{aligned}Cov(y_2, y_1) &= Cov(\mathbf{a_2}'\mathbf{x}, \mathbf{a_1}'\mathbf{x}) \\ &= \mathbf{a_2}' \cdot E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'] \cdot \mathbf{a_1} \\ &= \mathbf{a_2}'\Sigma\mathbf{a_1}\end{aligned}$$

es decir, se requiere que

$$\mathbf{a_2}'\Sigma\mathbf{a_1} = 0$$

Como se tenía que $\Sigma\mathbf{a_1} = \lambda\mathbf{a_1}$, lo anterior es equivalente a

$$\mathbf{a_2}'\Sigma\mathbf{a_1} = \mathbf{a_2}'\lambda\mathbf{a_1} = \lambda\mathbf{a_2}'\mathbf{a_1} = 0$$

esto equivale a que $\mathbf{a_2}'\mathbf{a_1} = 0$, es decir, que los vectores sean ortogonales.

De este modo, se tendrá que maximizar la varianza de y_2, es decir, $\mathbf{a_2}\Sigma\mathbf{a_2}$, sujeta a las siguientes restricciones

$$\mathbf{a_2}\Sigma\mathbf{a_2} = 1$$

$$\mathbf{a_2}\Sigma\mathbf{a_2} = 0$$

Se toma la función:

$$L(\mathbf{a_2}) = \mathbf{a_2}'\Sigma\mathbf{a_2} - \lambda(\mathbf{a_2}'\mathbf{a_2} - 1) - \delta\mathbf{a_2}'\mathbf{a_1}$$

y se deriva:

$$\frac{\partial L(\mathbf{a_2})}{\partial \mathbf{a_2}} = 2\Sigma\mathbf{a_2} - 2\lambda\mathbf{a_2} - \delta\mathbf{a_1} = 0$$

si se multiplica por $\mathbf{a_1}'$, entonces

$$2\mathbf{a_1}'\Sigma\mathbf{a_2} - \delta = 0$$

porque

$$\mathbf{a_1}'\mathbf{a_2} = \mathbf{a_2}'\mathbf{a_1} = 0$$

$$\mathbf{a_1}'\mathbf{a_1} = 1$$

Luego,

$$\delta = 2\mathbf{a_1}'\Sigma\mathbf{a_2} = 2\mathbf{a_2}'\Sigma\mathbf{a_1} = 0$$

ya que $Cov(y_2, y_1) = 0$.

De este modo, $\frac{\partial L(\mathbf{a_2})}{\mathbf{a_2}}$ queda finalmente como:

$$\frac{\partial L(\mathbf{a_2})}{\partial \mathbf{a_2}} = 2\Sigma\mathbf{a_2} - 2\lambda\mathbf{a_2} - \delta\mathbf{a_1} = 2\Sigma\mathbf{a_2} - 2\lambda\mathbf{a_2}$$

$$(\Sigma - \lambda I)\mathbf{a_2} = 0$$

Usando el mismo razonamiento de antes, se elige λ como el segundo mayor autovalor de la matriz Σ con su autovector asociado $\mathbf{a_2}$.

Estos razonamientos se pueden extender, de modo que al j-ésimo componente le correspondería el j-ésimo autovalor.

Entonces todos los componentes $\mathbf{y}$ (en total p) se pueden expresar como el producto de una matriz formada por los autovectores, multiplicada por el vector $\mathbf{x}$ que contiene las variables originales $x_1, ..., x_p$

$$\mathbf{y} = A\mathbf{x}$$

donde,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

Como

$$Var(y_1) = \lambda_1$$
$$Var(y_2) = \lambda_2$$
$$\cdots$$
$$Var(y_p) = \lambda_p$$

la matriz de covarianzas de $\mathbf{y}$ será

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}$$

porque $y_1, ..., y_p$ se han construido como variables incorreladas.

Se tiene que

$$\Lambda = Var(Y) = A^{'}Var(X)A = A^{'}\Sigma A$$

o bien

$$\Sigma = A\Lambda A^{'}$$

ya que A es una matriz ortogonal (porque $\mathbf{a_i^{'}a_i}) = 1$ para todas sus columnas) por lo que $AA^{'} = I$.

2.3 Máquinas de Soporte Vectorial

Una Máquina de Soporte Vectorial es un modelo de clasificación desarrollado por Vladimir Vapnik y su equipo en los laboratorios AT&T., este método primero mapea los puntos de entrada a un espacio de características de una dimensión mayor para luego encontrar un hiperplano que separe los puntos según sus clases y además maximice el margen "m" entre estas.

Si los datos de entrenamiento son linealmente separables, es posible seleccionar dos hiperplanos paralelos que separen las dos clases de datos, de modo que la distancia entre ellos sea lo más grande posible. La región delimitada por estos dos hiperplanos se llama "margen", y el hiperplano de margen máximo es el hiperplano que se encuentra a medio camino entre ellos. Cortes and Vapnik (1995)

Sean los datos de entrenamiento (x_i, y_i) $i = 1, .., n$ donde y_i es 1 o -1, indicando la clase del punto x_i.

El hiperplano buscado puede ser escrito de la siguiente forma

$$w^T x - b = 1$$

donde todo lo que esté en este límite o por encima de él es de la clase 1, y

$$w^T x - b = -1$$

donde todo lo que esté en este límite o por encima de él es de la clase -1.

Geométricamente, la distancia entre estos dos hiperplanos es $\frac{2}{||w||}$, según Cortes and Vapnik (1995), así para maximizar la distancia entre los planos se debe minimizar $||w||$. La distancia se calcula utilizando la ecuación de distancia de un punto a un plano. También se debe que evitar que los puntos de datos caigan en el margen, para esto se agrega la siguiente restricción

$$y_i(w^T x_i - b) \geq 1$$

para todo $i = 1, .., n$. Así, se obtiene el siguiente problema de optimización

$$\begin{aligned} \min \quad & ||w|| \\ \text{s.t.} \quad & y_i(w^T x_i - b) \geq 1 \text{ para todo } 1 \leq i \leq n \end{aligned}$$

Desafortunadamente, los fenómenos a estudiar no se suelen presentar en casos idílicos de dos dimensiones, sino que un algoritmo SVM debe tratar con varias particularidades como:

a. Más de dos variables predictoras

b. Curvas no lineales de separación

c. Casos donde los conjuntos de datos no pueden ser completamente separados

d. Clasificaciones en más de dos categorías.

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio de funciones Kernel ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de la máquinas de aprendizaje lineal. Hofmann et al. (2008)

Sean x y $x^{'}$ en el espacio de entrada X, la función tal que

$$k : X \ x \ X \rightarrow \mathbb{R}$$

se conoce como una función kernel, esta función satisface que, para todo x, $x^{'} \in X$

$$k(x, x^{'}) = \langle \phi(x), \phi(x^{'}) \rangle$$

donde ϕ mapea en algún espacio de producto interno $\mathcal{H}$ (espacio de Hilbert), también llamado el espacio de características.Hofmann et al. (2008)

2.3.1 Tipos de funciones Kernel (Núcleo)

- Polinomial-homogénea:

$$K(x_i, x_j) = (x_i \cdot x_j)^n$$

- Perceptron:

$$K(x_i, x_j) = ||x_i - x_j||$$

- Función de base radial Gaussiana: separado por un hiperplano en el espacio transformado.

$$K(x_i, x_j) = e^{-\frac{||x_i, x_j||}{2\sigma^2}}$$

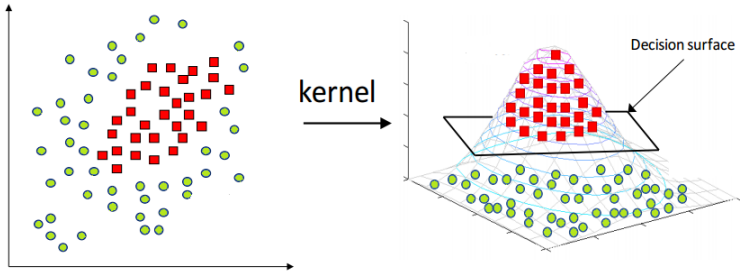


Figura. 2.1: Funcionamiento del Kernel en los SVM. **Fuente:** NewTechDojo (2017)

2.4 Máquinas de Soporte Vectorial de una clase (OCSVM)

Los problema de clasificación de una clase intentan identificar objetos de una clase específica entre todos los objetos, principalmente aprendiendo de un conjunto de datos de entrenamiento que contiene solo los objetos de esa clase.

La clasificación de una clase basada en SVM (OCC) se basa en identificar la hiperesfera más pequeña (con radio r y centro c) que consta de todos los puntos de datos. Formalmente, el problema se puede definir como el siguiente problema de optimización

$$\begin{aligned} \min_{r.c} \quad & r^2 \\ \text{s.t.} \quad & ||\Phi(x_i) - c||^2 \leq r^2 \text{ para todo } 1 \leq i \leq n \end{aligned}$$

Sin embargo, esta formulación es sensible a la presencia de valores atípicos. Es por esto que una formulación más flexible es propuesta

$$\begin{aligned} \min_{r,\, c,\, \xi} \quad & r^2 + \frac{1}{vn} \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & ||\Phi(x_i) - c||^2 \leq r^2 + \xi_i \text{ para todo } 1 \leq i \leq n \end{aligned}$$

donde ξ es una variable de holgura.

La compensación entre el radio de la hiperesfera y el número de muestras de entrenamiento que puede contener la hiperesfera se establece mediante el parámetro $v \in (0, 1)$. Cuando v es pequeño se intentar ubicar más datos en la hiperesfera, mientras que cuando v es cercano a 1 se intenta reducir el tamaño de la hiperesfera.

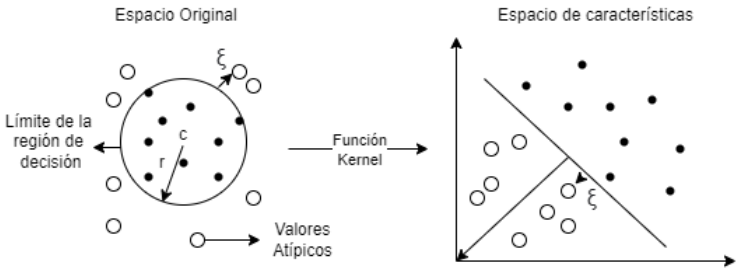


Figura. 2.2: Ejemplo de la funcionalidad del OCSVM.

2.5 K-medias

El método k-medias busca agrupar n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es el más cercano. En palabras sencillas se busca que los elementos de un mismo grupo sean lo más "similares" entre sí y "diferentes" de los elementos de otros grupos.

Dado un conjunto se observaciones $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$ donde cada observación es un vector real de d dimensiones, k-medias construye una partición de las observaciones en k conjuntos $(k \leq n)$ a fin de minimizar la suma de los cuadrados dentro de cada grupo

Sean $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ los conjuntos de datos el planteamiento del problema de optimización es el siguiente

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x_j} \in \S_i} ||\mathbf{x_j} - \mu_\mathbf{i}||^2$$

El problema es computacionalmente difícil (NP-hard) Aloise et al. (2009), sin embargo, existen eficiente heurísticas que convergen rápidamente a un óptimo local.

2.5.1 Algoritmo estándar

El algoritmo estándar que comúnmente se utiliza es una técnica de refinamiento iterativo. Debido a que es el más usado toma el nombre de algoritmo k-medias. MacKay (2003)

Dado un conjunto inicial de k centroides $\mathbf{m}_1^{(1)}, ..., \mathbf{m}_k^{(1)}$, el algoritmo alterna entre dos pasos

1. **Paso de asignación:** En este paso se asigna cada observación al grupo con la media más cercana

   $$S_i^t = \{x_p : ||x_p - m_i^{(t)}|| \leq ||x_p - m_j^{(t)}|| \quad \forall \quad 1 \leq j \leq k\}$$

   Donde cada x_p va exactamente dentro de un $S_i^{(t)}$, incluso aunque pudiera ir en dos de ellos.

2. **Paso de actualización:** Calcular los nuevos centroides como el centroide de las observaciones en el grupo

   $$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

El algoritmo se considera que ha convergido cuando las asignaciones ya no cambian.

2.6 Modelos Lineales Generalizados (GLM)

Puesto que los modelos de regresión lineal ordinaria predicen el valor esperado de una determinada variable respuesta como una combinación lineal de un conjunto de valores observados (predictores), esto conlleva a que un cambio constante en un predictor conduce a un cambio constante en la variable endógena Novales (2010). El uso de estos modelos es adecuado cuando la variable respuesta puede variar indefinidamente en cualquier dirección.

Sin embargo, este supuesto sobre la variable dependiente es inapropiado para el caso de estudio de este documento, puesto que la variable que se busca modelar es de tipo binaria

(presencia/ausencia), es decir la respuesta de la modelización de esta variable debe estar acotada entre 0 y 1.

Los modelos lineales generalizados nacen justamente de la necesidad de flexibilizar la regresión lineal ordinaria y generalizar su uso a distintos tipos de variables endógenas. Estos modelos permiten que las variables de respuesta tengan distribuciones arbitrarias (en lugar de simplemente distribuciones normales), y que una función arbitraria de la variable de respuesta (la función link) varíe linealmente con los predictores (en lugar de suponer que el la respuesta misma debe variar linealmente). Hastie and Pregibon (2017)

Para el caso de estudio se utiliza la función logit como la función link, estos modelos toman el nombre de modelos de regresión logística o logit.

2.6.1 Modelos de regresión logística

Los modelos de regresión logística forma parte de los métodos de regresión paramétricos, estos permiten modelar una variable dependiente categórica en función de las variables independientes que pueden ser cualitativas o cuantitativas, según Hosmer Jr et al. (2013). Como se menciona anteriormente en esta investigación se busca modelar una variable binaria (presencia/ausencia), que con fines prácticos las observaciones con etiquetas de "Presencia" tomarán el valor de 1 y las observaciones etiquetadas como "Ausencia" tomarán el valor de 0.

Considerando que la variable dependiente se distribuye binomialmente de la forma

$$Y_i \sim B(p_i, n_i), \text{ para } i = 1, \ldots, m,$$

donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas.

El modelo particular utilizado por la regresión logística, que la distingue de la regresión lineal estándar y de otros tipos de análisis de regresión utilizados para resultados con valores

binarios, es la forma en que la probabilidad de un resultado particular se vincula con la función predictora lineal, donde las variables explicativas del modelo pueden pensarse como un vector X_i k-dimensional (siendo k la cantidad de variables exógenas) y el modelo toma la forma

$$p_i = E\left(Y_i|X_i\right)$$

con, $E\left(Y_i|X_i\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$. Así, los logits de las probabilidades binomiales desconocidas son modeladas como una función lineal de las variables independientes X_i.

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} \tag{2.1}$$

Es así que la formulación del modelo es

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}$$

Finalmente, basta con estimar los coeficientes β_j del modelo para lo cual usualmente se utiliza el método de máxima verosimilitud Menard (2002).

2.7 Modelos Aditivos Generalizados (GAM)

Los modelos aditivos generalizados (GAM) fueron desarrollados originalmente por Trevor Hastie y Robert Tibshirani para combinar propiedades de modelos lineales generalizados con modelos aditivos.

Estos modelos relacionan una variable de respuesta univariada, Y, con algunas variables predictoras, X_i. Se especifica una distribución de familia exponencial para Y (en el caso de estudio distribución binomial) junto con una función link g (en este caso la función logit) que relacionan el valor esperado de Y con las variables predictoras a través de una estructura como

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

Las funciones f_i pueden ser funciones con una forma paramétrica específica (por ejemplo, un polinomio o una spline de regresión no penalizada de una variable) o pueden especificarse

de forma no paramétrica o semiparamétrica, simplemente como "funciones suaves", para ser estimadas por métodos no paramétricos. Estas relaciones suaves f_i se pueden estimar simultáneamente y luego predecir $g(E(Y)))$ sumándolas (de aquí toman el nombre de modelos aditivos).

Esta flexibilidad para permitir ajustes no paramétricos con suposiciones relajadas sobre la relación real entre la respuesta y el predictor, brinda el potencial para mejores ajustes a los datos que los modelos puramente paramétricos Hastie and Tibshirani (2017).

2.7.1 Modelo GAM logístico

Como se pudo ver anteriormente los modelos GAM son una generalización de los GLM, que permiten capturar relaciones no lineales entre la variable dependiente y cada una de las variables predictoras.

Así, una extensión natural del modelo logístico planteado en la ecuación (2.1) haciendo uso de la teoría de los modelos aditivos generalizados es la siguiente

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + f_1(x_{1,i}) + f_2(x_{2,i}) + \cdots + f_k(x_{k,i}) \tag{2.2}$$

notese que cada término lineal se reemplaza con una forma funcional f_i más general (función suave) y los logits de las probabilidades binomiales se modelan como una función aditiva de los predictores.

Las formas funcionales f_i son las piedras angulares de los modelos GAM, en este caso se hace uso de los splines cúbicos de regresión como las funciones suaves pues son computacionalmente más eficientes, según De Boor (1978) y del algoritmo de maxima verosimilitud restringida (REML) para estimar los parámetros de suavizado de las funciones suaves.

2.7.2 Funciones suaves usando splines cúbicos

Los splines de suavizado adoptan un enfoque completamente diferente para derivar curvas suaves. Estos permiten estimar la función suave minimizando la suma de cuadrados penalizada.

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int (s^{''}(x))^2 dx \tag{2.3}$$

donde la suma residual de cuadrados

$$\sum_{i=1}^{n}(y_i - s(x_i))^2$$

asegura que se ajuste a los datos observados, mientras que el término de penalización

$$\lambda \int (s^{''}(x))^2 dx$$

impone suavidad (es decir, penaliza la ondulación).

Tenga en cuenta que el término de penalización impone suavidad al calcular el cuadrado integrado de las segundas derivadas. Intuitivamente, esto tiene sentido, pues la segunda derivada mide las pendientes de las pendientes por lo tanto, una curva ondulada tendrá segundas derivadas grandes, mientras que una línea recta tendrá segundas derivadas de 0, según De Boor (1978). Es decir, escencialmente "se suman" las segundas derivadas al cuadrado para medir la ondulación de la curva.

El equilibrio entre el ajuste del modelo y la suavidad está controlado por el parámetro de suavizado . La función que minimiza la suma de cuadrados penalizada es un spline cúbico.

Notese que que si $\lambda \to 0$ se interpolarán las observaciones, es decir que la curva será muy ondulada y cuando $\lambda \to \infty$ el ajuste tenderá a una recta (con segunda derivada nula). Por lo tanto, es primordial la estimación del parámetro λ, según Wood (2006).

2.7.3 Máxima verosimilitud restringida (REML)

Como se concluyó en el apartado anterior, la estimación del parámetro λ cumple un papel trascendental en el ajuste del modelo GAM, pues podría no capturarse la relación no lineal que existe entre la variable respuesta y la variable independiente o podría llevar a un problema de sobreajuste del modelo.

Existen dos formas que comúmente se utilizan para estimar λ estas son:

- Validación cruzada o GCV
- Máxima verosimilitud restringida (REML)

En este caso de estudio utilizaremos el método REML puesto que este método es menos propenso a los mínimos locales, Gurka (2006).

Dado que los modelos GAM tienen una interpretación bayesiana, es posible tratarlo como un modelo mixto estándar separando los efectos fijos y estimando los parámetros de suavizado como parámetros de varianza.

Por lo tanto, dado el vector de parámetros de suavizado λ, se obtiene, la función de verosimilitud restringida, integrando β a partir de la densidad conjunta de los datos y los coeficientes

$$l_r(\hat{\beta}, \lambda) = \int f(y|\beta) f(\beta) d\beta$$

La función de verosimilitud restringida depende de λ y de las estimaciones $\hat{\beta}$ (a través de la penalización), pero no de los parámetros aleatorios β. Por lo tanto, se puede usar esta función para derivar vectores de prueba para λ para una iteración anidada de PIRLS (Mínimos cuadrados iterativos reponderados penalizados):

1. Dado un vector de prueba λ, estimar β utilizando PIRLS.

2. Actualice λ maximizando la probabilidad logarítmica restringida.

3. Repita los pasos 1 y 2 hasta la convergencia.

El método de mínimos cuadrados iterativos reponderados penalizados (PIRLS) que es una extensión de los mínimos cuadrados iterativos reponderados (IRLS), ampliamente utilizados para GLM, según Wood (2006), la estimación de los coeficientes a la iteración k está dada por

$$\hat{\beta}_{(k+1)} = (\mathbf{B}^{'} W_{(k)} \mathbf{B} + \mathbf{P})^{-1} \mathbf{B}^{'} W_{(k)} z_{(k)}$$

y continua incrementando k hasta que alcanza la convergencia. Nótese que el subíndice de iteración aplicado a $\mathbf{W}$ y z indica que ambos dependen de las probabilidades estimadas $\hat{p}$ que cambian en cada iteración.

2.8 Regresión Adaptiva Multivariante con Splines (MARS)

Los modelos de regresión adaptiva multivariante con splines (MARS) son una forma de análisis de regresión introducida por Jerome H. Friedman en 1991. Es una técnica de regresión no paramétrica y puede verse como una generalización tanto de la regresión lineal por pasos, como de los árboles de decisión CART.

El modelo MARS es un spline multivariante lineal:

$$m(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x)$$

(es un modelo lineal en transformaciones $h_m(x)$ de los predictores originales), donde las bases $h_m(x)$ se construyen de forma adaptativa empleando funciones hinge, las cuales son de la forma

$$h(x) = (x)_+ = \begin{cases} x & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

y considerando como posibles nodos los valores observados de los predictores,Friedman (1991).

A continuación, se explica el modelo MARS aditivo (sin interacciones), y después se extiende al caso con interacciones. Asumiendo que todas las variables predictoras son numéricas, el proceso de construcción del modelo es un proceso iterativo hacia delante (forward) que empieza con el modelo

$$\hat{m}(x) = \hat{\beta}_0$$

donde $\hat{\beta}_0$ es la media de todas las respuestas, para a continuación considerar todos los puntos de corte (knots) posibles x_{ji} con $i = 1, 2, ..., n$, $j = 1, 2, ..., p$, es decir, todas las observaciones de todas las variables predictoras de la muestra de entrenamiento. Para cada punto de corte x_{ji} (combinación de variable y observación) se consideran dos bases:

$$h_1(x) = h(x_j - x_{ji})$$

$$h_2(x) = h(x_{ji} - x_j)$$

y se construye el modelo

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 h_1(x) + \hat{\beta}_2 h_2(x)$$

La estimación de los parámetros $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ se realiza de la forma estándar como en una regresión lineal, minimizando RSS. De este modo se construyen muchos modelos alternativos y entre ellos se selecciona aquel que tenga un menor error de entrenamiento. En la siguiente iteración se conservan $h_1(x)$ y $h_2(x)$ y se añade una pareja de términos nuevos siguiendo el mismo procedimiento. Y así sucesivamente, añadiendo de cada vez dos nuevos términos. Este procedimiento va creando un modelo lineal segmentado (piecewise) donde cada nuevo término modeliza una porción aislada de los datos originales.

El tamaño de cada modelo es el número de términos (funciones $h_m(x)$) que este incorpora. El proceso iterativo se para cuando se alcanza un modelo de tamaño M, que se consigue

después de incorporar $\frac{M}{2}$ cortes. Este modelo depende de $M+1$ parámetros β_m con $m = 0, 1, ..., M$. El objetivo es alcanzar un modelo lo suficientemente grande para que sobreajuste los datos, para a continuación proceder a su poda en un proceso de eliminación de variables hacia atrás (backward) en el que se van eliminando las variables de una en una (no por parejas, como en la construcción), según Casal et al. (2021). En cada paso de poda se elimina el término que produce el menor incremento en el error. Así, para cada tamaño $\lambda = 0, 1, ..., M$ se obtiene el mejor modelo estimado $\hat{m}_\lambda$.

Generalizando esta idea el modelo general MARS sólo se diferencia de este caso aditivo en que se permiten interacciones, es decir, multiplicaciones entre las variables $h_m(x)$, según Friedman (1991). Para ello, en cada iteración durante la fase de construcción del modelo, además de considerar todos los puntos de corte, también se consideran todas las combinaciones con los términos incorporados previamente al modelo. Esta interacción se puede ver de la siguiente manera

$$\hat{\beta}_{m+1} h_l(x) h(x_j - x_{ji}) + \hat{\beta}_{m+2} h_l(x) h(x_{ji} - x_j)$$

Al igual que λ, también el grado de interacción máxima permitida se considera un hiperparámetro del problema, aunque lo habitual es trabajar con grado 1 (modelo aditivo) o interacción de grado 2. Una restricción adicional que se impone al modelo es que en cada producto no puede aparecer más de una vez la misma variable X_j. A continuación se muestra un algoritmo para la estimación de los parámetros del modelo.

2.8.1 Validación cruzada generalizada (GCV)

En cualquier modelo de regresión lo que se busca es estimar qué tan bien funciona un modelo con datos nuevos, no con los datos de entrenamiento. Sin embargo, estos datos nuevos generalmente no están disponibles en el momento de la construcción del modelo, por lo que se usa GCV para estimar qué rendimiento tendría el modelo con los datos nuevos. La suma de cuadrados residual (RSS) en los datos de entrenamiento es inadecuada para comparar modelos, porque el RSS siempre aumenta a medida que se eliminan los términos MARS. En

otras palabras, si el RSS se usara para comparar modelos, el paso hacia atrás siempre elegiría el modelo más grande, pero el modelo más grande normalmente no tiene el mejor rendimiento de generalización. Es por esto que el GCV penaliza la flexibilidad porque los modelos que son demasiado flexibles modelarán la realización específica de ruido en los datos en lugar de solo la estructura sistemática de los datos, según Casal et al. (2021).

La fórmula del GCV es

$$GCV(\lambda) = \frac{RSS}{(1 - M(\lambda)/n)^2}$$

donde el RSS es la suma de los residuos al cuadrado medida en la data de entrenamiento, n es el número de observaciones y $M(\lambda)$ es el número de parámetros efectivos del modelo, que depende del número de términos más el número de puntos de corte utilizados penalizado por un factor (2 en el caso aditivo sin interacciones, 3 cuando hay interacciones).

La validación cruzada generalizada se llama así porque utiliza una fórmula para aproximar el error que se determinaría mediante la validación de dejar uno fuera. Es solo una aproximación, pero funciona bien en la práctica. Esta técnica fue introducida por Craven y Wahba y Friedman los amplió para el uso en los modelos MARS.

2.9 Kriging ordinario

Kriging se refiere a una familia de algoritmos de regresión lineal de mínimos cuadrados que intentan predecir valores de una variable en ubicaciones donde los datos no están disponibles según el patrón espacial que presentan los datos disponibles. La descripción de la teoría kriging y su aplicación fueron detalladas por Delhomme (1978).

Según Webster and Oliver (2007), el kriging ordinario es la única técnica que considera dos fuentes de información respecto al atributo estas son: la variación y la distancia entre puntos.

La suposición básica es que los datos son realizaciones de una función aleatoria $Z(x) : x \in D$, donde "x" es un índice espacial y "D" es un dominio fijo en un espacio bidimensional. Se hace un supuesto de estacionariedad que consta de dos partes:

1. La media del proceso se considera constante y debe ser una función de la distancia de retraso h. Esto significa que la diferencia esperada entre los puntos x y $x + h$ (donde h es la distancia entre los dos puntos) es cero, es decir,

$$E[Z(x) - Z(x + h)] = 0$$

2. Se supone que la varianza de la diferencia entre dos valores x y $x + h$ depende solo de la distancia h entre los dos puntos, lo que significa que están autocorrelacionados espacialmente y que es más probable que los puntos cercanos tengan valores similares o diferencias similares que los puntos distantes, es decir,

$$Var[Z(x + h) - Z(x)] = 2\gamma(h)$$

   Esta función $\gamma(h)$ se llama semivariograma o variograma Burgess and Webster (1980).

El objetivo del Kriging ordinario es estimar los valores de datos en ubicaciones no muestreadas x_0 usando la información disponible en otras partes del espacio $(x_1, x_2, ..., x_n)$. Esto se puede realizar expresando $Z(x_0)$ como una combinación lineal de los datos $(Z(x_1), Z(x_2), ..., Z(x_n))$:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_0)$$

Los pesos óptimos λ_i se calculan asumiendo que la estimación $Z(x_0)$ y $\hat{Z}(x_0)$ es insesgada, es decir, el valor esperado de las estimaciones es el mismo que el de los datos conocidos. La condición necesaria para que el estimador sea insesgado es $\sum_{i=1}^{n} \lambda_i = 1$.

El semivariograma es un modelo estadístico que representa cómo los datos varían espacialmente en el área de interés. La variación entre puntos se mide mediante la

semivarianza. Combinando pares de datos a una distancia geográfica h, la semivarianza $\gamma(h)$ de la muestra se puede escribir como:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i + h) - Z(x_i)]^2$$

donde $N(h)$ representa al número de pares de puntos separados por la distancia h.

Una vez que se ha calculado la función del semivariograma a partir de los valores muestreados en diferentes ubicaciones, el siguiente paso es ajustar un semivariograma paramétrico $\gamma(h)$. Un método común utilizado para ajustar modelos de semivariogramas paramétricos al semivariograma de muestra es el método de mínimos cuadrados ponderados, propuesto por Cressie (1985).

A continuación, se muestran algunos de los modelos téoricos de semivariograma más usados por el método Kriging.

2.9.1 Modelos de Semivariogramas

Se definen:

- c_0 : nugget, valor que resulta al extrapolar la curva del semivariograma para la distancia cero.
- c_1 : silo o meseta, es el límite del semivariograma cuando la distancia h tiende al infinito
- a : rango, la distancia en la que el valor del variograma alcanza el 95% del valor de la meseta.
- h : distancia de separación

Los modelos de semivariograma que usualmente se utilizan para Kriging son:

- **Modelo esférico:** El modelo de variograma teórico con meseta más utilizado es el modelo esférico, el cual tiene un crecimiento rápido cerca del origen mientras que para distancias superiores al rango los incrementos son nulos, el modelo se expresa de la siguiente forma:

$$\gamma(h) = \begin{cases} c_0 + c_1 \left(1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right) & \text{si } h \leq a \\ c_0 + c_1 & \text{si } h > a \end{cases}$$

- **Modelo exponencial:** Se aplica cuando la dependencia espacial crece exponencialmente respecto a la distancia entre las observaciones, la formulación es la siguiente:

$$\gamma(h) = c_0 + c_1 \left(1 - exp\left(\frac{-3h}{a}\right)\right)$$

- **Modelo gaussiano:** Este modelo tiene meseta y la principal característica de este modelo es su forma parabólica cerca del origen, su fórmula es la siguiente:

$$\gamma(h) = c_0 + c_1 \left(1 - exp\left(\frac{-h^2}{a^2}\right)\right)$$

Estos modelos teóricos de semivariograma son detallados por Giraldo et al. (2007).

2.10 Kriging residual

El kriging residual o kriging con regresión propuesto por Gambolati and Volpi (1979), se utiliza con la finalidad de lidiar con la presencia de tendencia en el valor medio de la variable respuesta, es decir cuando la variable a modelar no es estacionaria.

El procedimiento kriging residual consiste en estimar la tendencia $m(x)$ usando un método de regresión, a partir de esto se obtienen los residuos $r(x)$ del modelo, estos residuos conservan la variabilidad espacial de la variable dependiente, según Odeh et al. (1995), por lo que al aplicar kriging ordinario sobre los residuos se obtiene una función $\hat{r}(x)$ que representa una corrección al modelo de regresión.

Finalmente, la predicción buscada es entonces igual a la tendencia estimada más la predicción del error, la fórmula final $Z^*(x)$ es:

$$Z^*(x_0) = \hat{m}(x_0) + \hat{r}(x_0)$$
$$\hat{r}(x_0) = \sum_{i=1}^{n} \lambda_i r(x_0)$$

2.11 Estadísticos de evaluación y desempeño de los modelos de regresión

2.11.1 Matriz de confusión

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo de clasificación donde la respuesta puede tener dos o más clases. En esta matriz cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. En el caso de una clasificación binaria se tiene una tabla con cuatro combinaciones diferentes de valores predichos y reales tal como se presenta en la figura 2.3.

		Clase predicha	
		Positiva	**Negativa**
Clase real	**Positiva**	TP	FN
	Negativa	FP	TN

Figura. 2.3: Matriz de Confusión

El significado de las entradas de esta matriz es la siguiente:

- **TP:** Son los verdaderos positivos, es decir, el número de veces que el valor real es positivo y la prueba predijo también que era positivo.
- **TN:** Son los verdaderos negativos, es decir, el número de veces que el valor real es negativo y la prueba predijo también que el resultado era negativo.
- **FP:** Son los falsos positivos, el número de veces que la prueba predijo un valor positivo cuando el valor real es negativo.
- **FN:** Son los falsos negativos, el número de veces que la prueba predijo un valor negativo cuando el valor real es positivo.

Esta matriz es muy útil al medir el desempeño de un modelo de clasificación, pues a partir de esta es posible calcular la sensibilidad, especificidad y precisión del modelo.

Sensibilidad

La sensibilidad es un indicativo de la capacidad del estimador para discriminar los casos positivos, de los negativos. Esta también se conoce como la "Tasa de Verdaderos Positivos" y es la proporción de casos positivos que fueron correctamente clasificados por el modelo. Su ecuación es:

$$Sensibilidad = \frac{TP}{TP + FN}$$

Especificidad

La especificidad al igual que la sensibilidad es una métrica de poder de discriminación. Esta es también conocidad como la "Tasa de Verdaderos Negativos" y es la proporción de casos negativos que fueron correctamente clasificados por el modelo. Su ecuación es:

$$Especificidad = \frac{TN}{TN + FP}$$

Precisión

Consiste en la proporción de número de predicciones correctas con el total de predicciones hechas por el algoritmo. Es decir, toma la exactitud de los datos correctos y la compara con el total de datos arrojados (sean estos exactos o no). Su ecuación viene dada de la siguiente forma:

$$Precision = \frac{TP + TN}{TP + TN + FP + FN}$$

2.11.2 Curva ROC

La curva ROC es una herramienta que permite conocer el rendimiento global de un modelo de clasificación, esta es una representación gráfica de la sensibilidad frente a la especificidad para un modelo de clasificación binario. Este gráfico representa los pares (1-especificidad, sensibilidad) obtenidos al considerar todos los posibles valores de corte de la prueba, Lobo et al. (2008).

La curva ROC es necesariamente creciente, propiedad que refleja el compromiso existente entre sensibilidad y especificidad: si se modifica el valor de punto de corte para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir la especificidad. Si el modelo no discriminara entre grupos, la curva ROC sería la diagonal que une los vértices inferior izquierdo y superior derecho. La exactitud de un modelo se refleja en la forma de la curva pues a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo mejor es la dicrminación del modelo. Si la discriminación fuera perfecta (100% de sensibilidad y 100% de especificidad) la curva pasaría por el vértice superior izquierdo López-de Ullibarri (1998).

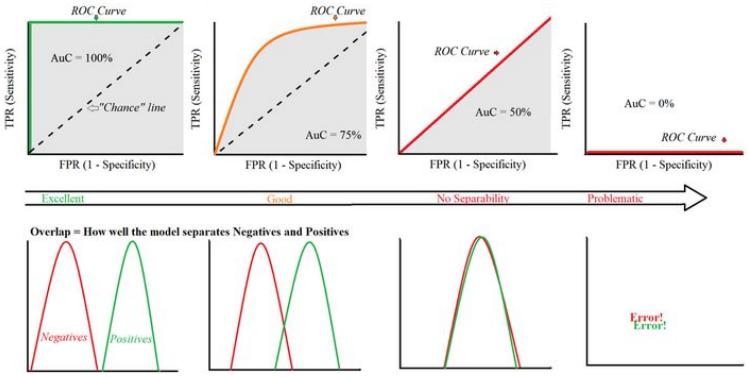


Figura. 2.4: Casos de la curva ROC. **Fuente:** Glen (2019)

Área bajo la curva (AUC)

El valor AUC es la probabilidad de que el modelo clasifique una observación positiva aleatoria más alto que una observación negativa aleatoria. Este valor es justamente como su nombre lo dice el cálculo del área bajo la curva ROC. El AUC toma valores entre 0 y 1. Un modelo cuyas predicciones son 100% incorrectas tiene un AUC de 0 y uno cuyas predicciones son 100% correctas tiene un AUC de 1.

Para los modelos de clasificación lo esperado es tener valores de AUC por encima del 0.7 y entonces se aceptará que estos modelos tienen una buena capacidad de discriminación, según Moran (1948).

2.11.3 Estadístico de Kolmogorov-Smirnov (KS)

El estadístico KS es una de los más utilizados en la evaluación de modelos de lasificación binaria, esta es una medida no paramétrica que se usa para evaluar la "bondad de ajuste" de entre curvas, haciendo uso de las distribución empírica acumulada (ECDF).

En palabras simples el estadístico de Kolmogorov-Smirnov se puede entender como la mayor distancia entre las ECDF de cada muestra. Sean $F_{n1}(x)$ y $F_{n2}(x)$ dos ECDF de la muestra de presencias acumuladas y la muestra de ausencias acumuladas respectivamente, se

debe calcular su máxima diferencia absoluta sobre todos los valores x, es decir,

$$D = \max_{x} \quad |F_{n1}(x) - F_{n2}(x)|$$

Mientras mayor sea el valor de KS mejor discriminación presenta el modelo pues se puede decir que las distribuciones de los eventos (presencia/ausencia) son diferentes, usualmente valores de KS mayores a 0.4 son aceptables.

2.11.4 Coeficiente Gini

El coeficiente de Gini es una métrica que indica el poder de discriminación del modelo, es decir, la eficacia del modelo para diferenciar las clases de un modelo binario. En el contexto de los modelos de clasificación, el coeficiente Gini mide la relación ordinal entre las predicciones de los modelos y el resultado real. Si el modelo es útil, las probabilidades bajas deberían estar más asociadas con la ausencia del evento y las probabilidades altas con la presencia del evento, según Anderson (2007).

El cálculo de este estadístico se puede derivar de la curva ROC vista en el apartado 2.9.2. pues como demuestra Schechtman (2016) existe una relación lineal entre el valor de Gini y el AUC, esta relación es de la siguiente forma:

$$Gini = 2 * AUC - 1$$

Esta métrica se usa a menudo para comparar la calidad de diferentes modelos y evaluar su poder de predicción, usualmente un modelo con valor de Gini mayor a 0.6 es considerado como bueno.

Capítulo 3

Metodología y Resultados

Este capítulo se centra en describir a detalle la metodología utilizada en la construcción del modelo de distribución de las ranas aposemáticas en el Ecuador continental, además se muestran los resultados obtenidos en términos de poder de predicción y discriminación entre presencias/ausencias de ranas Dendrobatidaes.

Observación 1: En esta sección se presentan los resultados de las métricas de poder predictivo, obtenidas con cada uno de los modelos propuestos, con la finalidad de validar los modelos, no obstante, estas no deben ser interpretadas como en un modelo común de clasificación, ya que en este caso se construye la variable objetivo binaria de manera artificial.

Observación 2: En esta sección se presentan predicciones mensuales para la distribución de las ranas aposemáticas en el Ecuador, puesto que según El-Gabbas et al. (2021), un enfoque para modelar la estacionalidad de la distribución de especies es calibrar un modelo estacional (o mensual), dado que para cada temporada, los avistamientos de especies y las condiciones ambientales observadas se utilizan para predecir la idoneidad del hábitat en la temporada respectiva.

3.1 Análisis y tratamiento de la base de datos

En este apartado se presentan los resultados del análisis descriptivo sobre las variables climáticas, ambientales y geográficas que aportan como regresoras del modelo.

Dado que se tienen observaciones de la presencia de ranas aposemáticas entre los años 2012 a 2019, se trabaja con rasters mensuales de las variables regresoras en este mismo lapso

temporal, por lo tanto para cada variable regresora se tienen 84 rasters.

Estos datos se obtuvieron de las bases de datos de la NASA, por lo que ya pasaron por un preproceso y corrección de datos realizado por esta entidad, sin embargo aún existe la presencia de máximo 20% de valores pérdidos en algunos rasters, estos son tratados con la técnica de interpolación con la distancia inversa ponderada, se utiliza el software QGIS3 para la realización de esta interpolación. Como se observa en la figura 3.1, el mapa de calor después de realizar el algoritmo de interpolación mantiene la estructura general del mapa de calor de los datos observados.

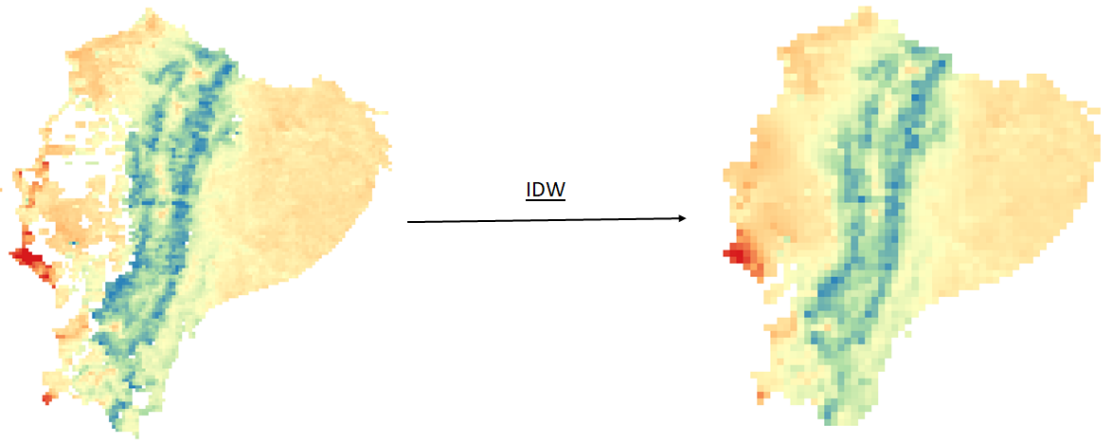


Figura. 3.1: Mapa de calor de variable climática antes y después del IDW

Además, el algoritmo IDW no altera la distribución de los valores de observados de las variables climáticas y ambientales, como se puede observar en la figura 3.2.

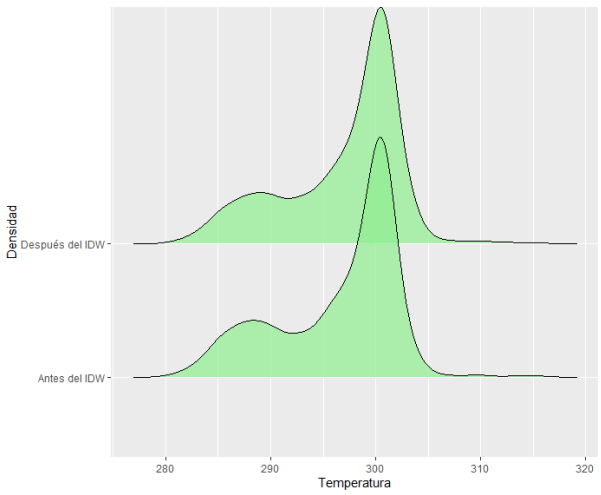


Figura. 3.2: Distribución de variable climática antes y después del IDW

Una vez realizada la imputación de los valores pérdidos de las variables regresoras, se emparejan las ubicaciones de los datos de presencias de ranas con la información climática, ambiental y geográfica correspondiente, a continuación se presenta un resumen de los estadísticos descriptivos de las variables regresoras.

	Temperatura(°K)	NDVI	Elevación(m.s.n.m)	Precipitación(mm)	Evapotranspiración(mm)
Media	298.005142	0.764599	650.456013	195.831649	98.676922
Std	3.701617	0.098068	678.661445	158.731439	37.806251
Min	283.720000	0.352625	3.000000	0.000000	0.285431
25%	296.815002	0.734038	243.000000	41.413392	71.102502
50%	298.680000	0.792400	359.000000	169.703289	111.635870
75%	300.269593	0.829462	962.000000	314.919941	126.159088
Max	307.924988	0.901425	3822.000000	715.640732	167.054088

Tabla. 3.1: Estadísticos descriptivos de las variables regresoras

3.2 Creación de las pseudoausencias

Como se menciona anteriormente, la creación de pseudoausencias para los modelos de distribución de especie nacen de la necesidad de corregir el sesgo y sobreoptimismo que

presentan las predicciones de los modelos que tan solo se ajustan con observaciones de presencias de la especie.

Los modelos que incluyen observaciones de ausencia de especies presentan mejores resultados en términos de predicción (Engler et al. (2004), Chefaoui and Lobo (2008), Hengl et al. (2009)), sin embargo, los datos de ausencia de las especies suelen ser muy limitada y por lo tanto inexistente en la mayoría de las bases de datos biológicas, como lo indica Araújo and Williams (2000), es por esto que varias técnicas de generación de pseudo-ausencias se han incorporado con la finalidad de generar modelos más robustos y predicciones más precisas, mucha de esta metodología utiliza justamente las llamadas técnicas de perfil (técnicas que tan solo utilizan los datos de presencia) para crear un primer modelo de la distribución de la especie y a partir de esto crear las pseudoausencias.

En el presente documento se usa la metodología tres pasos con selección k-medias (TSKM) propuesto por Iturbide et al. (2015). Este proceso para la generación de pseudo-ausencias busca estimar la distribución del nicho fundamental[1] de la especie mediante la introducción de pseudo-ausencias dentro del espacio del nicho correspondiente a áreas fuera del nicho realizado[2] de la especie.

Observación: Entre las técnicas de perfil más conocidas se encuentra el Análisis Factorial de Nicho Ecológico (ENFA) el cual permite perfilar el nicho ecológico en el que se encuentra presente la especie a través de las similitudes que presenta el ambiente en el que vive la especie con el ambiente general del espacio geográfico que se está utilizando en la modelización. Sin embargo, en este documento no se hace uso de esta técnica ya que en la actualidad existen metodologías que permiten una mayor flexibilidad y presentan mejores predicciones como concluye Munoz-Mari et al. (2007).

[1]El nicho fundamental corresponde a las condiciones físicas bajo las cuales una especie podría vivir, en ausencia de interacciones con otras especies.

[2]El nicho realizado es el nicho real u observado de una especie.

3.2.1 Análisis de componentes principales

Antes de comenzar con la modelización de la máquina de soporte vectorial de una clase se realiza un paso previo con ACP de los cuales se extraen las primeras dos componentes principales, las cuales explican el 77.2% de la variabilidad de los datos como se puede ver en la figura 3.3.

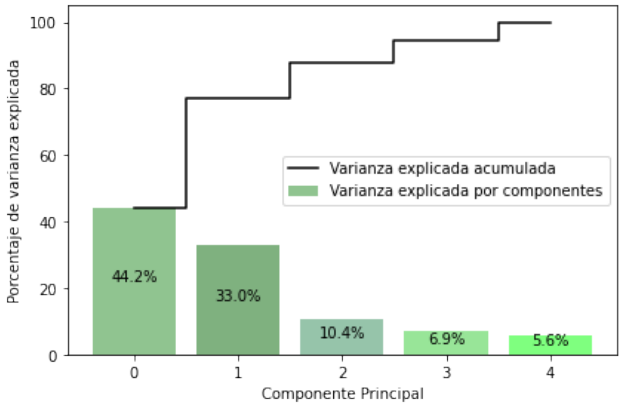


Figura. 3.3: Representación de la varianza explicada

En la figura 3.4 se presenta un biplot que es una representación bidimensional de las primeras dos componentes, en este gráfico es posible notar la dirección de crecimiento que tienen las variables dependientes en las nuevas componentes, además de que es posible notar que las variables precipitación, evapotranspiración y NDVI aportan en gran proporción a la componente 0 del ACP, y la variable temperatura aporta en la componente 1. Además, también es posible observar la correlación que existe entre la variable evapotranspiración y NDVI.

Figura. 3.4: Representación biplot sobre las dos primeras componentes principales

Las ventajas de este proceso antes de realizar el OCSVM son las siguientes:

1. Permite reducir la complejidad computacional y por lo tanto el tiempo de ejecución del algoritmo según Liang et al. (2022), considerando que el algoritmo OC-SVM tienen alta complejidad algorítmica y amplios requisitos de memoria debido al uso de programación cuadrática, Pedregosa et al. (2011). En el caso de uso de este documento y con las herramientas de hardware disponibles el tiempo de ejecución del OC-SVM se redujo significativamente de 4.7 minutos antes de realizar el ACP a 1 minuto después de realizar el ACP.

2. Al reducir la dimensión del problema se logra una mejor visualización de los resultados de la máquina de soporte vectorial, pues nos permite crear curvas de nivel para observar la frontera que produce la máquina de soporte vectorial.

3. Finalmente, realizar un ACP antes de utilizar el algoritmo OC-SVM ayuda a mejorar el resultado de las predicciones que se obtienen con este pues mejora la precisión y la eficiencia del entrenamiento, además de preservar las características de los datos iniciales, como lo sugiere Sundarkumar and Ravi (2013) y Yu et al. (2014), estos beneficios se deben principalmente a que corrige el problema de dimensionalidad y multicolinealidad de los datos.

3.2.2 Máquina de soporte vectorial de una clase

Pues bien, ahora se entrena un modelo solo para las presencias de ranas utilizando máquinas de soporte vectorial de una clase y el kernel de base radial Gaussiano para la estimación, la base se divide en base de entrenamiento (75%) y validación (25%).

Este modelo es un primer acercamiento a la distribución de la especie, y brinda una visión sobreoptimista de la presencia de la especie en la región que se estudia. En la figura 3.5 se puede observar la frontera del hiperplano creado y los puntos tanto de la base de entrenamiento como de la base de validación que quedan dentro y fuera del hiperplano.

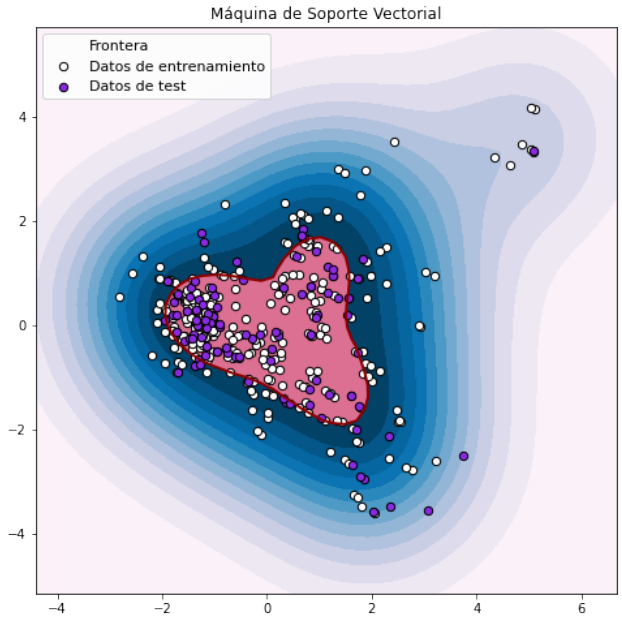


Figura. 3.5: Curva de nivel del hiperplano estimado con el OCSVM

Observación 1: La precisión del modelo OCSVM que se entrenó es de 71%, entendiendo que la precisión en este caso no tiene el mismo significado que en un modelo usual de clasificación.

Observación 2: Se creó un estimador bootstrap para medir la estabilidad de las predicciones del modelo ante diferentes bases de entrenamiento, siendo el valor de este

estimador del 95.4% por lo que se puede concluir que este modelo produce estimaciones estables de la distribución de la especie.

3.2.3 Algoritmo K-medias

Luego de aplicar el modelo OCSVM sobre todas las mallas temporales se obtienen las predicciones de presencias/ausencias. En cada malla temporal se filtran las predicciones de ausencias, lo cual otorga un área inhabitable para la especie. Sobre este espacio ambiental y geográfico se realiza un agrupamiento utilizando el algoritmo k-medias, para este proceso se define k (el número de clústeres) igual al número de puntos de presencia en cada temporalidad, y se retienen los valores de las coordenadas de cada centroide del grupo que vendrán a definirse como las "pseudo-ausencias", por lo tanto se obtiene una distribución regular de puntos disímiles en el entorno geográfico y ambiental del estudio que constituye una muestra representativa del entorno inadecuado para la especie, según Senay et al. (2013).

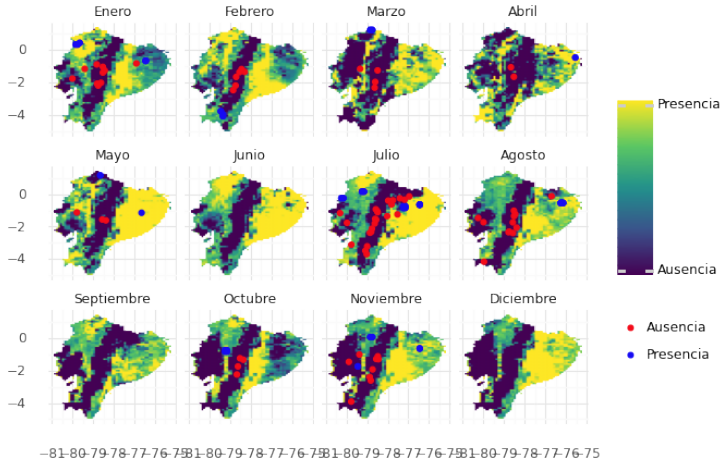


Figura. 3.6: Predicciones y pseudoausencias creadas con el OCSVM para el año 2016

En la figura 3.6 se puede observar el mapa de distribución del nicho potencial que se obtiene con la OCSVM, se aprecia que en la región Sierra del Ecuador es donde se encuentran las áreas inhabitables para la especie, del gráfico también se observa que los puntos rojos, que vendrían a ser las pseudoausencias creadas, se ubican mayoritariamente en la región Sierra.

Como se puede observar, el hábitat adecuado de la especie se encuentra en las zonas cálidas, tropicales y subtropicales del Ecuador, estas predicciones obtenidas con la OCSVM son coherentes con la descripción del hábitat adecuado de la espcie que Grant et al. (2006) da sobre la familia Dendrobatidae.

Los gráficos de todas las capas temporales se encuentran en el Apéndice B.1.

3.3 Modelos de Regresión

Una vez que se crean las pseudoausencias, finalmente se obtiene una variable dependiente binaria (presencia/ausencia) que puede ser modelada por un modelo de clasificación. En esta sección se presentan tres modelos (GLM/GAM/MARS) de clasificación que permiten estimar mapas de distribución potencial de las especie, pues como menciona Chefaoui and Lobo (2008) los modelos de clasificación que usan pseudoausencias y presencias presentan mejores resultados en términos de predicción que las técnicas de perfil (modelos que solo utilizan presencias).

3.3.1 Modelo de regresión lineal logístico

Puesto que no se tiene una gran cantidad de variables exógenas se hace uso del algoritmo "backward" para seleccionar el modelo logístico, esta técnica consiste en que, a partir de un modelo inicial que incluya todas las variables independientes, en cada iteración del algoritmo se va eliminando una variable utilizando el criterio de Akaike, hasta que se obtenga un modelo parsimonioso.

	Coeficiente	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	2.386e+02	7.027e+01	3.395	0.000685	***
Año	-8.794e-02	3.245e-02	-2.710	0.006732	**
Latitud	1.880e-01	5.640e-02	3.333	0.000858	***
NDVI	7.692e+00	1.047e+00	7.345	2.06e-13	***
Temperatura	-2.134e-01	3.166e-02	-6.739	1.60e-11	***
Elevación	-2.575e-03	1.827e-04	-14.094	<2e-16	***
Precipitación	-5.761e-03	4.993e-04	-11.537	<2e-16	***
Evapotranspiración	5.281e-03	2.941e-03	1.796	0.072561	.

Tabla. 3.2: Modelo obtenido usando el algoritmo backward

Una vez que se obtiene el modelo con el algoritmo backward se puede notar que el p-valor de la variable evapotranspiración es ligeramente mayor a 0.05 que es el valor usual utilizado para la prueba de significancia, además, se comprueba que no existe afectación al poder de discriminación del modelo y se elimina esta variable. De esta forma, el modelo obtenido es el siguiente:

	Coeficiente	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	2.382e+02	7.032e+01	3.388	0.000704	***
Año	-8.769e-02	3.245e-02	-2.702	0.006888	**
Latitud	1.788e-01	5.624e-02	3.180	0.001472	**
NDVI	8.885e+00	8.163e-01	10.885	<2e-16	***
Temperatura	-2.153e-01	3.175e-02	-6.782	1.19e-11	***
Elevación	-2.605e-03	1.831e-04	-14.231	<2e-16	***
Precipitación	-5.417e-03	4.575e-04	-11.842	<2e-16	***

Tabla. 3.3: Modelo de regresión lineal logístico (GLM)

En la tabla 3.3 se observa que todas las variables seleccionadas para el modelo son

estadísticasmente significativas y realizando un breve análisis de los coeficientes estos son consistentes con los supuestos sobre el nicho ecológico de la especie. Sin embargo, es necesario realizar un análisis de multicolinearidad del modelo, pues en caso de existir se presentaría problemas en la estimación de los parámetros.

Para calcular el grado de multicolinearidad es usual utilizar el Factor de Inflación de la Varianza (VIF), los resultados se presentan en la tabla 3.4, se concluye que no existe un problema de multicolinearidad puesto que los valores del VIF son menores a 10, que es considerado como el valor máximo aceptado para el VIF según Montgomery (2017).

Variable	**VIF**
Año	1.424386
Latitud	1.236115
NDVI	1.625783
Temperatura	5.696177
Elevación	5.429635
Precipitación	2.018228

Tabla. 3.4: VIF Modelo de regresión lineal logística (GLM)

Evaluación del modelo

Con la finalidad de realizar la validación del modelo planteado, en la tabla 3.5 se presentan varios estadísticos que permiten medir el poder de predicción del modelo.

Partición	**KS**	**AUC**	**Gini**
Entrenamiento	0.6789	0.8796	0.7592
Validación	0.7056	0.8865	0.7731

Tabla. 3.5: Medidas de discriminación del modelo GLM

De la tabla 3.5 es posible notar que los valores de los estadísticos para la data de entrenamiento y validación son muy buenos pues se tiene un valor de KS por encima del 0.6, valores de AUC por encima del 0.8 y valores de Gini por encima del 0.7, que son umbrales usualmente utilizados para determinar que un modelo tiene un buen poder de discriminación. Estos resultados se puede ver graficamente en la figura 3.7 donde se presenta la curva ROC tanto para la base de entrenamiento como para la de validación, que presenta una forma adecuada, lo que sugiere que el modelo tiene un buen poder de predicción.

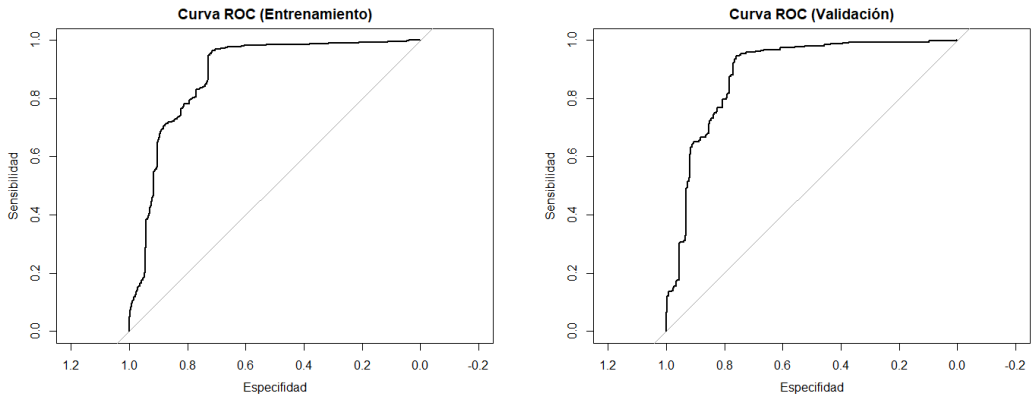


Figura. 3.7: Curva ROC del modelo GLM

Haciendo uso de la base de entramiento se escoge un valor óptimo de punto de corte que es aquel que en la curva ROC maximiza la distancia entre la curva y la diagonal. Este punto se estima en 0.4210 por lo tanto, utilizando este valor se construye la matriz de confusión, la cual se presenta en la tabla 3.6.

	Real	
Predicción	**0**	**1**
0	865	46
1	348	1157

(a) Matriz de confusión de la base de entramiento

	Real	
Predicción	**0**	**1**
0	223	16
1	74	291

(b) Matriz de confusión de la base de validación

Tabla. 3.6: Matrices de confusión del modelo GLM

De la matriz de confusión se puede intuir que el modelo clasifica bien tanto para la data de entramiento como para la de validación, sin embargo, es necesario calcular ciertas métricas que se pueden obtener de esta matriz de confusión, estas se presentan en la tabla 3.7.

Partición	**Sensibilidad**	**Especificidad**	**Precisión**
Entrenamiento	0.9618	0.7131	0.8369
Validación	0.9479	0.7508	0.851

Tabla. 3.7: Métricas de poder de discriminación del modelo GLM

De la tabla 3.7 se tiene que la sensibilidad (porcentaje de presencias clasificados correctamente) es muy buena tanto en la base de entrenamiento como en la base de validación, sin embargo, la especificidad del modelo (porcentaje de ausencias clasificadas correctamente) no es tan buena como la sensibilidad, es decir que el modelo tiene mayor probabilidad de equivocarse cuando predice una ”ausencia”.

Finalmente, es posible aceptar este modelo como adecuado pues los estadísticos

presentados para la validación indican que este tiene un buen desempeño y poder de discriminación. Por lo tanto, en la figura 3.8 se muestra el mapa de calor de la distribución potencial de la ranas dendrobatidaes en el Ecuador.

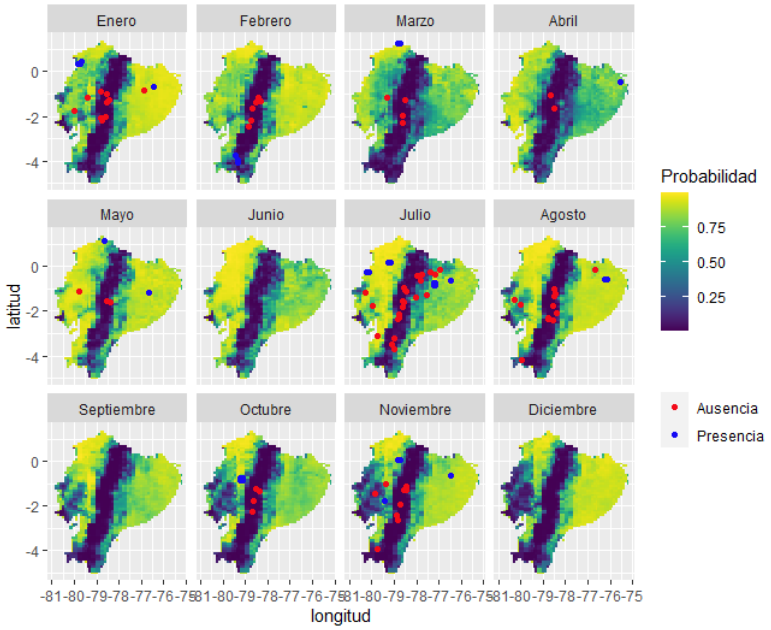


Figura. 3.8: Distribución potencial de las ranas Dendrobatidaes en el Ecuador en el año 2016 (GLM)

Los mapas de calor de todos los años estudiados se encuentran en el Apéndice B.2.

3.3.2 Modelo de Regresión Aditivo Generalizado Logístico

Se comienza esta modelización escogiendo las mismas variables independientes que se seleccionaron con el algoritmo backward en el modelo GLM, se utilizan estas variables preseleccionadas por el modelo anterior para evitar un posible sobreajuste del modelo aditivo generalizado. Sin embargo, esta vez se aplican funciones suaves sobre ciertas variables regresoras continuas, lo cual permite incluir en el modelo las relaciones no lineales que existe entre la variable dependiente y las variables independientes.

Se estiman funciones suaves sobre las variables temperatura, latitud y precipitación, no se

Coeficientes paramétricos					
	Coeficiente	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	3.019e+02	8.093e+01	3.730	0.000191	***
Año	-1.492e-01	4.001e-02	-3.730	0.000192	***
NDVI	2.235e+00	1.113e+00	2.009	0.044573	*
Elevación	-2.389e-03	2.053e-04	-11.635	<2e-16	***
Importancia aproximada de los términos suaves					
	edf	**Ref.df**	**Chi.sq**	**p-value**	
s(Latitud)	7.258	9	183.6	<2e-16	***
s(Temperatura)	6.936	9	252.6	<2e-16	***
s(Precipitación)	6.562	9	249.0	<2e-16	***

Tabla. 3.8: Modelo logístico aditivo generalizado (GAM)

estiman splines para las variables elevación y NDVI puesto que por la descripción del hábitat de la especie se asume que la relación entre la presencia de ranas Dendrobatidaes y estas variables es lineal. Los resultados del modelo obtenido se presentan a continuación:

En la tabla 3.8 se observa que todas las variables del modelo GAM son estadísticamente significativas, incluso aquellas que fueron suavizadas con una función suave. Sobre los coeficientes paramétricos es posible decir que tienen coeherencia con lo expuesto sobre el hábitat de la ranas Dendrobatidaes, pues a mejor calidad de la vegetación (mayor valor del índice NDVI) mayor probabilidad de tener una presencia de la rana y a altitudes más altas la probabilidad de presencia de la rana disminuye.

La desventaja, de este modelo es que no es posible realizar un análisis de la congruencia de los signos de los coeficientes de las variables regresoras suavizadas, pues los coeficientes mostrados en la tabla 3.8 son coeficientes de una forma funcional de las variables, pero no de las variables; por lo que se pierde interpretabilidad en el modelo. En cambio, se realiza un

análisis de las curvas de predicción parcial de las funciones suaves de las variables regresoras.

Los gráficos de las curvas de predicción parcial que se presentan a continuación son una representación de las funciones suaves predichas por GAM de la variable de respuesta presencia/ausencia de ranas en función de las variables explicativas. Los grados de libertad para los ajustes no lineales están entre paréntesis en el eje y, las marcas sobre el eje x indican la distribución de las observaciones (presencia/ausencia) y las zonas sombreadas representan los intervalos de confianza del 95% de las funciones spline suaves.

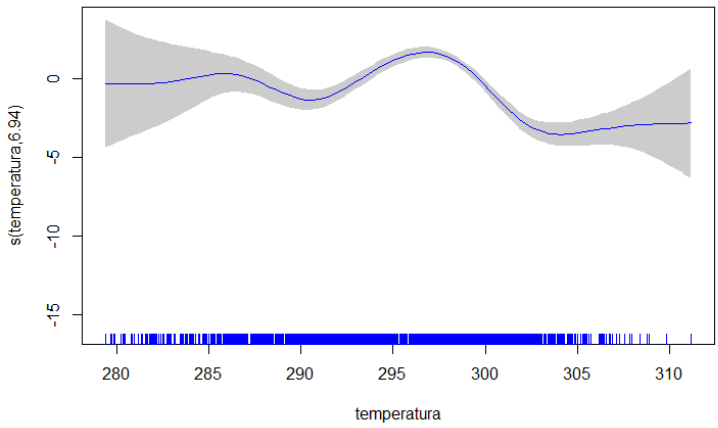


Figura. 3.9: Curva de predicción parcial de la variable Temperatura

En la figura 3.9 se aprecia que, en efecto, se tiene una relación no lineal entre la variables temperatura y la variable objetivo por lo que se justifica el uso de splines, además se alcanza a distinguir que la probabilidad de tener presencia de la ranas Dendrobatidaes es mayor si la temperatura se encuentra entre los $295°K - 300°K$; y la probiblidad disminuye si la temperatura se encuentra cercana a los extremos fríos y calientes.

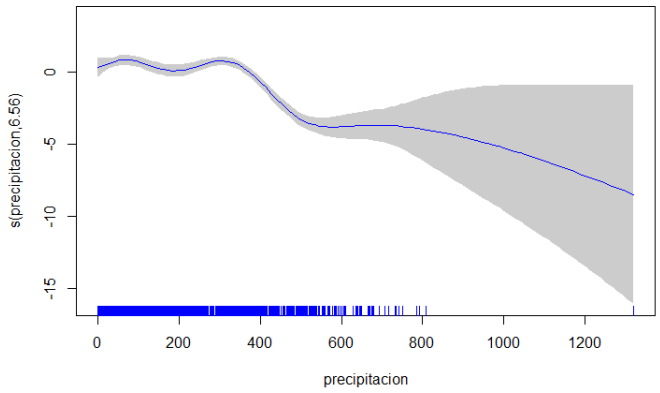


Figura. 3.10: Curva de predicción parcial de la variable Precipitación

En la figura 3.10 nuevamente se percibe que la relación entre la variable respuesta y la variable independiente precipitación es efectivamente no lineal. En esta variable se observa claramente que el exceso de lluvias provoca una fuerte disminución en la probabilidad de presencia de ranas.

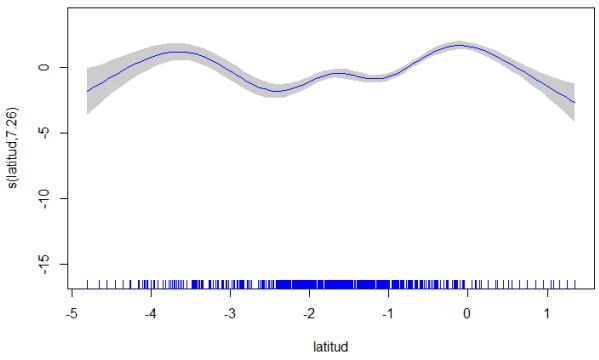


Figura. 3.11: Curva de predicción parcial de la variable Latitud

En la figura 3.11, el caso de la variable latitud, se aprecia que el uso de splines es útil en la modelización por la relación que existe entre la variable Latitud y la variable respuesta. En esta variable se percibe que en las zonas cercanas a la latitud 0 es donde se aumenta la probabilidad de presencia.

Finalmente, es posible indicar que el modelo está bien específicado pues la interpretación de las variables es coherente con la definición del nicho ecológico potencial de la familia de especies de ranas Dendrobatidaes, que describen autores como Twomey and Brown (2008).

Por otro lado, se conoce que las variables no presentan multicolinearidad, sin embargo, en los modelos GAM como se ajustan funciones suaves, es necesario determinar si la función suave de una variable se puede producir usando una combinación de las funciones suaves de los otros términos en el modelo, es decir, es necesario verificar la existencia de la concurvidad.

La concurvidad puede entenderse como la forma no lineal de la colinealidad, como lo propone Buja et al. (1989). Para la concurvidad se calcularán tres índices todos se basan en la idea de que un término suave, f, en el modelo se puede descomponer en una parte, g, que se encuentra completamente en el espacio de uno o más términos en el modelo, y una parte restante que es completamente dentro del propio espacio del término. Si g constituye una gran parte de f, entonces hay un problema de concurvidad. Todos los índices utilizados se basan en el cuadrado de $\frac{||g||}{||f||}$, que es la relación de las normas euclidianas al cuadrado de los vectores de f y g evaluados en los valores de las covariables observadas.

A continuación se presentan tres medidas para la concurvidad estas son

- **worst:** Este es el mayor valor que el cuadrado de $\frac{||g||}{||f||}$ podría tomar para cualquier vector de coeficientes, siendo esta una medida bastante pesimista de la concurvidad.
- **observed:** Esto simplemente devuelve el valor del cuadrado de $\frac{||g||}{||f||}$ según los coeficientes estimados, está es una medida optimista de la concurvidad.

- **estimate:** Este es el cuadrado de la norma de Frobenius de la base de g dividida por la norma de Frobenius de la base de f. Es una medida de hasta qué punto la base f puede ser explicada por la base g. No posee el pesimismo o el exceso de optimismo de las dos medidas anteriores, pero es menos sencillo de entender.

Aún cuando se presentan todos los cálculos, el valor usualmente aceptado como valor máximo de concurvidad es 0.8 para el peor caso (worst), pero, en la tabla 3.9 se observa que el valor de la concurvidad para el término $s(temperatura)$ es mayor a 0.8, en este caso, Wood (2006) indica que si existe convergencia en la estimación de los parámetros del GAM es posible sentir seguridad de los resultados del modelo aún en presencia de la concurvidad.

	para	s(latitud)	s(temperatura)	s(precipitacion)
worst	0.9999993	0.4199983	0.8623521	0.6295651
observed	0.1475647	0.1475647	0.3428618	0.3564683
estimate	0.2336208	0.2336208	0.7016593	0.4428470

Tabla. 3.9: Concurvidad del modelo GAM

Puesto que, el modelo no tuvo problemas de convergencia y la interpretación de los términos es adecuada, se procede a realizar la evaluación del comportamiento de este.

Evaluación del modelo

Con el objetivo de evaluar el modelo se hace uso de los mismos estadísticos utilizados en el modelo GLM.

En la tabla 3.10 se aprecia que tanto para la base de entramiento como la base de validaión los estadísticos son buenos, pues todos superan el umbral que marca un modelo ineficiente de un modelo con buen poder de discriminación.

Partición	**KS**	**AUC**	**Gini**
Entrenamiento	0.7967	0.9502	0.9004
Validación	0.7945	0.9457	0.8915

Tabla. 3.10: Medidas de discriminación del modelo GAM

Nuevamente, en la figura 3.12 se muestran las curvas ROC de la base de entrenamiento y de la base de validación que permite observar gráficamente la eficacia del modelo planteado.

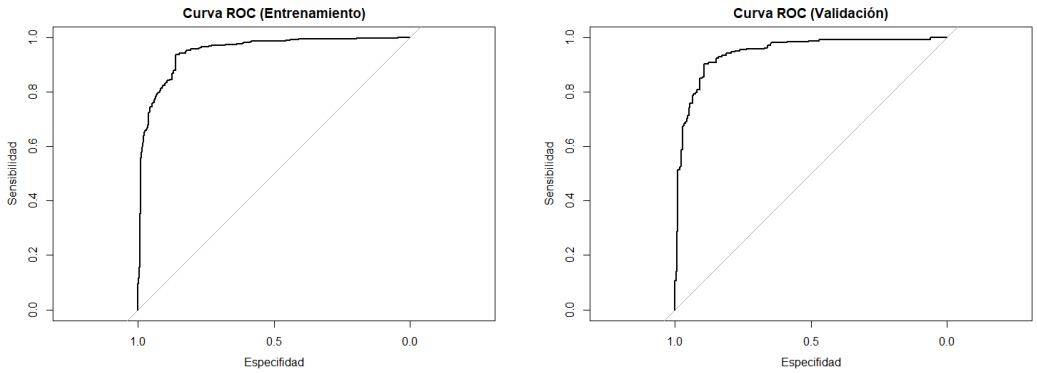


Figura. 3.12: Curva ROC del modelo GAM

Se estima el punto de corte óptimo usando la base de entrenamiento de la misma forma en como se realizó para el modelo GLM, en este caso, el punto de corte óptimo es 0.46888, usando este valor se procede a armar la matriz de confusión, esta se presenta en la tabla 3.11.

	Real	
Predicción	**0**	**1**
0	1041	77
1	172	1126

(a) Matriz de confusión de la base de entramiento

	Real	
Predicción	**0**	**1**
0	263	30
1	34	277

(b) Matriz de confusión de la base de validación

Tabla. 3.11: Matrices de confusión del modelo GAM

A continuación, en la tabla 3.12 se presentan las métricas que se desprenden de las matrices de confusión

Partición	**Sensibilidad**	**Especificidad**	**Precisión**
Entrenamiento	0.9360	0.8582	0.8969
Validación	0.9023	0.8855	0.894

Tabla. 3.12: Métricas de poder de discriminación del modelo GAM

De la tabla 3.12 se observa que tanto la sensibilidad como la especificidad del modelo son buenas, tanto para la base de entrenamiento como para la base de validación.

De esta manera, se procede a mostrar, en la figura 3.13, el mapa de calor de la distribución potencial de la ranas Dendrobatidaes en el Ecuador utilizando el modelo GAM.

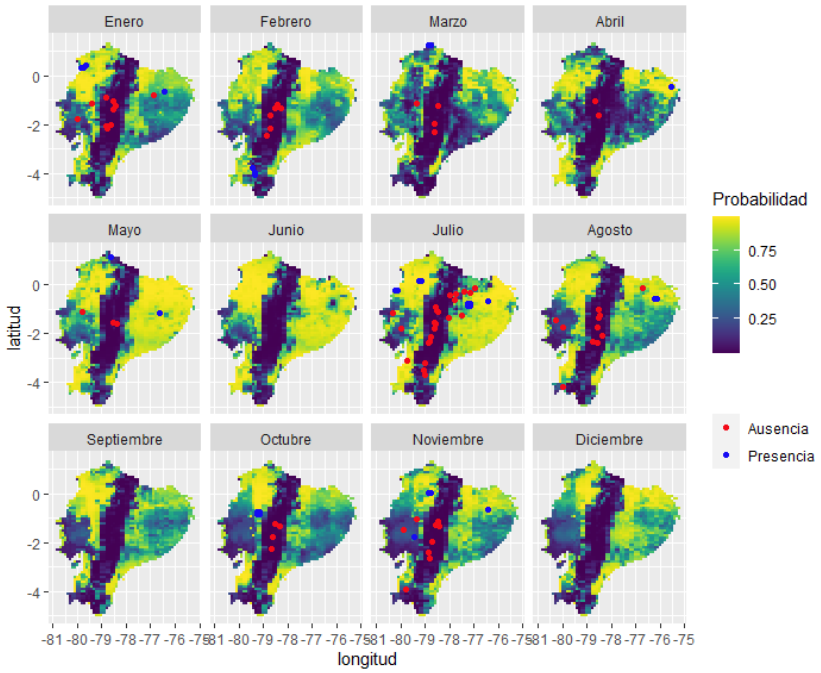


Figura. 3.13: Distribución potencial de las ranas Dendrobatidaes en el Ecuador en el año 2016 (GAM)

Los mapas de calor de todos los años estudiados se encuentran en el Apéndice B.3.

3.3.3 Modelo de Regresión Adaptiva Multivariante con Splines

Para el entrenamiento de este modelo se usan en principio todas las variables regresoras disponibles, sin embargo, el modelo obtenido por el algoritmo GCV no fue el idóneo pues los términos que se estimaban presentaba valores de VIF muy superiores al umbral impuesto de 10. Nótese que en este caso no son las variables las que presentan la multicolinearidad, sino más bien las funciones hinge creadas por el modelo. Dado que, este modelo tiende al sobreajuste es de suma importancia evitar que existan términos colineales para que evitar inflar los estadísticos del modelo, es así que es necesario eliminar las variables NDVI, latitud, longitud y evapotranspiración para ajustar el modelo MARS. La ecuación final del modelo se presenta en la tabla 3.13.

Término	**Coeficiente**
(Intercept)	3.2840147
h(temperatura-298.725)	-0.9560454
h(elevacion-521)	-0.0051108
h(elevacion-1806)	0.0041574
h(2018-anio) * h(310.113-precipitacion)	0.0021812
h(precipitacion-310.113) * h(852-elevacion)	-0.0000393
h(precipitacion-514.095) * h(521-elevacion)	0.0000825

Tabla. 3.13: Modelo logístico adaptivo multivariante (MARS)

En la tabla 3.14, se presentan los valores de VIF para los términos del modelo MARS, se aprecia que todos son menores a 10 por lo que se verifica que no existe colinealiedad.

Variable	**VIF**
h(elevacion-521)	9.833610
h(elevacion-1806)	8.920614
h(temperatura-298.725)	1.391943
h(precipitacion-310.113)*h(852-elevacion)	2.637964
h(precipitacion-514.095)*h(521-elevacion)	2.433283
h(2018-anio)*h(310.113-precipitacion)	1.252889

Tabla. 3.14: VIF Modelo de regresión lineal adaptiva multivariante (MARS)

En el modelo MARS no es posible realizar un análisis de significancia de los términos, sin embargo, es posible verificar la importancia de las variables en términos de GCV para el modelo,en este caso la variable más importante para la discriminación entre presencia o ausencia de ranas Dendrobatidaes es la elevación y la que menos aporta al modelo es la variable temportal año, este resultado se presenta en el figura 3.14.

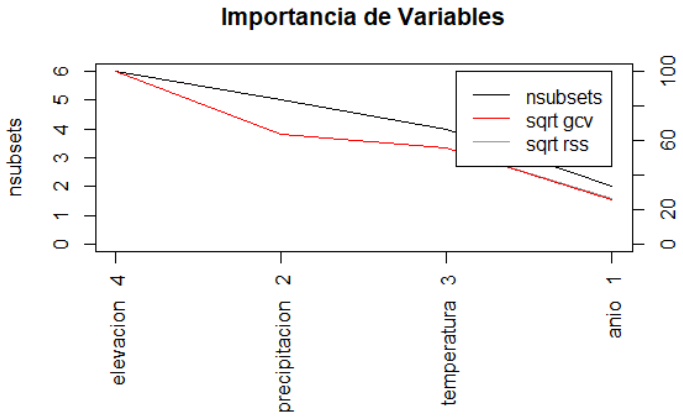


Figura. 3.14: Importancia de las variables en el modelo MARS

La interpretabilidad de este modelo se dificulta por la presencia de los splines, sin embargo, es posible representar gráficamente los efectos de las variables importantes en la probabilidad de presencia, estos efectos se presentan en la figura 3.15. Como se puede observar en la variable temperatura se nota un efecto constante de esta variable en la probabilidad de presencia de la especie hasta el valor de temperatura $298.73°K$, desde aquí la probabilidad de presencia de la rana disminuye dramáticamente, lo mismo sucede para la variable elevación la cual muestra mayor probabilidad de encontrar una rana a altitudes menor a los 1000 m.s.n.m. y probabilidades muy bajas de presencia de la especie a altitudes mayores a los 2000 m.s.n.m..

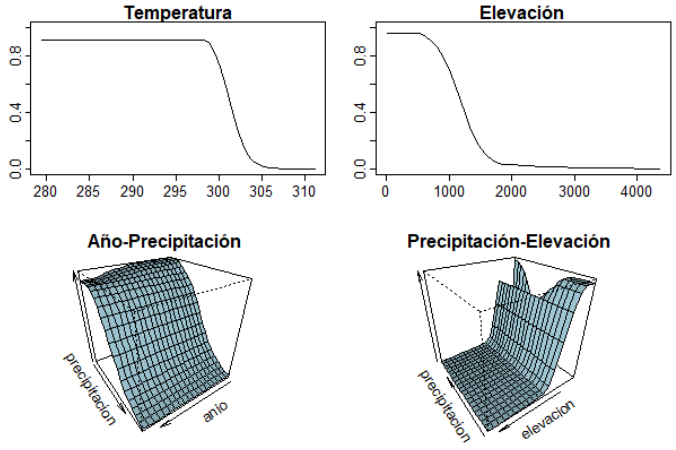


Figura. 3.15: Efecto de las variables en el modelo MARS

Evaluación del modelo

Se realiza la validación del MARS con los mismos estadísticos que se calcularon para los anteriores modelos GLM y GAM.

Partición	**KS**	**AUC**	**Gini**
Entrenamiento	0.7250	0.9428	0.8857
Validación	0.7407	0.9395	0.8791

Tabla. 3.15: Medidas de discriminación del modelo MARS

En la tabla 3.15 se presentan los resultados de los estadísticos de eficacia del modelo, tanto para la base de entrenamiento como para la base de validación estos valores son muy buenos, pues superan los umbrales impuestos.

Se presentan las curvas ROC de la base de entrenamiento y de la base de validación, figura 3.16, que permiten observar gráficamente los resultados obtenidos sobre la eficacia del modelo.

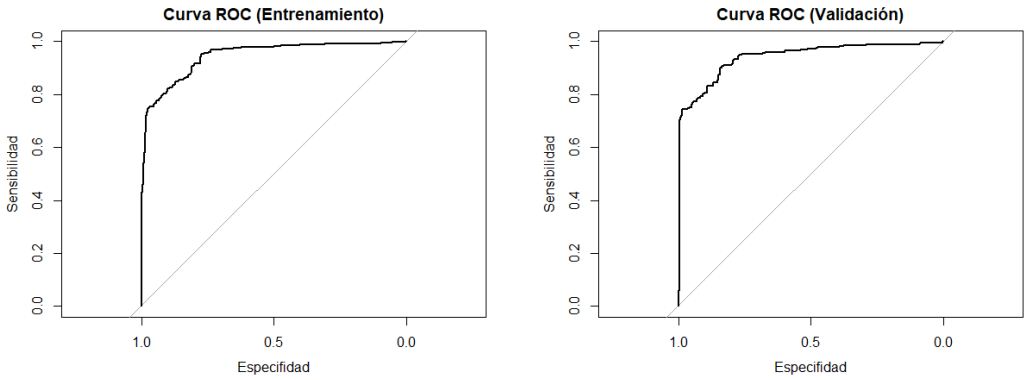


Figura. 3.16: Curva ROC del modelo MARS

Se estima el punto de corte óptimo de la misma forma en como se realizó en los casos anteriores, este valor es aproximadamente 0.7990683, usando este valor se procede a armar

la matriz de confusión, la cual se presenta en la tabla 3.16.

	Real	
Predicción	**0**	**1**
0	1174	298
1	39	905

(a) Matriz de confusión de la base de entramiento

	Real	
Predicción	**0**	**1**
0	293	85
1	4	222

(b) Matriz de confusión de la base de validación

Tabla. 3.16: Matrices de confusión del modelo MARS

En la tabla 3.17 se presentan las métricas de poder de discriminación que se desprenden de la matriz de confusión.

Partición	**Sensibilidad**	**Especificidad**	**Precisión**
Entrenamiento	0.7523	0.9678	0.8605
Validación	0.7231	0.9865	0.8526

Tabla. 3.17: Métricas de poder de discriminación del modelo MARS

En la tabla 3.17 se aprecia que los valores de las métricas son satisfactorias, tanto para la base de entrenamiento como para la base de validación, sin embargo, es posible que se equivoque más en los casos en los que predice una presencia.

Finalmente, se mestra el mapa de calor de la distribución potencial de la ranas Dendrobatidaes en el Ecuador utilizando el modelo MARS, en la figura 3.17.

Los mapas de calor de todos los años estudiados se encuentran en el Apéndice B.4.

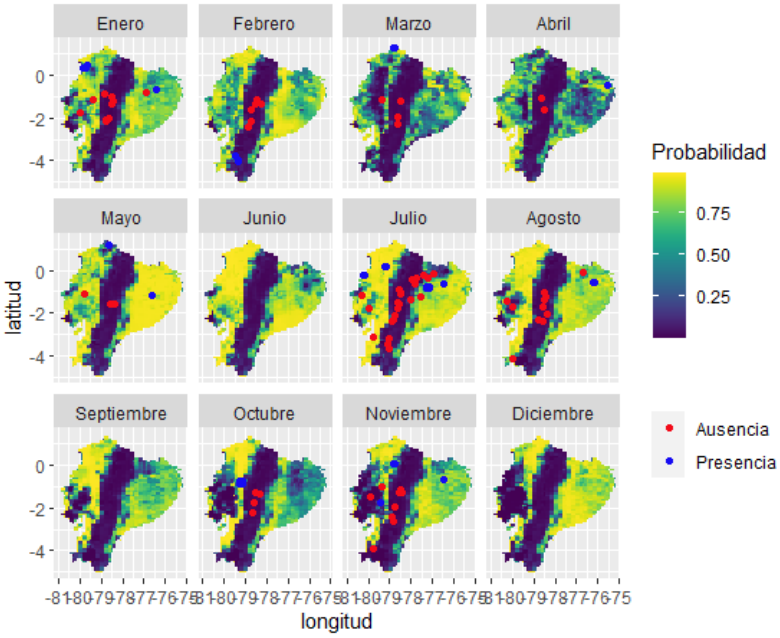


Figura. 3.17: Distribución Potencial de las ranas Dendrobatidaes en el Ecuador en el año 2016 (MARS)

3.4 Elección del mejor modelo de clasificación

Se elige el ”mejor” modelo en base a los estadísticos de discriminación y poder de clasificación. Es decir, que en este caso se llama el mejor modelo al que presenta mejores predicciones que los otros.

En la tabla 3.18 se presenta la comparación de los resultados de los estadísticos KS, AUC y Gini de los tres modelos de clasificación propuestos, como se puede observar el modelo hecho utilizando regresión logística GAM es el que presenta mejores resultados tanto en la base de entrenamiento como en la base de validación.

Modelo	Participación	KS	AUC	Gini
GLM	Entrenamiento	0.6789	0.8796	0.7592
	Validación	0.7056	0.8865	0.7731
GAM	Entrenamiento	0.7967	0.9502	0.9004
	Validación	0.7945	0.9457	0.8915
MARS	Entrenamiento	0.7250	0.9428	0.8857
	Validación	0.7407	0.9395	0.8791

Tabla. 3.18: Comparación de medidas de poder de discriminación

En la tabla 3.19 se aprecia que los modelos que presentan los mejores resultados son el modelo MARS y el GAM, siendo el GAM mejor en identificar las presencias (mayor sensibilidad) y el MARS ligeramente mejor en identificar las ausencias (mayor especificidad), sin embargo, la precisión del modelo GAM es mejor. Es así, que se decide aceptar como el mejor al modelo GAM.

Modelo	Participación	Sensibilidad	Especificidad	Precisión
GLM	Entrenamiento	0.9618	0.7131	0.8369
	Validación	0.9479	0.7508	0.851
GAM	Entrenamiento	0.9360	0.8582	0.8969
	Validación	0.9023	0.8855	0.894
MARS	Entrenamiento	0.7523	0.9678	0.8605
	Validación	0.7231	0.9865	0.8526

Tabla. 3.19: Comparación de métricas de poder de discriminación

3.5 Análisis Kriging Residual

Una vez que se ha seleccionado el mejor modelo en términos de poder predictivo, se modelan los residuos del modelo GAM usando Kriging, esto con la finalidad de abarcar la relación

espacial que existe entre la distribución de las ranas Dendrobatidaes y el espacio.

Dado que se tienen capas de predicciones mensuales de las presencias/ausencias de las ranas, pues existe variabilidad temporal en las variables climáticas, geográficas y ambientales, se realizará un modelo Kriging para cada capa temporal. Sin embargo, debido a que no se tiene la misma cantidad de observaciones de presencia/pseudoausencia en todas las capas temporales, en algunas se presenta un déficit de información para la realización del Kriging residual. Por esto, solo se modelarán con Kriging los residuos de las capas en las que se tengan más de 14 observaciones de presencia/pseudoausencia, pues solo en estas capas temporales se pudo ajustar variogramás teóricos a los datos observados, en el resto de las capas solo se considera la predicción realizada por el modelo GAMS.

Para cada capa temporal, que cumple con la restricción impuesta, el ajuste del variograma se realiza utilizando la librería "*automap*" del software **R**, el algoritmo de la función "*autofitVariogram*" itera sobre los modelos de variograma enumerados en la subsección 2.9.1 y elige el modelo de variograma que tiene la suma de cuadrados residual más pequeña con el variograma empírico.

Finalmente, para las 21 capas que cumplen con la restricción planteada respecto a la cantidad de información disponible, se ajustan variogramas y se realiza la interpolación Kriging. La información de los variogramas modelados se encuentra en el Apéndice C.1.

A continuación se presentan la evaluación del modelo final obtenido con la técnica del Kriging residual.

Evaluación del modelo

A continuación, se realiza la validación del modelo (esta vez solo para la base de test) con los mismos estadísticos que se calcularon para los anteriores modelos de clasificación.

Partición	KS	AUC	Gini
Validación	0.9212	0.9813	0.9626

Tabla. 3.20: Medidas de discriminación del modelo Kriging Residual

En la tabla 3.20 se presentan los resultados de los estadísticos de eficacia del modelo, se observa que efectivamente se presenta una clara mejora de estos estadísticos respecto a los obtenidos para la base de validación utilizando solo el modelo GAM, que se presentan en la tabla 3.10.

Se presentan la curva ROC de la base de validación , figura 3.18, que permite observar gráficamente los resultados obtenidos sobre la eficacia del modelo.

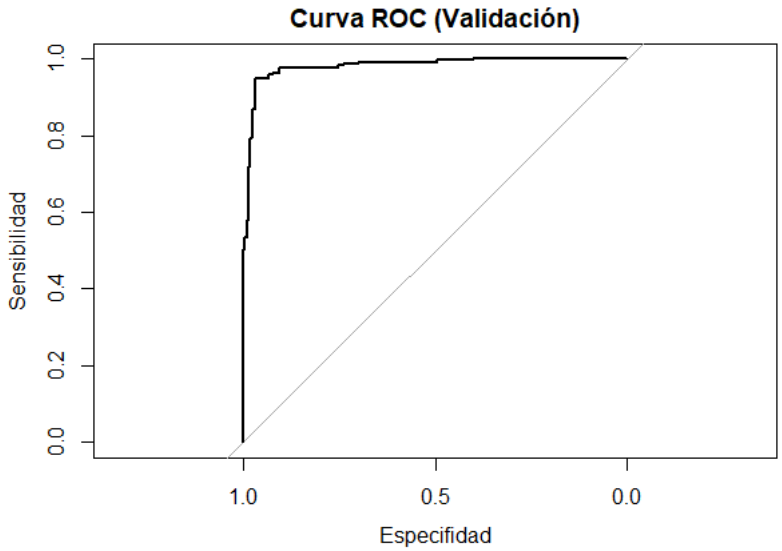


Figura. 3.18: Curva ROC del modelo Kriging Residual

Se estima el punto de corte óptimo de la misma forma en como se realizó en los modelos anteriores, este valor es aproximadamente 0.7070075, usando este valor se procede a armar la matriz de confusión, la cual se presenta en la tabla 3.21.

En la tabla 3.22 se presentan las métricas de poder de discriminación que se desprenden

	Real	
Predicción	**0**	**1**
0	235	12
1	7	229

Tabla. 3.21: Matriz de confusión del modelo Kriging Residual

de la matriz de confusión, en esta se aprecia que los valores de las métricas para el modelo Kriging Residual son satisfactorias, y mejores respecto a las obtenidas con el modelo obtenido usando solo regresión GAM mostradas en la tabla 3.12.

Partición	**Sensibilidad**	**Especificidad**	**Precisión**
Validación	0.9502	0.9711	0.9607

Tabla. 3.22: Métricas de poder de discriminación del modelo Kriging Residual

Finalmente, se muestra el mapa de calor obtenido con el modelo Kriging residual, en la figura 3.19.

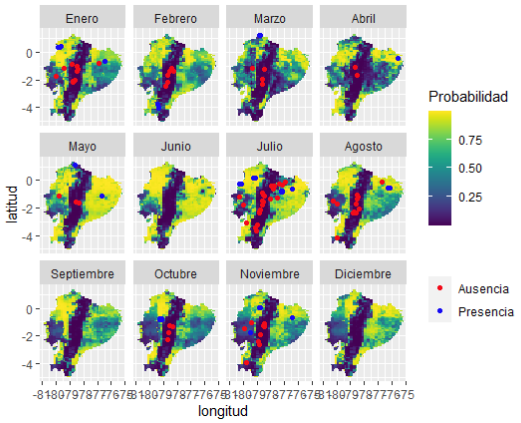


Figura. 3.19: Distribución potencial de las ranas Dendrobatidaes en el Ecuador en el año 2016 (Kriging Residual)

Los mapas de calor de todos los años estudiados se encuentran en el Apéndice B.5.

Capítulo 4

Conclusiones y Recomendaciones

En conclusión, el uso combinado de las distintas técnicas geo y bio estadísticas aquí propuestas permitieron obtener exitosamente estimaciones espacio-temporales de la distribución potencial de las ranas aposemáticas en la zona continental del Ecuador. Estas estimaciones resultan ser esenciales en el entendimiento de los factores ambientales que condicionan la supervivencia de este tipo de especies, vulnerables a cambios mínimos en factores climáticos como temperatura y precipitación, además permiten anticiparse a posibles cambios futuros en estas condiciones.

Con la finalidad de solventar el problema de tener una variable objetivo (presencia/ausencia) incompleta, en esta investigación se uso satisfactoriamente un algoritmo de creación de pseudoausencias que como se puede apreciar en los valores finales de validación de los modelos propuestos permite realizar predicciones más fiables que las técnicas usuales de los modelos de distribución de especies. El uso de la máquina de soporte vectorial de una clase en conjunto con el análisis de componentes principales permitió minimizar el riesgo de inlcuir falsas ausencias en la muestra, dado que se crea un primer mapa de habitabilidad de las ranas aposemáticas, con predicciones robustas y estables, y no se ubican las pseudoausencias aleatoriamente en el espacio. Además, el uso de la técnica de clusterización k-means hizo posible que las pseudoausencias se distribuyan uniformente en el mapa geográfico y ambiental, abarcando así la variedad de climas que posee el Ecuador.

Se pudieron probar satisfactoriamente varias técnicas de clasificación usando modelos GAM, GLM y MARS, comparables gracias a métricas de evaluación del modelo compatibles (como AUC, Gini, KS, etc), que hicieron posible elegir uno de los modelos que optimice estas

métricas, siendo este el modelo GAM, este captura de mejor manera la relación existente entre las variables regresoras y la variable objetivo, puesto que como se observa en las variables temperatura y precipitación claramente no existe una relación lineal con la presencia o ausencia de las ranas aposemáticas. Sin embargo, en todos los modelos realizados se tiene la presencia de las mismas variables regresoras, lo que indica que estas claramente son parte fundamental del hábitat de las ranas aposemáticas, de entra ellas se aprecia que justamente el tiempo ha afectado negativamente a su hábitat potencial, por lo que es posible esperar que este se reduzca en el futuro.

Otro aspecto importante de este trabajo es el uso del Kriging residual, el uso de Kriging en los residuos del modelo GAM mejoró los resultados obtenidos con el modelo GAM (en términos de poder de predicción), el uso de esta combinación de técnicas permitió considerar las diferentes relaciones que tiene la variable respuesta, pues se pudo modelar las relaciones que tiene la variable objetivo con el medioambiente, el tiempo y el espacio.

Se espera que las conclusiones obtenidas del desarrollo de esta investigación ayuden en la creación de políticas ambientales más eficientes en tareas de protección de esta clase de anfibios endémicos del país. Sin embargo, cabe resaltar que es posible extender y mejorar esta investigación con la recolección de una mayor cantidad y mejor calidad de datos de presencia de las ranas aposemáticas, el uso de un rango de tiempo más largo y más variables regresoras tales como la distancia a ciudades o ríos, radiación, fenómenos naturales tales como los incendios forestales, el fenómeno del niño, etc.

Adicionalmente, hay que recalcar que la complejidad computacional de implementar esta clase de técnicas, a una escala mayor con más observaciones de presencias de la especie y más fuentes de información que se introduzcan como capas covariantes adicionales a las mostradas aquí, o más aún mayor granularidad en las capas raster, requeriría recursos de hardware más sofisticados, como pueden ser los clusters de cómputo que usan computación en paralelo y por ende implementaciones más eficientes de los algoritmos usados en este trabajo.

Bibliografía

[Aloise et al. 2009] Aloise, Daniel ; Deshpande, Amit ; Hansen, Pierre ; Popat, Preyas: NP-hardness of Euclidean sum-of-squares clustering. In: *Machine learning* 75 (2009), Nr. 2, S. 245–248

[Ministerio del Ambiente 2017] Ambiente, Agua y Trancisión E. Ministerio del: *Ecuador es el país más diverso en especies de anfibios.* 2017

[Anderson 2007] Anderson, Raymond: *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation.* Oxford University Press, 2007

[Araújo and Williams 2000] Araújo, Miguel B. ; Williams, Paul H.: Selecting areas for species persistence using occurrence data. In: *Biological Conservation* 96 (2000), Nr. 3, S. 331–345

[Brovelli et al. 2008] Brovelli, Maria A. ; Crespi, Mattia ; Fratarcangeli, Francesca ; Giannone, Francesca ; Realini, Eugenio: Accuracy assessment of high resolution satellite imagery orientation by leave-one-out method. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 63 (2008), Nr. 4, S. 427–440

[Buja et al. 1989] Buja, Andreas ; Hastie, Trevor ; Tibshirani, Robert: Linear smoothers and additive models. In: *The Annals of Statistics* (1989), S. 453–510

[Burgess and Webster 1980] Burgess, TM ; Webster, Richard: Optimal interpolation and isarithmic mapping of soil properties: i the semi-variogram and punctual kriging. In: *Journal of soil science* 31 (1980), Nr. 2, S. 315–331

[Casal et al. 2021] Casal, Rubén F. ; Bouzas, Julián C. ; Fuente, Manuel O. de la: Aprendizaje Estadístico. (2021)

[Chefaoui and Lobo 2008] Chefaoui, Rosa M. ; Lobo, Jorge M.: Assessing the effects

of pseudo-absences on predictive distribution model performance. In: *Ecological modelling* 210 (2008), Nr. 4, S. 478–486

[Cortes and Vapnik 1995] Cortes, Corinna ; Vapnik, Vladimir: Support-vector networks. In: *Machine learning* 20 (1995), Nr. 3, S. 273–297

[Cressie 1985] Cressie, Noel: Fitting variogram models by weighted least squares. In: *Journal of the international Association for mathematical Geology* 17 (1985), Nr. 5, S. 563–586

[De Boor 1978] De Boor, Carl: *A practical guide to splines*. Bd. 27. springer-verlag New York, 1978

[Delhomme 1978] Delhomme, Jean P.: Kriging in the hydrosciences. In: *Advances in water resources* 1 (1978), Nr. 5, S. 251–266

[El-Gabbas et al. 2021] El-Gabbas, Ahmed ; Van Opzeeland, Ilse ; Burkhardt, Elke ; Boebel, Olaf: Dynamic species distribution models in the marine realm: Predicting year-round habitat suitability of baleen whales in the Southern Ocean. In: *Frontiers in Marine Science* 8 (2021), Nr. 802276

[Engler et al. 2004] Engler, Robin ; Guisan, Antoine ; Rechsteiner, Luca: An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. In: *Journal of applied ecology* 41 (2004), Nr. 2, S. 263–274

[EOS 2020] EOS, Earth Observing S.: *Índice de Vegetación de Diferencia Normalizada (NDVI)*. 2020. – URL `https://eos.com/es/make-an-analysis/ndvi/`

[ESA Climate Office 2021] ESA Climate Office: *Land Surface Temperature (LST)*. Sep 2021. – URL `https://climate.esa.int/media/documents/LST_cci_Science-Highlights_issue3.pdf`

[Friedman 1991] Friedman, Jerome H.: Estimating functions of mixed ordinal and categorical variables using adaptive splines / Stanford Univ CA Lab for Computational Statistics. 1991. – Forschungsbericht

[Gambolati and Volpi 1979] Gambolati, Giuseppe ; Volpi, Giampiero: Groundwater contour mapping in Venice by stochastic interpolators: 1. Theory. In: *Water Resources Research* 15 (1979), Nr. 2, S. 281–290

[Giraldo et al. 2007] Giraldo, Ramón ; Delicado Useros, Pedro F. ; Mateu, Jorge: Geostatistics for functional data: An ordinary kriging approach. (2007)

[GISTEMP Team 2022] GISTEMP Team: *GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies.* 2022. – URL `https://data.giss.nasa.gov/gistemp/`

[Glen 2019] Glen, Stephanie: *ROC Curve Explained in One Picture.* 2019

[Grant et al. 2006] Grant, T. ; Frost, Caldwell ; J. P., Gagliardo ; R. O. N., Haddad ; C. F., Kok ; P. J., W. C.: Phylogenetic systematics of dart-poison frogs and their relatives (Amphibia: Athesphatanura: Dendrobatidae). In: *Bulletin of the American Museum of natural History* 2006 (2006), Nr. 299, S. 1–262

[Gurka 2006] Gurka, Matthew J.: Selecting the best linear mixed model under REML. In: *The American Statistician* 60 (2006), Nr. 1, S. 19–26

[Hastie and Pregibon 2017] Hastie, Trevor J. ; Pregibon, Daryl: Generalized linear models. In: *Statistical models in S.* Routledge, 2017, S. 195–247

[Hastie and Tibshirani 2017] Hastie, Trevor J. ; Tibshirani, Robert J.: *Generalized additive models.* Routledge, 2017

[Hengl et al. 2009] Hengl, Tomislav ; Sierdsema, Henk ; Radović, Andreja ; Dilo, Arta: Spatial prediction of species' distributions from occurrence-only records: combining

point pattern analysis, ENFA and regression-kriging. In: *Ecological modelling* 220 (2009), Nr. 24, S. 3499–3511

[Hijmans et al. 2005] Hijmans, Robert J. ; Cameron, Susan E. ; Parra, Juan ; L. Jones, Peter G. ; Jarvis, Andy: Very high resolution interpolated climate surfaces for global land areas. In: *International Journal of Climatology: A Journal of the Royal Meteorological Society* 25 (2005), Nr. 15, S. 1965–1978

[Hofmann et al. 2008] Hofmann, Thomas ; Schölkopf, Bernhard ; Smola, Alexander J.: Kernel methods in machine learning. In: *The annals of statistics* 36 (2008), Nr. 3, S. 1171–1220

[Hosmer Jr et al. 2013] Hosmer Jr, David W. ; Lemeshow, Stanley ; Sturdivant, Rodney X.: *Applied logistic regression*. Bd. 398. John Wiley & Sons, 2013

[Intergovernmental Panel On Climate Change (IPOCC) 2007] Intergovernmental Panel On Climate Change (IPOCC): Climate change 2007: the physical science basis. In: *Agenda* 6 (2007), Nr. 07, S. 333

[Iturbide et al. 2015] Iturbide, Maialen ; Bedia, Joaquín ; Herrera, Sixto ; Hierro, Oscar del ; Pinto, Miriam ; Gutiérrez, Jose M.: A framework for species distribution modelling with improved pseudo-absence generation. In: *Ecological Modelling* 312 (2015), S. 166–174

[IUCN] IUCN: *The IUCN red list of threatened species.* – URL `https://www.iucnredlist.org/`

[Liang et al. 2022] Liang, Naisheng ; Tuo, Youcai ; Deng, Yun ; He, Tianfu: PCA-based SVM classification for simulated ice floes in front of sluice gates. In: *Polar Science* (2022), S. 100839

[Lobo et al. 2008] Lobo, Jorge M. ; Jiménez-Valverde, Alberto ; Real, Raimundo:

AUC: a misleading measure of the performance of predictive distribution models. In: *Global ecology and Biogeography* 17 (2008), Nr. 2, S. 145–151

[MacKay 2003] MacKay, David: An example inference task: Clustering. In: *Information theory, inference and learning algorithms* 20 (2003), S. 284–292

[Menard 2002] Menard, Scott: *Applied logistic regression analysis*. Bd. 106. Sage, 2002

[Montgomery 2017] Montgomery, Douglas C.: *Design and analysis of experiments*. John wiley & sons, 2017

[Moran 1948] Moran, Patrick A.: The interpretation of statistical maps. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (1948), Nr. 2, S. 243–251

[Munoz-Mari et al. 2007] Munoz-Mari, Jordi ; Camps-Valls, Gustavo ; Gómez-Chova, Luis ; Calpe-Maravilla, Javier: Combination of one-class remote sensing image classifiers. In: *2007 IEEE International Geoscience and Remote Sensing Symposium* IEEE (Veranst.), 2007, S. 1509–1512

[NASA Earthdata 2020] NASA Earthdata: *Species Distribution Modeling Data*. Abril 2020. – URL `https://earthdata.nasa.gov/learn/pathfinders/biodiversity/species-distribution`

[NewTechDojo 2017] NewTechDojo: *Learn Support Vector Machine using Excel – Machine Learning Algorithm*. 2017

[Novales 2010] Novales, Alfonso: Análisis de regresión. In: *Universidad Complutense de Madrid: Madrid, Spain* (2010), S. 116

[Odeh et al. 1995] Odeh, Inakwu O. ; McBratney, AB ; Chittleborough, DJ: Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. In: *Geoderma* 67 (1995), Nr. 3-4, S. 215–226

[Pedregosa et al. 2011] Pedregosa, F. ; Varoquaux, G. ; Gramfort, A. ; Michel, V. ; Thirion, B. ; Grisel, O. ; Blondel, M. ; Prettenhofer, P. ; Weiss, R. ; Dubourg, V. ; Vanderplas, J. ; Passos, A. ; Cournapeau, D. ; Brucher, M. ; Perrot, M. ; Duchesnay, E.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830

[Ron 2021] Ron, Merino-Viteri A. Ortiz D.: *Anfibios del Ecuador. Version 2021.0. Museo de Zoología, Pontificia Universidad Católica del Ecuador*. 2021

[Schechtman 2016] Schechtman, Gideon: The Relationship between Gini Methodology and the ROC curve. In: *Available at SSRN 2739245* (2016)

[Schölkopf et al. 2001] Schölkopf, Bernhard ; Platt, John C andShawe-Taylor ; Smola, John ; Alex J. Williamson, Robert C.: Estimating the support of a high-dimensional distribution. In: *Neural computation* 13 (2001), Nr. 7, S. 1443–1471

[Senay et al. 2013] Senay, Senait D. ; Worner, Susan P. ; Ikeda, Takayoshi: Novel three-step pseudo-absence selection technique for improved species distribution modelling. In: *PloS one* 8 (2013), Nr. 8, S. e71218

[Su et al. 2017] Su, Jinya ; Yi, Dewei ; Liu, Cunjia ; Guo, Lei ; Chen, Wen-Hua: Dimension reduction aided hyperspectral image classification with a small-sized training dataset: experimental comparisons. In: *Sensors* 17 (2017), Nr. 12, S. 2726

[Sundarkumar and Ravi 2013] Sundarkumar, G G. ; Ravi, Vadlamani: Malware detection by text and data mining. In: *2013 IEEE International Conference on Computational Intelligence and Computing Research* IEEE (Veranst.), 2013, S. 1–6

[Susskind et al. 2019] Susskind, J. ; Schmidt, GA. ; Lee, JN. ; Iredell, L: Recent global warming as confirmed by AIRS. In: *Environmental Research Letters* 14 (2019), Nr. 4, S. 044030

[Twomey and Brown 2008] Twomey, Evan ; Brown, Jason L.: Spotted poison frogs: rediscovery of a lost species and a new genus (Anura: Dendrobatidae) from northwestern Peru. In: *Herpetologica* 64 (2008), Nr. 1, S. 121–137

[López-de Ullibarri 1998] Ullibarri, Pita-Fernández S. López-de: Curvas Roc. In: *Cuadernos de atención primaria* 5 (1998), Nr. 4, S. 229–235

[Webster and Oliver 2007] Webster, Richard ; Oliver, Margaret A.: *Geostatistics for environmental scientists*. John Wiley & Sons, 2007

[Wood 2006] Wood, Simon N.: *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2006

[Yu et al. 2014] Yu, Huanhuan ; Chen, Rongda ; Zhang, Guoping: A SVM stock selection model within PCA. In: *Procedia computer science* 31 (2014), S. 406–412

[Zaniewski et al. 2002] Zaniewski, A E. ; Lehmann, Anthony ; Overton, Jacob M.: Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. In: *Ecological modelling* 157 (2002), Nr. 2-3, S. 261–280

Apéndice A:

Implementación

En esta sección se muestra el flujo de creación del modelo de distribución potencial de ranas aposemáticas en el Ecuador continental y la implementación en los diferentes softwares utilizados. Este flujo está compuesto principalmente por tres partes: el preprocesamiento de datos, creación de pseudoausencias y el modelo Kriging residual; tal como se muestra en la figura A.1.

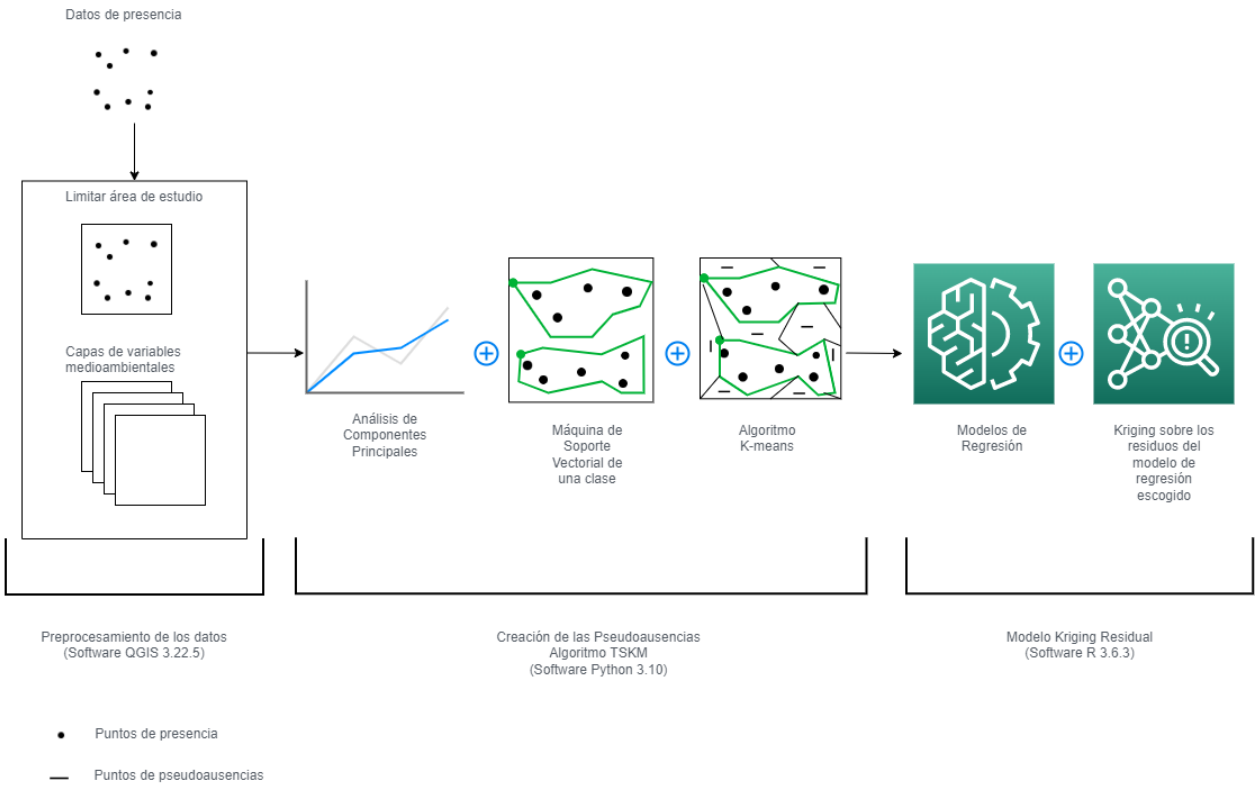


Figura. A.1: Flujo para la creación del modelo de distribución de ranas aposemáticas en el Ecuador continental

Para la fase de preprocesamiento de datos utilizando el software QGIS3 no se creó código sino que tan solo se hizo uso de la herramienta IDW ya implementada en este software.

Los códigos realizados para este proyecto de investigación se detallan a continuación.

A.1 Creación de Pseudoausencias

Para esta parte del flujo se utilizó el lenguaje de programación Python 3.10.

A.1.1 ACP y OCSVM

Este código realiza las rutinas de análisis de componentes principales, máquina de soporte vectorial de una clase y las predicciones con el modelo del OCSVM.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.font_manager
from sklearn import svm
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# Modelo SVM model
# ----------------
def svm_ranas(df = None,
variables = ['ndvi','temperatura','elevacion','precipitacion','evapo'],
path = 'data/dendrobatidaes.csv',reducir_acp = True,escala_x=0.2,
escala_y=0.2,plot = True, print_results=False):
```

```
    "Modelo One Class SVM de base alojada en path"
    if df is None:
        df = pd.read_csv(path)
    df = df.drop(columns = ['Unnamed: 0'])
    df = df.drop_duplicates(['x','y','mes','anio'])
    #variables = ['temperatura','ndvi','elevacion']

    X = estandarizar(df,variables)
    if reducir_acp:
        df_acp, acp_fitted = acp(X)
        variables = list(df_acp.columns)
    else:
        df_acp = pd.DataFrame(X, columns=variables)
        acp_fitted = None
        del df



    X_train, X_test = train_test_split(df_acp, test_size=0.25,
    random_state=0)
    #print('********** Xtrain: \n',X_train)
    X_train = X_train[variables].to_numpy()
    X_test = X_test[variables].to_numpy()
    #X_outliers = np.random.uniform(low=-4, high=4, size=(20, 3))

    model = svm.OneClassSVM(nu=0.35, gamma = 0.3, kernel="rbf",
    shrinking=True)
    model.fit(X_train)
    print('Intercepto',model.intercept_)
    print('w:',model.dual_coef_.dot(model.support_vectors_))
    #print(model.n_support_)
```

```
    #print(model.score_samples(X_train))
    #
    #y_pred_train = model.predict(X_train)
    y_pred_test = model.predict(X_test)
    y_true_test = np.ones(y_pred_test.size)

    if print_results:
        print(classification_report(y_true_test, y_pred_test))

    #
    if plot:
        mi,ma,rng = [],[],[]
        k = 0

        mi.append(df_acp[variables[k]].min())
        ma.append(df_acp[variables[k]].max())
        rng.append(ma[k] - mi[k])
        x1 = np.linspace(start=mi[k]-rng[k]*escala_x,
        stop=ma[k]+rng[k]*escala_x)

        k=+1
        mi.append(df_acp[variables[k]].min())
        ma.append(df_acp[variables[k]].max())
        rng.append(ma[k] - mi[k])
        x2 = np.linspace(start=mi[k]-rng[k]*escala_y,
        stop=ma[k]+rng[k]*escala_y)

        #d1, d2, d3 = np.meshgrid(x1,x2,x3)
        d1, d2 = np.meshgrid(x1,x2)
```

```
        #Z = model.decision_function(np.c_[d1.ravel(), d2.ravel(),
        d3.ravel()])
        Z = model.decision_function(np.c_[d1.ravel(), d2.ravel()])
        Z = Z.reshape(d1.shape)

        plot_svm(X_train,X_test,x1,x2,d1,d2,Z)

    return model, acp_fitted


def plot_svm(X_train,X_test,x1,x2,d1,d2,Z):

    plt.figure(figsize=(8, 8))
    plt.title("Máquina de Soporte Vectorial")
    plt.contourf(d1, d2, Z,
    levels=np.linspace(start=Z.min(), stop = 0, num = 14), cmap='PuBu')
    a = plt.contour(d1, d2, Z, levels=[0], linewidths=2, colors='darkred')
    plt.contourf(d1, d2, Z, levels=[0, Z.max()], colors='palevioletred')

    s = 40
    b1 = plt.scatter(X_train[:, 0], X_train[:, 1], c='white',
    s=s, edgecolors='k')
    b2 = plt.scatter(X_test[:, 0], X_test[:, 1], c='blueviolet',
    s=s, edgecolors='k')

    plt.axis('tight')
    plt.xlim((min(x1), max(x1)))
    plt.ylim((min(x2), max(x2)))
    plt.legend([a.collections[0], b1, b2],
            ["Frontera", "Datos de entrenamiento",
```

```
            "Datos de test"],
            loc="upper left",
            prop=matplotlib.font_manager.FontProperties(size=11))
    plt.show()


def estandarizar(df, variables):
    "Estandarizacion de variables, se centra y escala"
    df = df.dropna()
    X = df.loc[:, variables].values# Separating out the target
    X = StandardScaler().fit_transform(X)
    #df = pd.DataFrame(x, columns=variables)
    return X


def acp(X, n_components=2):
    acp_fitted = PCA(n_components=n_components)
    principalComponents = acp_fitted.fit_transform(X)
    var_names = ['PC_'+ str(k+1) for k in range(n_components)]
    df_acp = pd.DataFrame(data = principalComponents,
    columns = var_names)
    print('PCA: Varianza explicada por las nuevas componentes',
    acp_fitted.explained_variance_ratio_)
    #print('PCA: Absolute value of each component',
    abs( acp_fitted.components_ ))
    return df_acp,acp_fitted


def predict_svm(model,acp_fitted,df_new,
variables=['ndvi','temperatura','elevacion','precipitacion','evapo'],
```

```
calcular_dist = True):
    X_new = estandarizar(df_new,variables)
    acp_projeccion = acp_fitted.transform(X_new)

    n_components = acp_projeccion.shape[1]

    var_names = ['PC_'+ str(k+1) for k in range(n_components)]
    Xdf_acp = pd.DataFrame(data = acp_projeccion,columns = var_names)

    # Prediccion
    y_pred_new = model.predict(Xdf_acp)
    if calcular_dist:
        distancias = model.decision_function(Xdf_acp)
    else:
        distancias = None
    return y_pred_new,distancias,Xdf_acp
```

A.1.2 K-means y extracción de pseudoausencias

Este código realiza las rutinas de clusterización K-medias y la extracción de las pseudoausencias.

```
from sklearn.cluster import KMeans
import pandas as pd
import numpy as np
from oc_svm import *

def pseudoausencias(df,df_new,
variables=['ndvi','temperatura','elevacion','precipitacion','evapo']):
    df = df.dropna()
```

```
    df_new = df_new.dropna()

    model, acp_fitted = svm_ranas(df = df,variables=variables,plot=False)
    y_pred_new, distancia,Xdf_acp = predict_svm(model = model,
    acp_fitted = acp_fitted, df_new = df_new, variables= variables)
    df_new['y_prediccion'] = y_pred_new

    anio = list(range(2012,2020))
    mes = list(range(1,13))

    df_pseudo = pd.DataFrame(columns = ['x','y','ndvi','temperatura',
    'elevacion','precipitacion','evapo','cluster','anio','mes'])
    df_cluster = pd.DataFrame(columns = ['x','y','ndvi','temperatura',
    'elevacion','precipitacion','evapo','cluster','anio','mes'])

    for j in anio:
        for k in mes:
            pseudo_aux, cluster_aux = pseudo_kmeans(df = df,
            df_new = df_new, anio=j, mes=k)
            df_pseudo = df_pseudo.append(pseudo_aux,ignore_index=True)
            df_cluster = df_cluster.append(cluster_aux,ignore_index=True)

    return df_pseudo, df_cluster, Xdf_acp


def pseudo_kmeans(df,df_new,anio = 2012, mes = 1):
    ranas = df[(df['anio']==anio) & (df['mes']==mes)]
    #print('** RANAS: \n',ranas)
    if ranas.shape[0]>0:
        background = df_new[['mes','anio','x', 'y','ndvi','temperatura',
```

```
        'elevacion','precipitacion','evapo','y_prediccion']]
        clusters = background[(background['anio']==anio) &
        (background['mes']==mes) &
        (background['y_prediccion']==-1)][['x','y','ndvi','temperatura',
        'elevacion','precipitacion','evapo']]
        #print("**************************************")
        #print('anio:',anio,'mes:',mes,'presencias:',ranas.shape[0])
        if(clusters.shape[0]<ranas.shape[0]):
            kmeans = KMeans(n_clusters=clusters.shape[0],
            random_state=0).fit(clusters.to_numpy())
        else:
            kmeans = KMeans(n_clusters=ranas.shape[0],
            random_state=0).fit(clusters.to_numpy())

        clusters['anio'] = anio
        clusters['mes'] = mes
        clusters['cluster'] = kmeans.labels_

        pseudo = pd.DataFrame(kmeans.cluster_centers_,
        columns=['x','y','ndvi','temperatura','elevacion',
        'precipitacion','evapo'])
        pseudo['anio'] = anio
        pseudo['mes'] = mes
        pseudo['cluster'] = (-1)*np.array(range(kmeans.cluster_centers_.shape[0]))
    else:
        pseudo = pd.DataFrame(columns=['x','y','ndvi','temperatura',
        'elevacion','precipitacion','evapo','cluster'])
        clusters = pd.DataFrame(columns=['x','y','ndvi','temperatura',
        'elevacion','precipitacion','evapo','cluster'])
```

```
    return pseudo, clusters
```

A.2 Modelos de Regresión

Los diferentes modelos de regresión expuestos en este trabajo se realizaron en el lenguaje de programación R 3.6.3.

A.2.1 Modelo GLM

```
library(dplyr)
library(readr)
library(MASS)
library(caret)
library(mgcv)
library(ggplot2)
library(viridis)

#### GLM
df <- read_csv("data/baseRegresion.csv")
set.seed(123)

train_ind <- sample.int(n = nrow(df),size = floor(.8*nrow(df)),replace = F)
train <- df[train_ind, ]
test <- df[-train_ind, ]

model_train = glm(formula = y~.-evapo,data = train,family = "binomial")

model <- stepAIC(model_train, direction = "backward", trace = FALSE)
summary(model)
```

```
car::vif(model)

## Cálculo de los estadísticos de desempeño

prediction_train <- predict(model, type = "response")
prediction_test <- predict(model, newdata = test, type = "response")

ks_train = MLmetrics::KS_Stat(y_pred = prediction_train,y_true = train$y)
ks_test = MLmetrics::KS_Stat(y_pred = prediction_test, y_true = test$y)
AUC_train = MLmetrics::AUC(y_pred = prediction_train, y_true = train$y)
AUC_test = MLmetrics::AUC(y_pred = prediction_test, y_true = test$y)
gini_train = MLmetrics::Gini(y_pred = prediction_train,y_true = train$y)
gini_test = MLmetrics::Gini(y_pred = prediction_test, y_true = test$y)

library(pROC)
curva_ROC <- roc(train$y,model$fitted.values)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Entrenamiento)")
curva_ROC <- roc(test$y,prediction_test)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Validación)")

library(InformationValue)
optCutOff<-optimalCutoff(train$y,predictedScores = prediction_train)[1]

pred_train <- ifelse(prediction_train > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_train, train$y),positive = '1')
```

```
pred_test <- ifelse(prediction_test > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_test, test$y),positive = '1')
```

A.2.2 Modelo GAM

```
library(dplyr)
library(readr)
library(MASS)
library(caret)
library(mgcv)
library(ggplot2)
library(viridis)
##### GAM

df <- read_csv("data/baseRegresion.csv")
set.seed(123)
#df = df %>% rename("evapotranspiracion"=evapo)
train_ind <- sample.int(n = nrow(df),size = floor(.8*nrow(df)),replace = F)
train <- df[train_ind, ]
test <- df[-train_ind, ]

model = mgcv::gam(formula = y~s(latitud)+anio+
                  +s(temperatura)+ ndvi
                   +s(precipitacion)+elevacion,
                  data = train, family = binomial,
                  select = TRUE,method="REML")
concurvity(model)
summary(model)
plot(model, se=TRUE, col="blue",shade = TRUE)
```

```
## Cálculo de los estadísticos de desempeño

prediction_train <- predict(model, type = "response")
prediction_test <- predict(model, newdata = test, type = "response")

ks_train = MLmetrics::KS_Stat(y_pred = prediction_train,y_true=train$y)
ks_test = MLmetrics::KS_Stat(y_pred = prediction_test,y_true = test$y)
AUC_train = MLmetrics::AUC(y_pred = prediction_train, y_true = train$y)
AUC_test = MLmetrics::AUC(y_pred = prediction_test, y_true = test$y)
gini_train = MLmetrics::Gini(y_pred = prediction_train,y_true = train$y)
gini_test = MLmetrics::Gini(y_pred = prediction_test, y_true = test$y)

library(pROC)
curva_ROC <- roc(train$y,model$fitted.values)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Entrenamiento)")
curva_ROC <- roc(test$y,prediction_test)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Validación)")

library(InformationValue)
optCutOff<-optimalCutoff(train$y,predictedScores = prediction_train)[1]

pred_train <- ifelse(prediction_train > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_train, train$y),positive = '1')

pred_test <- ifelse(prediction_test > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_test, test$y),positive = '1')
```

A.2.3 Modelo MARS

```
library(earth)
library(caret)
library(dplyr)
library(readr)
library(MASS)
library(caret)
library(ggplot2)
library(viridis)
##### MARS

df <- read_csv("data/baseRegresion.csv")
set.seed(123)
train_ind<-sample.int(n=nrow(df),size=floor(.8*nrow(df)),replace = F)
train <- df[train_ind, ]
test <- df[-train_ind, ]

model=earth(factor(y)~anio+precipitacion+temperatura+elevacion,
            data = train,
            keepxy=TRUE,
             glm=list(family=binomial),
            degree=2, nprune=7)

plot(model)
summary(model)
## Cálculo de los estadísticos de desempeño
```

```
#prediction_train <- predict(model$finalModel, type = "response")
#prediction_test<-predict(model$finalModel,newdata=test,type="response")
prediction_train <- predict(model, type = "response")
prediction_test <- predict(model, newdata = test, type = "response")

ks_train = MLmetrics::KS_Stat(y_pred=prediction_train,y_true = train$y)
ks_test = MLmetrics::KS_Stat(y_pred = prediction_test, y_true = test$y)
AUC_train = MLmetrics::AUC(y_pred = prediction_train, y_true = train$y)
AUC_test = MLmetrics::AUC(y_pred = prediction_test, y_true = test$y)
gini_train = MLmetrics::Gini(y_pred = prediction_train,y_true =train$y)
gini_test = MLmetrics::Gini(y_pred = prediction_test, y_true = test$y)

library(pROC)
curva_ROC <- roc(train$y,prediction_train)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Entrenamiento)")
curva_ROC <- roc(test$y,prediction_test)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Validación)")

library(InformationValue)
optCutOff<-optimalCutoff(train$y,predictedScores=prediction_train)[1]
pred_train <- ifelse(prediction_train > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_train, train$y),positive = '1')
pred_test <- ifelse(prediction_test > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_test, test$y),positive = '1')
```

A.2.4 Kriging Residual

```
#### Kriging
library(gstat)
library(raster)
library(automap)
library(dplyr)
library(readr)
library(MASS)
library(caret)
library(mgcv)
library(ggplot2)
library(viridis)

load("data/modeloGAM.RData")
df <- read_csv("data/baseRegresion.csv")
set.seed(123)
train_ind<-sample.int(n=nrow(df),size = floor(.8*nrow(df)),replace=F)
train.xy_entrada <- df[train_ind, ]
train.xy_entrada = train.xy_entrada %>%
dplyr::select(anio,mes,longitud,latitud)
test <- df[-train_ind, ]
test.xy_entrada = test #%>% dplyr::select(anio,mes,longitud,latitud)
train.xy_entrada$residuos = residuals(model)
#aux = train.xy %>% dplyr::select(anio,mes,longitud,latitud) %>%
group_by(anio,mes) %>% count()

### Entrenamiento

anio = c(2012,2013,2014,2015,2016,2017,2018,2019)
```

```
mes = c(1,2,3,4,5,6,7,8,9,10,11,12)
aux = data.frame("y"=NA,"anio"=NA,"mes"=NA,"longitud"=NA,
"latitud"=NA,"gam"=NA,"gam_kr"=NA)
for (i in anio) {
  for (j in mes) {
    train.xy = train.xy_entrada %>% filter(anio==i,mes==j)
    train.xy = train.xy[!duplicated(train.xy),]
    n = nrow(train.xy)
    test.xy = test.xy_entrada %>% filter(anio==i,mes==j)
    if((is.data.frame(train.xy) && nrow(train.xy)>14)&&
    (is.data.frame(test.xy) && nrow(test.xy)>0)){
      test.xy_kr = test.xy
      coordinates(train.xy) = ~longitud+latitud
      coordinates(test.xy_kr) = ~longitud+latitud
      tryCatch(
        expr = {
          variograma=autofitVariogram(residuos~ 1, input_data = train.xy)
          png(file = paste0("figuras/KRIGING/Variograma año",i,
          " mes",j,".png"))
          plot(variograma, pch=19, col="green",
          sub=paste0("Variograma del mes= ",j," año= ",i," n=",n,"
          sserr=",round(variograma$sserr,2)))
          dev.off()
          kr = autoKrige(residuos~ 1, input_data = train.xy,
          new_data = test.xy_kr)
          predicciones_kr = kr$krige_output$var1.pred
          prediction_test <- predict(model, newdata = test.xy)
          gam <- predict(model, newdata = test.xy,type="response")
          gam_kr <- prediction_test + predicciones_kr
          gam_kr = 1/(1 + exp(-gam_kr))
```

```
          test.xy$gam_kr = gam_kr
          test.xy$gam = gam
          test.xy = test.xy %>%
          dplyr::select(y,anio,mes,longitud,latitud,gam,gam_kr)
          aux = rbind(aux,test.xy)
        },
        error = function(e){
          gam <- predict(model, newdata = test.xy,type="response")
          test.xy$gam = gam
          test.xy$gam_kr = gam
          test.xy = test.xy %>%
          dplyr::select(y,anio,mes,longitud,latitud,gam,gam_kr)
          aux = rbind(aux,test.xy)
          message(paste0('Caught an error! mes',j,"año",i))
          print(e)
        }
      )
    }
  }
}
aux = aux[complete.cases(aux),]
evalData <- sqrt(mean((aux$gam_kr - aux$y)^2))
ks_test = MLmetrics::KS_Stat(y_pred = aux$gam_kr, y_true = aux$y)
AUC_test = MLmetrics::AUC(y_pred = aux$gam_kr, y_true = aux$y)
gini_test = MLmetrics::Gini(y_pred = aux$gam_kr, y_true = aux$y)
library(pROC)
curva_ROC <- roc(aux$y,aux$gam_kr)
plot(curva_ROC,col = "black",xlim = c(1,0),ylim = c(0,1),
     xlab = "Especifidad", ylab ="Sensibilidad",
     main = "Curva ROC (Validación)")
```

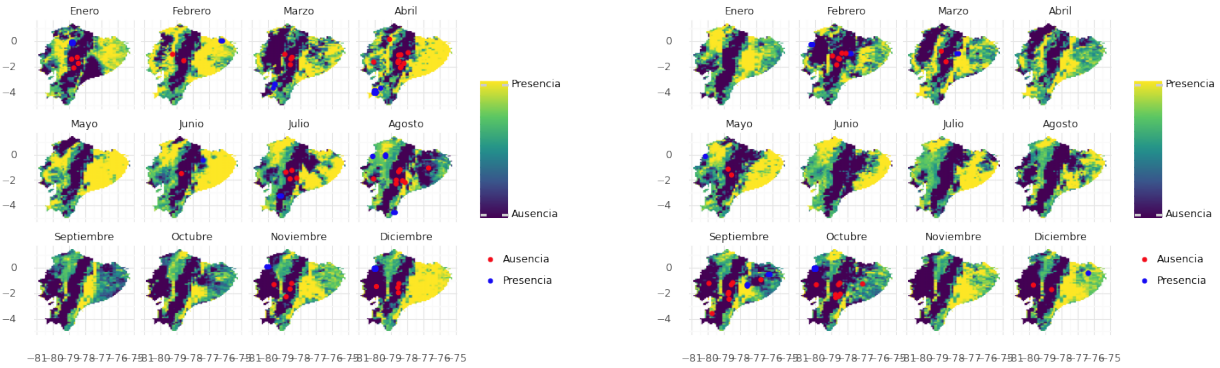
```
library(InformationValue)
optCutOff <- optimalCutoff(aux$y,predictedScores = aux$gam_kr)[1]
pred_test <- ifelse(aux$gam_kr > optCutOff, 1, 0)
caret::confusionMatrix(table(pred_test, aux$y),positive = '1')
```

Apéndice B:

Mapas de calor de la distribución potencial de las ranas aposemáticas

B.1 Distribución potencial usando OCSVM

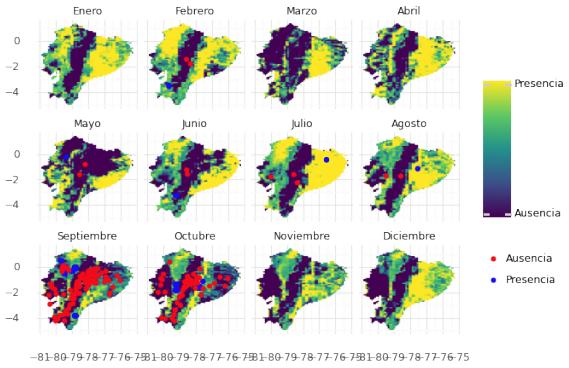
En esta sección se presentan los mapas de calor sobre el Ecuador, que se obtuvieron al realizar la técnica de perfil OCSVM modelando solo las presencias de las ranas.



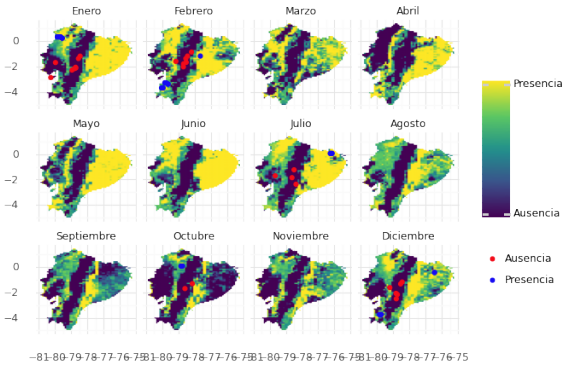
(a) Año 2012

(b) Año 2013

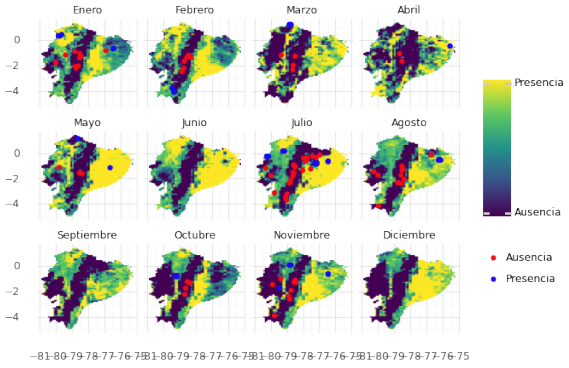
Figura. B.1: Distribución potencial de las ranas Dendrobatidaes usando OCSVM



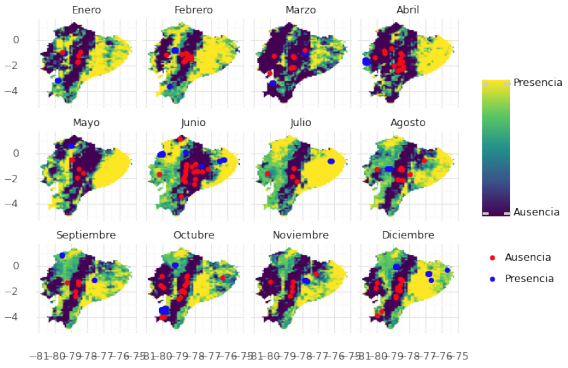
(c) Año 2014



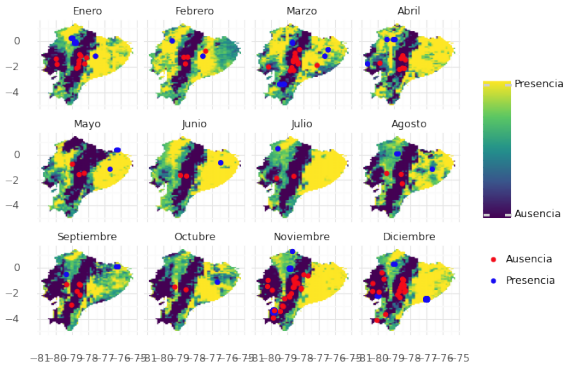
(d) Año 2015



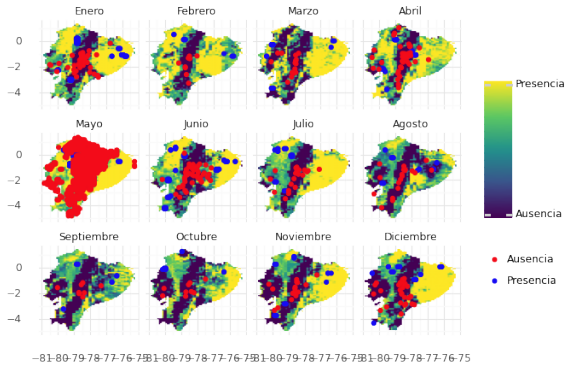
(e) Año 2016



(f) Año 2017



(g) Año 2018

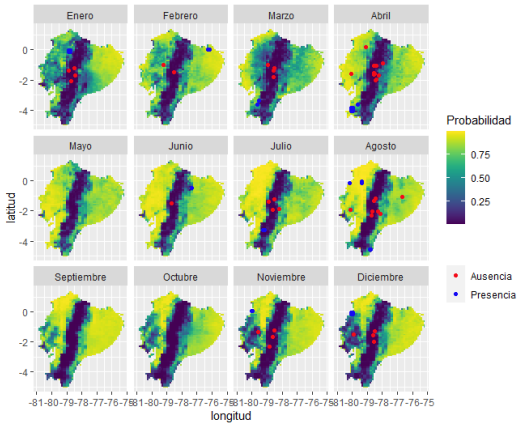


(h) Año 2019

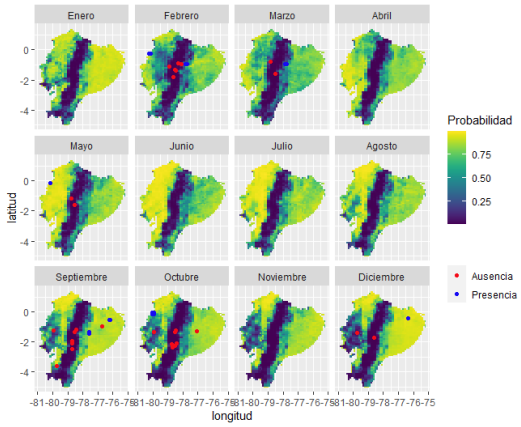
Figura. B.1: Distribución potencial de las ranas Dendrobatidaes usando OCSVM

B.2 Distribución potencial usando modelos GLM

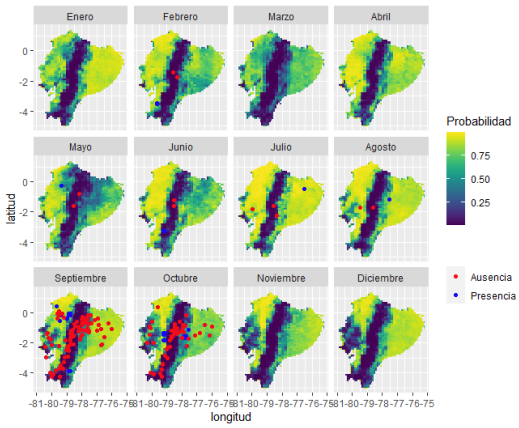
En esta sección se presentan los mapas de calor que se obtienen utilizando modelos de regresión lineal logística para la estimación de la distribución de las ranas aposemáticas en el Ecuador.



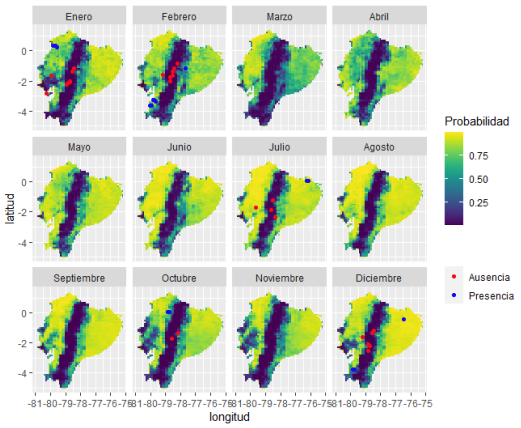
(a) Año 2012



(b) Año 2013

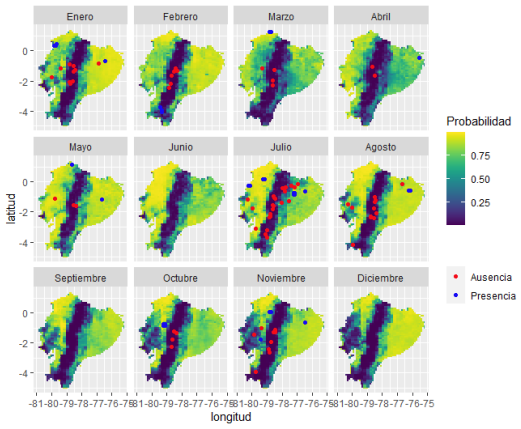


(c) Año 2014

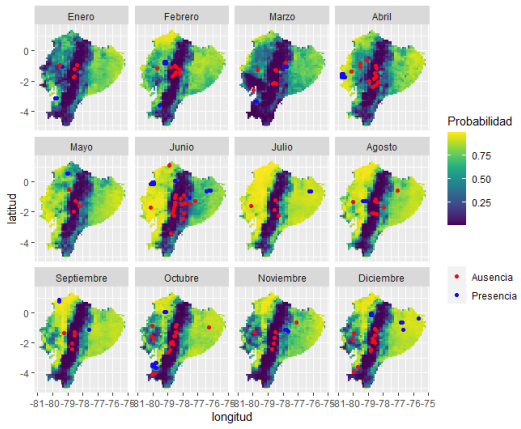


(d) Año 2015

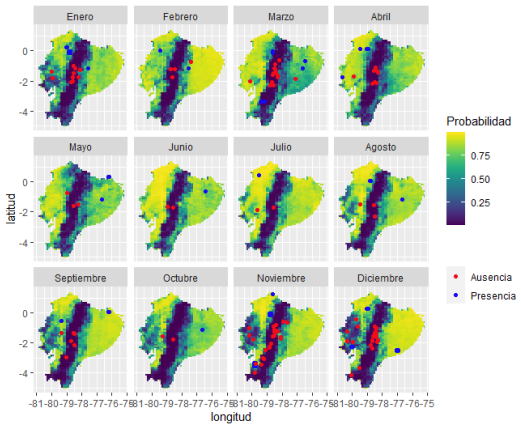
Figura. B.2: Distribución potencial de las ranas Dendrobatidaes usando GLM



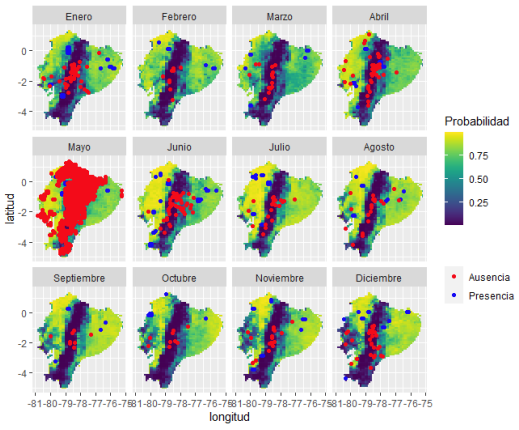
(e) Año 2016



(f) Año 2017



(g) Año 2018

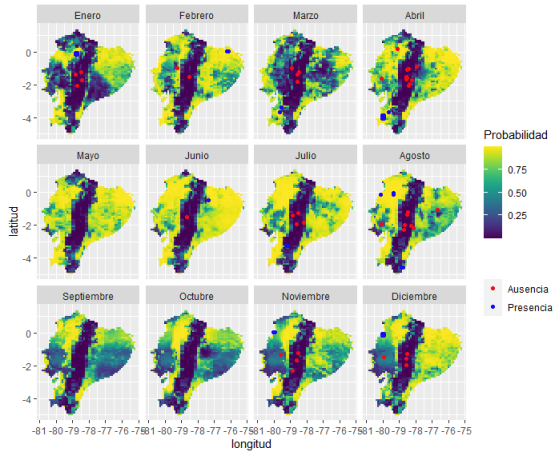


(h) Año 2019

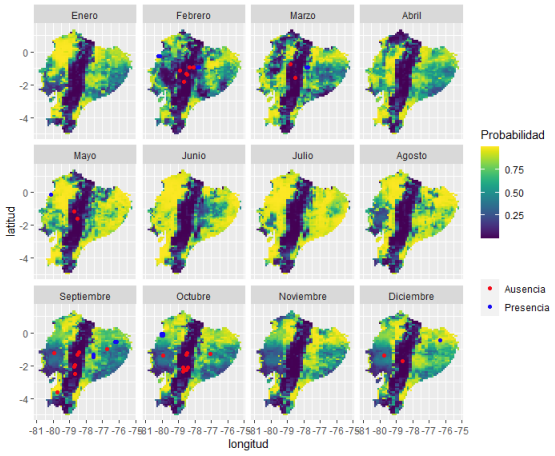
Figura. B.2: Distribución potencial de las ranas Dendrobatidaes usando GLM

B.3 Distribución potencial usando modelos GAM

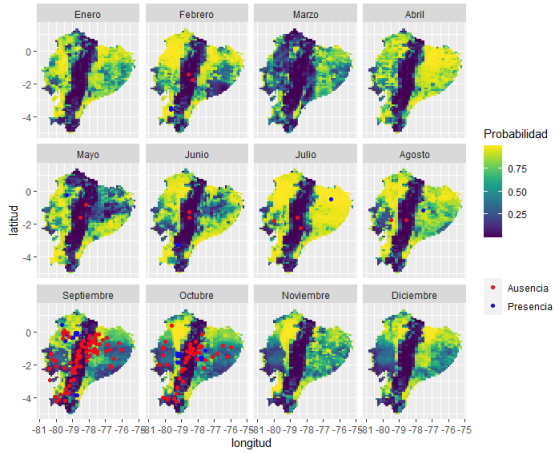
En esta sección se presentan los mapas de calor que se obtienen utilizando modelos de regresión aditiva logística para la estimación de la distribución de las ranas aposemáticas en el Ecuador.



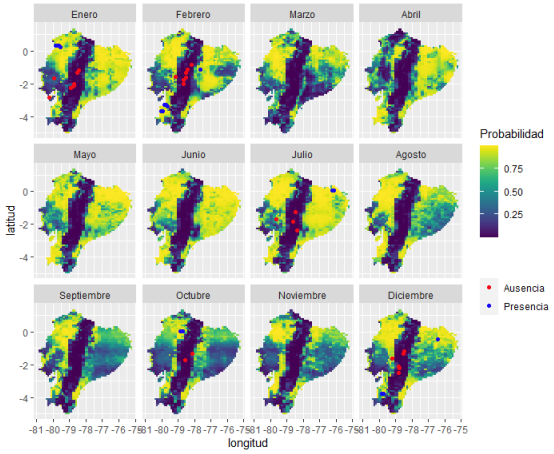
(a) Año 2012



(b) Año 2013

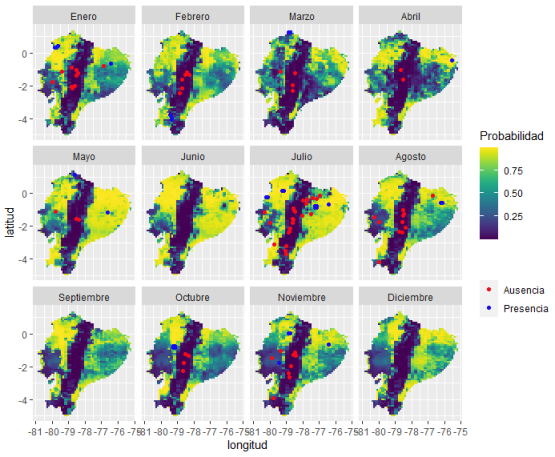


(c) Año 2014

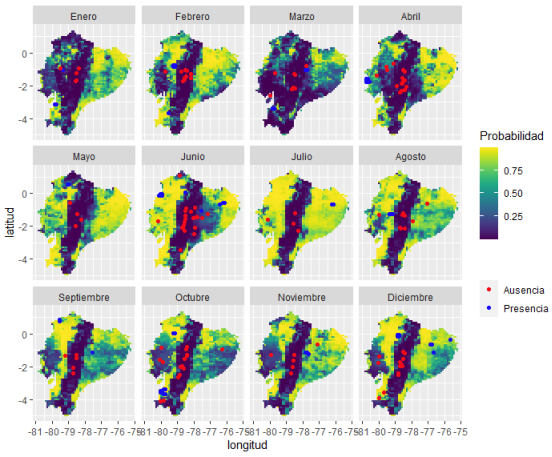


(d) Año 2015

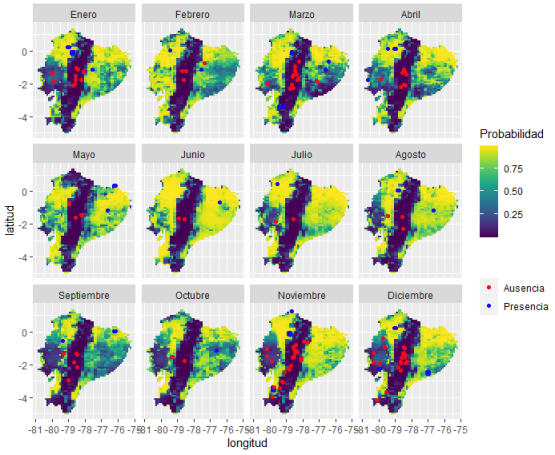
Figura. B.3: Distribución potencial de las ranas Dendrobatidaes usando GAM



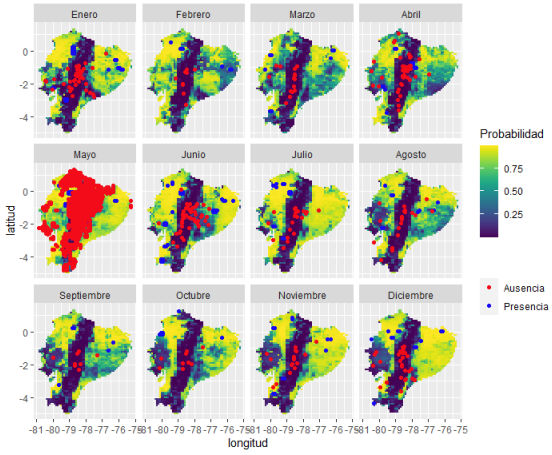
(e) Año 2016



(f) Año 2017



(g) Año 2018

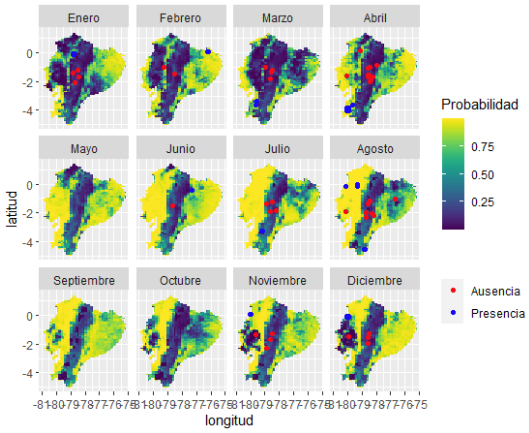


(h) Año 2019

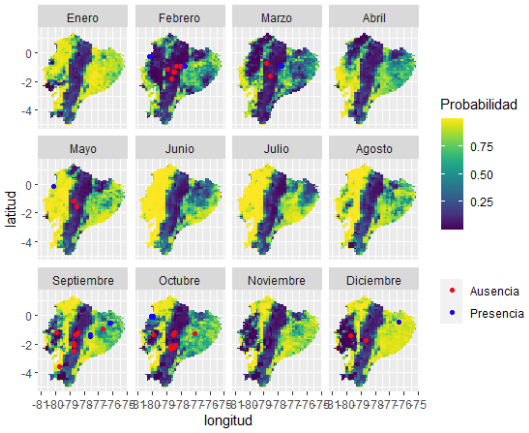
Figura. B.3: Distribución potencial de las ranas Dendrobatidaes usando GAM

B.4 Distribución potencial usando modelos MARS

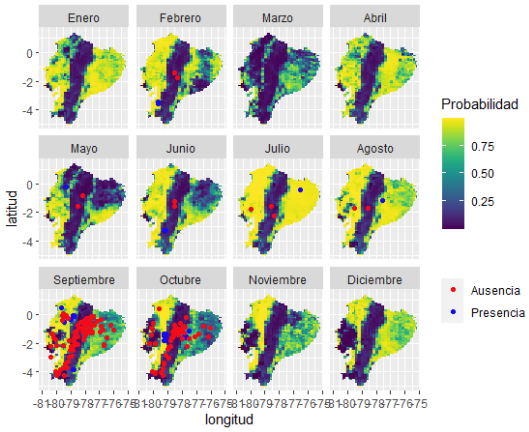
En esta sección se presentan los mapas de calor que se obtienen utilizando modelos de regresión adaptiva multivariante con splines logística para la estimación de la distribución de las ranas aposemáticas en el Ecuador.



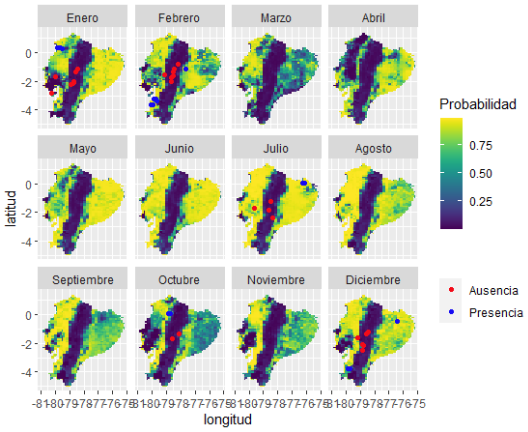
(a) Año 2012



(b) Año 2013

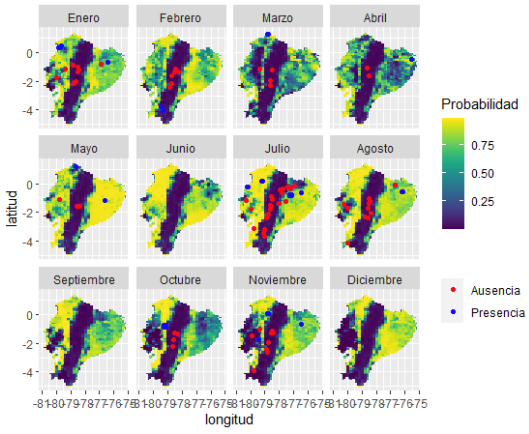


(c) Año 2014

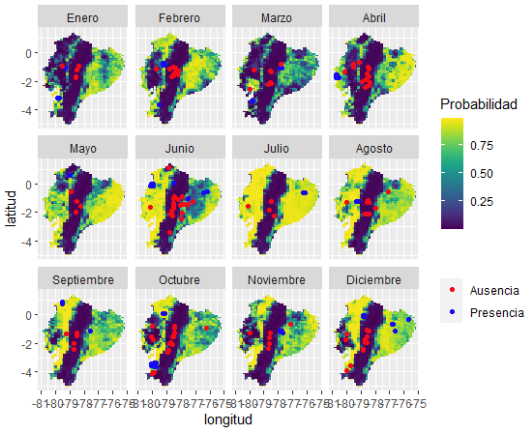


(d) Año 2015

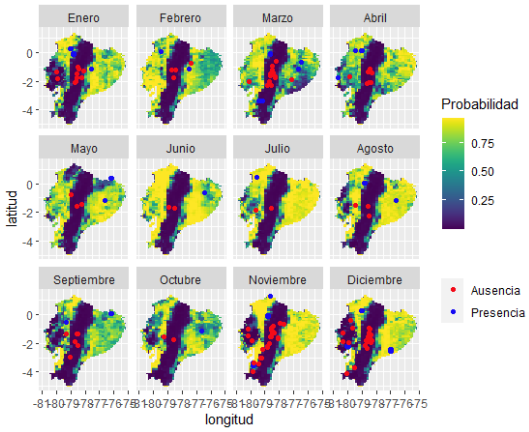
Figura. B.4: Distribución potencial de las ranas Dendrobatidaes usando MARS



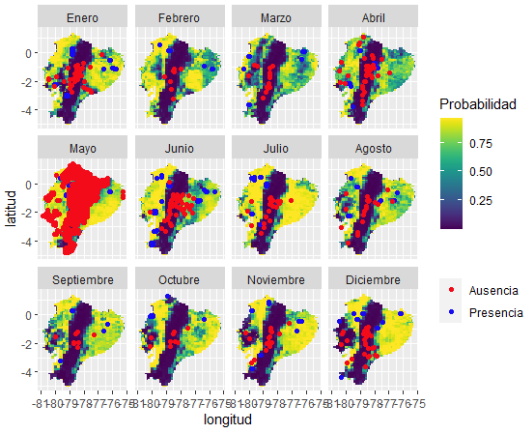
(e) Año 2016



(f) Año 2017



(g) Año 2018

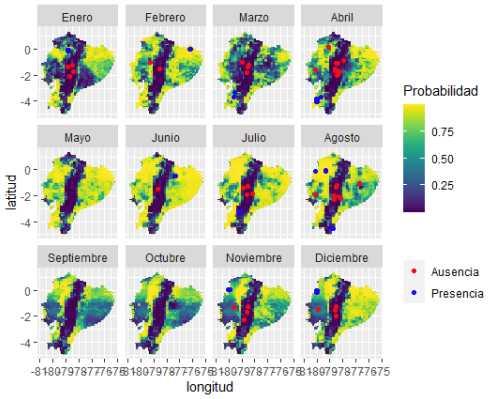


(h) Año 2019

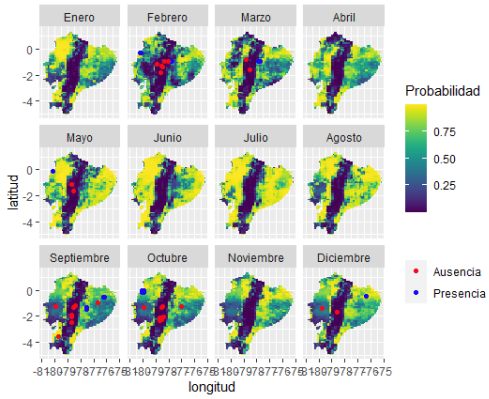
Figura. B.4: Distribución potencial de las ranas Dendrobatidaes usando MARS

B.5 Distribución potencial usando el modelo Kriging Residual

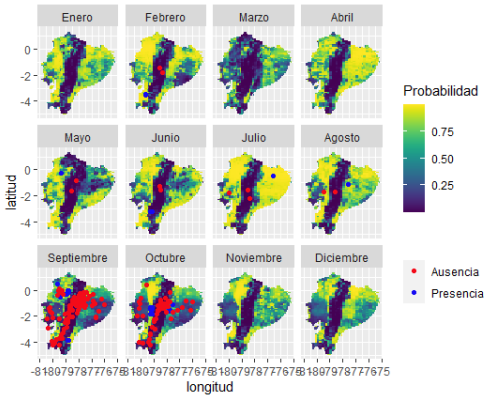
En esta sección se presentan los mapas de calor que se obtienen utilizando el modelo Kriging Residual para la estimación de la distribución de las ranas aposemáticas en el Ecuador.



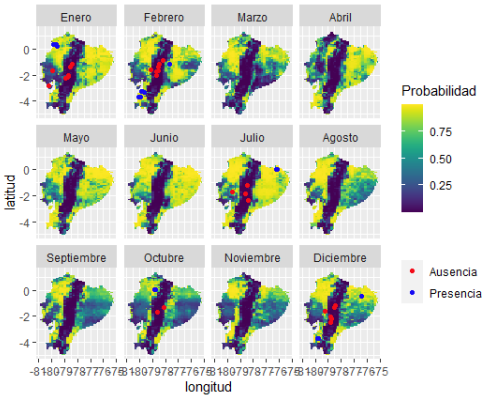
(a) Año 2012



(b) Año 2013

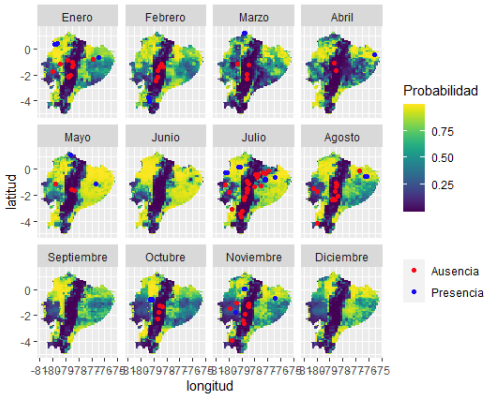


(c) Año 2014

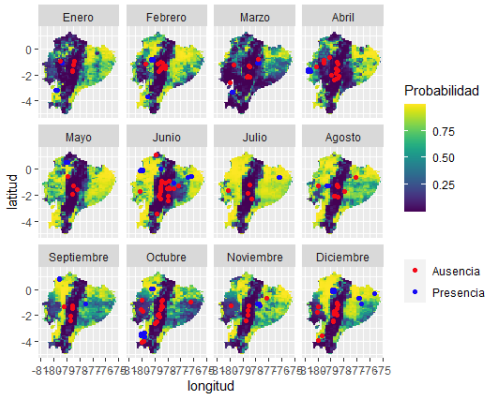


(d) Año 2015

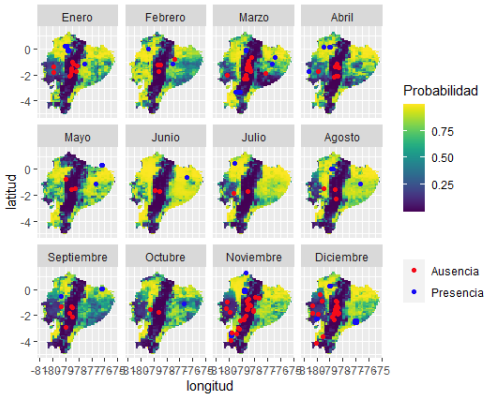
Figura. B.5: Distribución potencial de las ranas Dendrobatidaes usando Kriging Residual



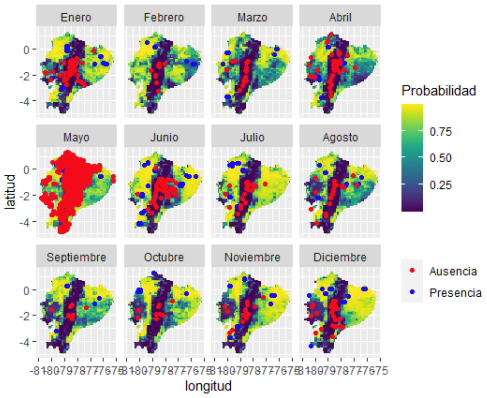
(e) Año 2016



(f) Año 2017



(g) Año 2018



(h) Año 2019

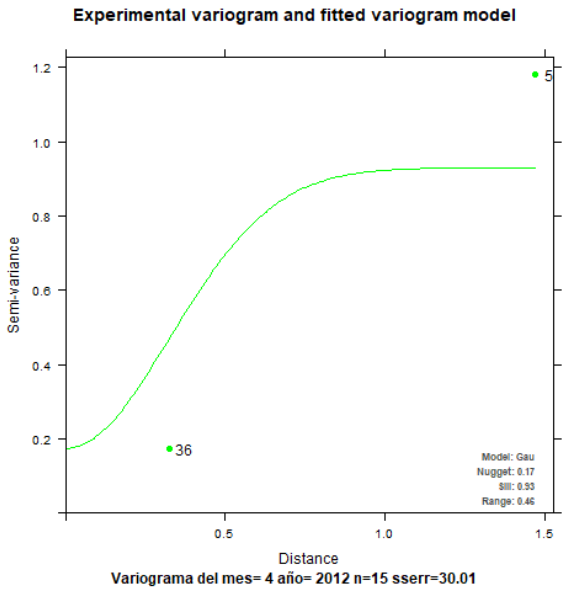
Figura. B.5: Distribución potencial de las ranas Dendrobatidaes usando Kriging Residual

Apéndice C:

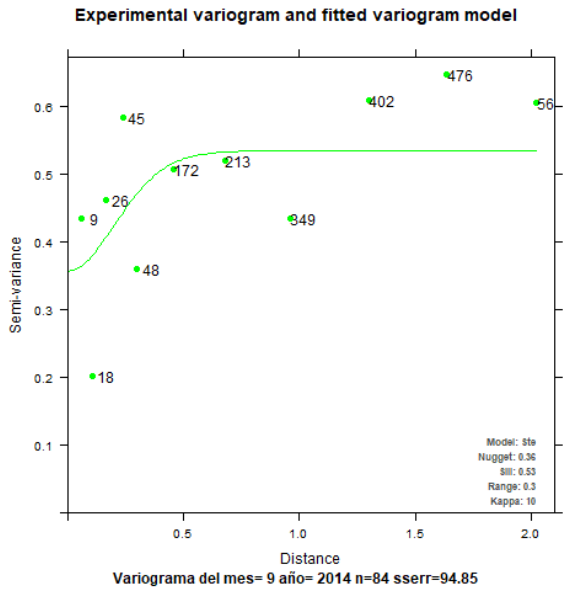
Variogramas Modelados

C.1 Gráficos de los Variogramas

En esta sección se presentan los gráficos de los modelos de variogramas.

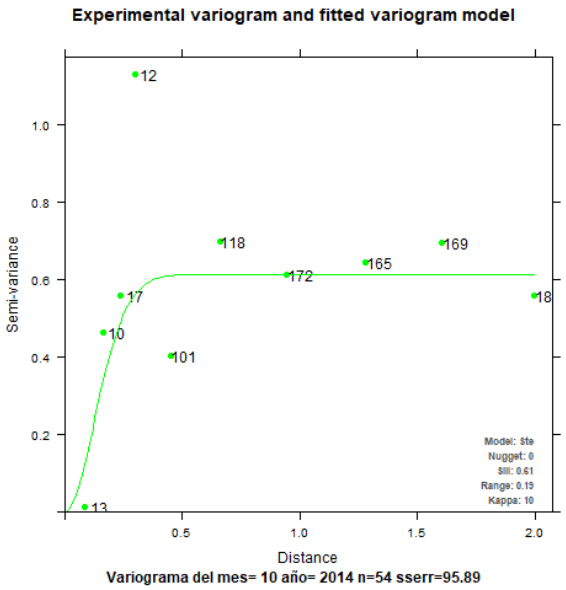


(a) Año: 2012, Mes: Abril

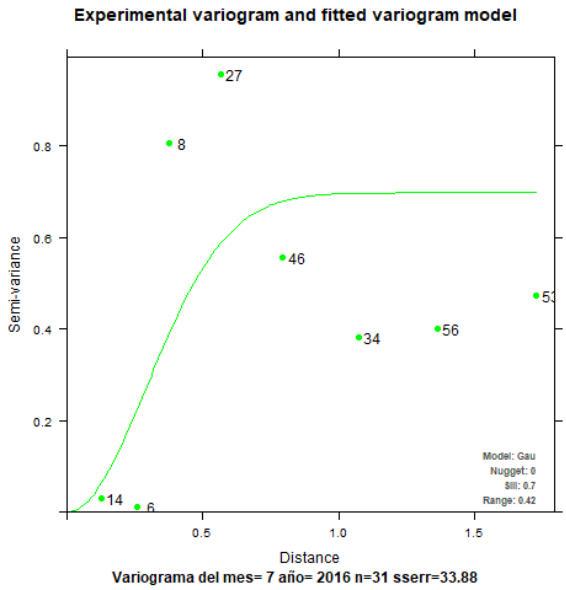


(b) Año: 2014, Mes: Septiembre

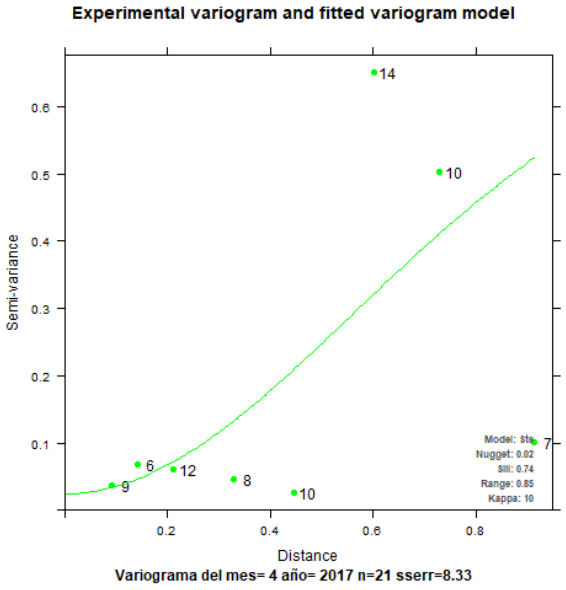
Figura. C.1: Variogramas modelados



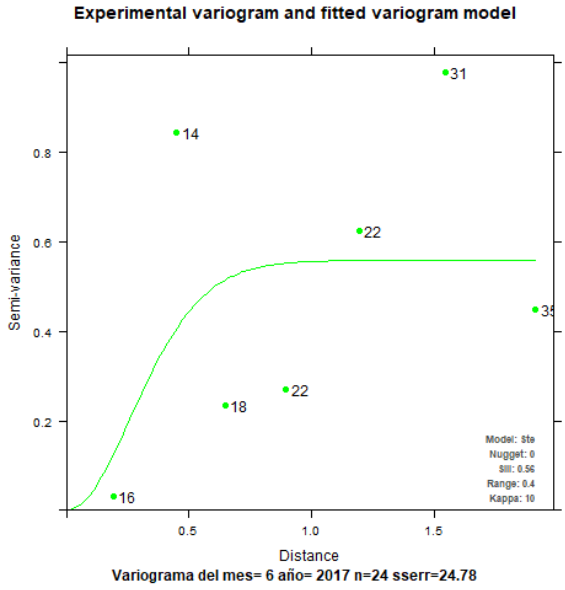
(c) Año: 2014, Mes: Octubre



(d) Año: 2016, Mes: Julio

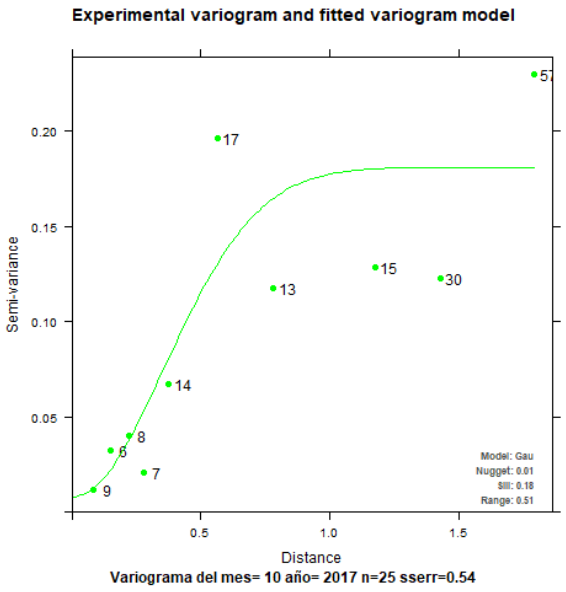


(e) Año: 2017, Mes: Abril

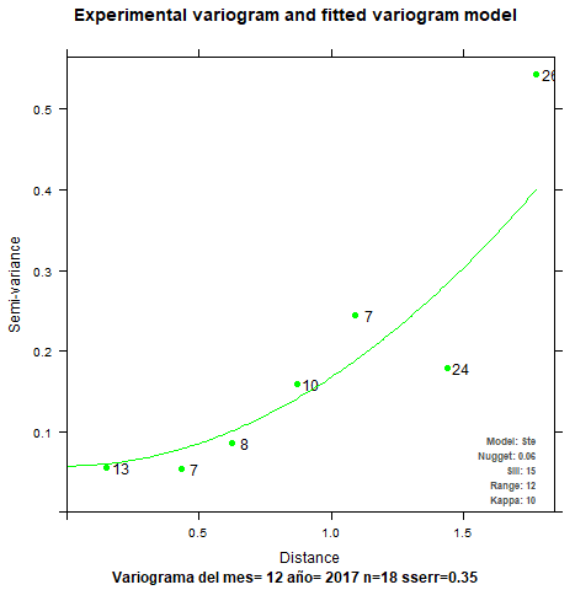


(f) Año: 2017, Mes: Junio

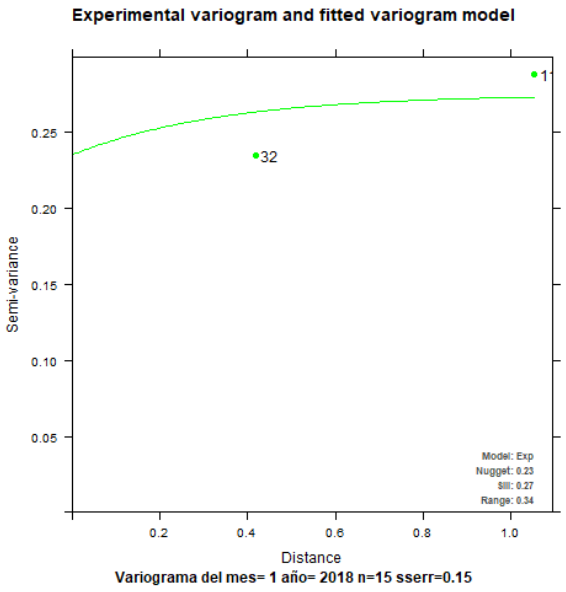
Figura. C.1: Variogramas modelados



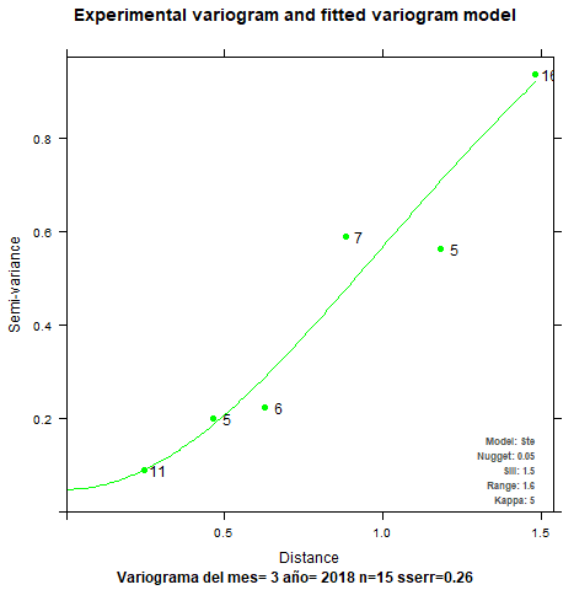
(g) Año: 2017, Mes: Octubre



(h) Año: 2017, Mes: Diciembre

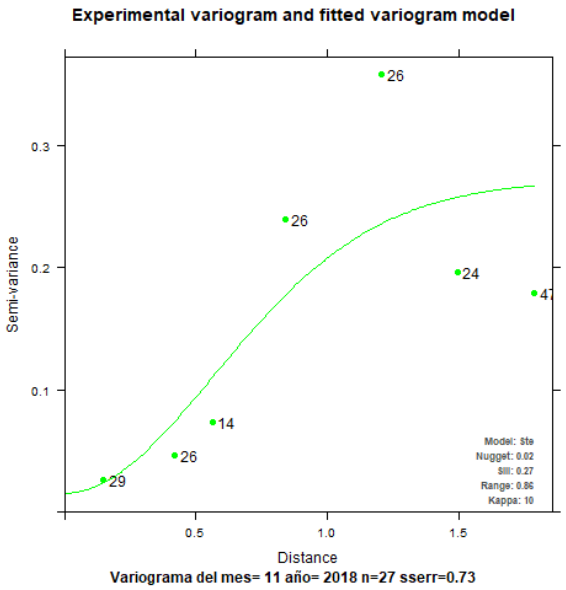


(i) Año: 2018, Mes: Enero

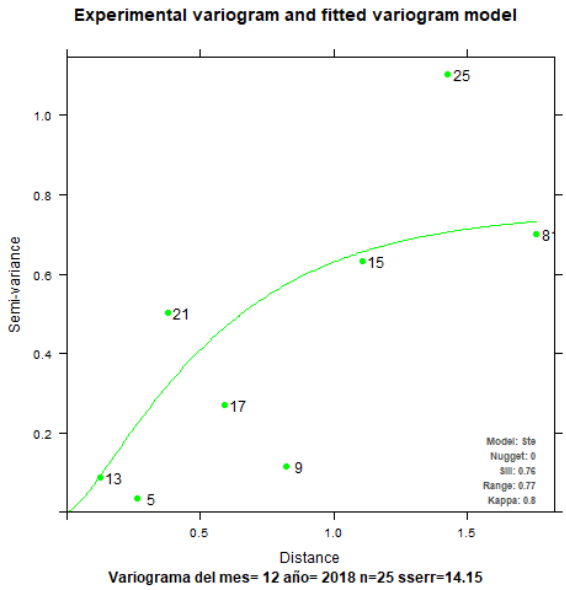


(j) Año: 2018, Mes: Marzo

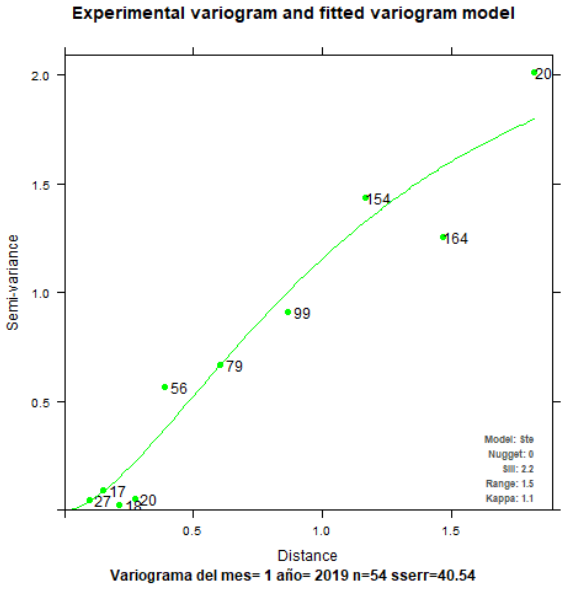
Figura. C.1: Variogramas modelados



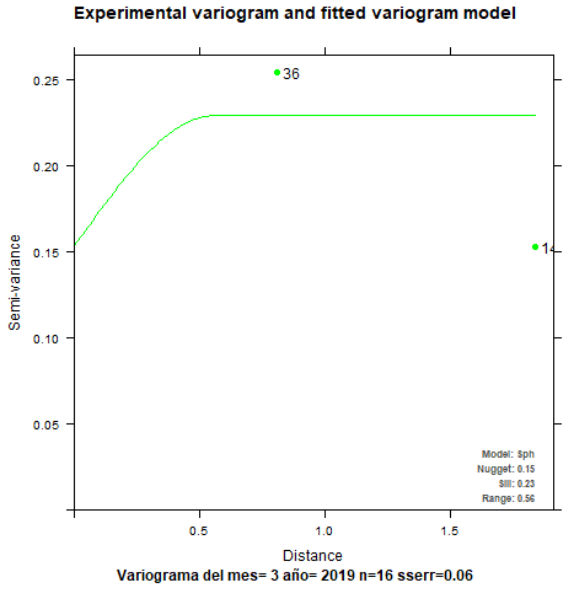
(k) Año: 2018, Mes: Noviembre



(l) Año: 2018, Mes: Diciembre

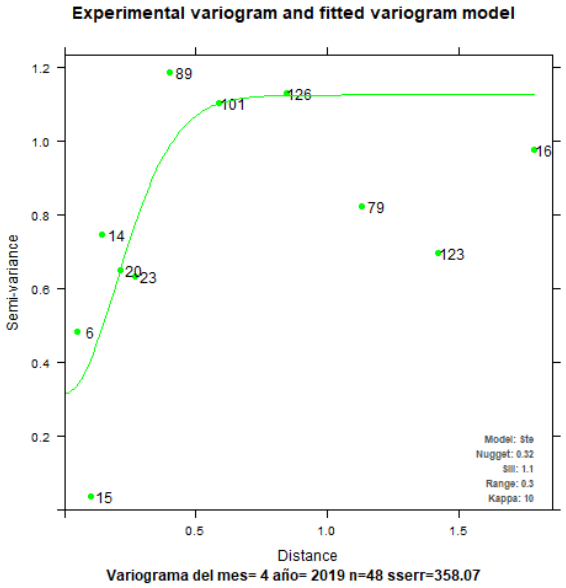


(m) Año: 2019, Mes: Enero

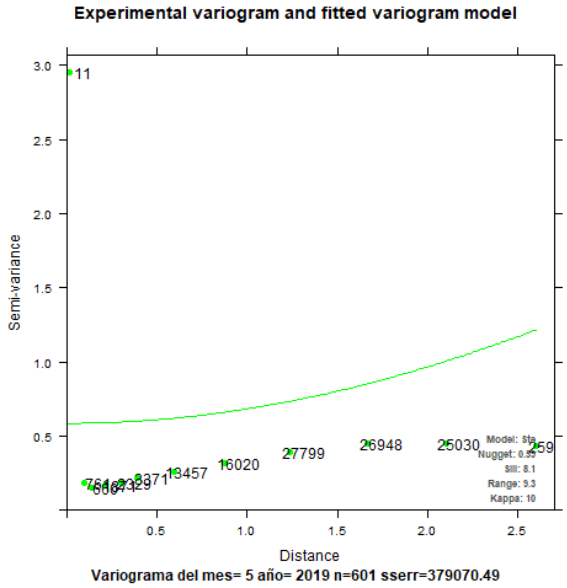


(n) Año: 2019, Mes: Marzo

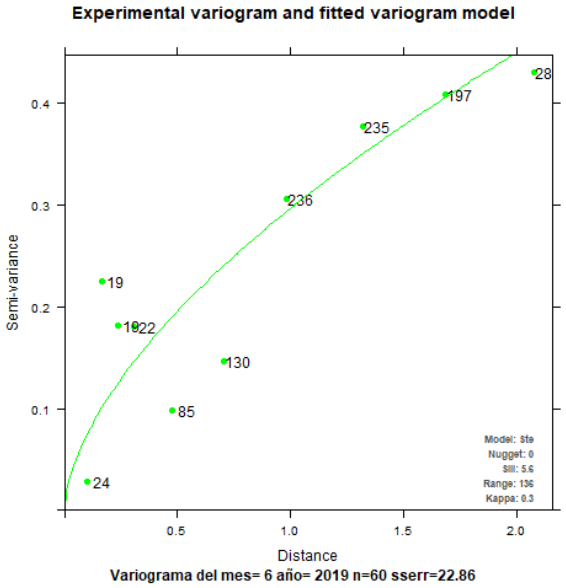
Figura. C.1: Variogramas modelados



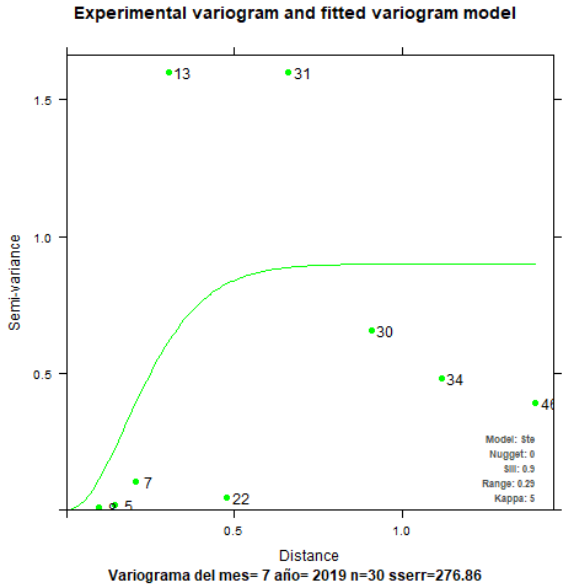
(o) Año: 2019, Mes: Abril



(p) Año: 2019, Mes: Mayo

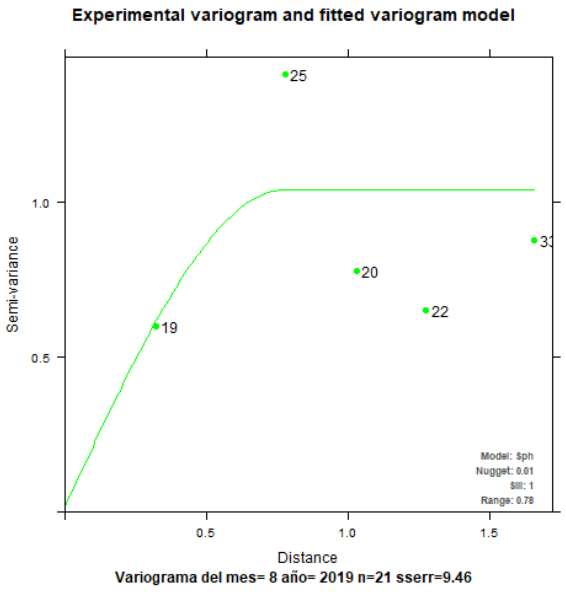


(q) Año: 2019, Mes: Junio

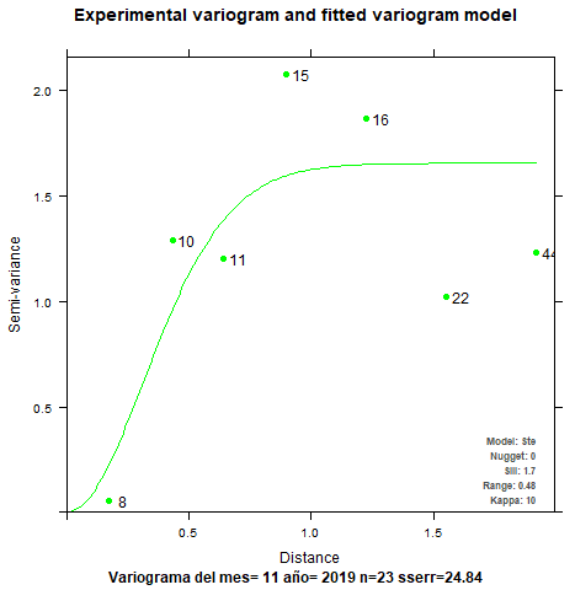


(r) Año: 2019, Mes: Julio

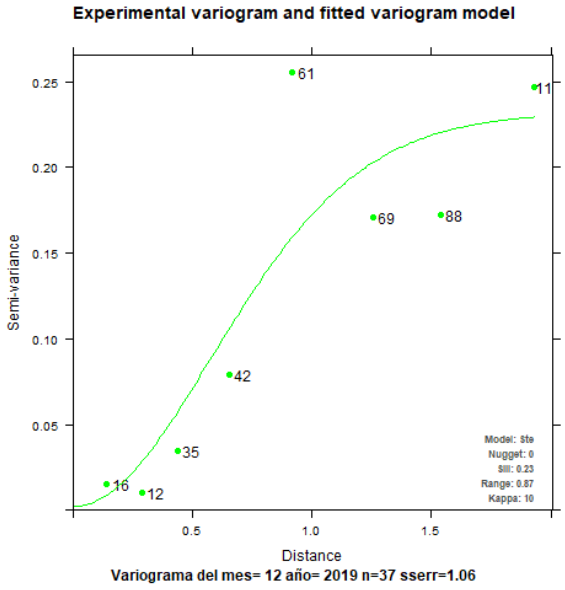
Figura. C.1: Variogramas modelados



(s) Año: 2019, Mes: Agosto



(t) Año: 2019, Mes: Noviembre



(u) Año: 2019, Mes: Diciembre

Figura. C.1: Variogramas modelados