Centro de Investigación en Matemáticas, A.C.

CIMAT

Estimación de traslape entre nichos climáticos para dilucidar la relación suculencia-aridez

T E S I S

Que para obtener el grado de
Maestro en Ciencias
con especialidad en
Probabilidad y Estadística

P r e s e n t a:
Melina Nohemí Del Ángel Martínez

Director de tesis:
Dr. Miguel Nakamura Savoy

Autorización de la versión final

Guanajuato, Gto. 6, noviembre del 2019.

Agradecimientos

Agradezco al Dr. Miguel Nakamura por su contribución en mi formación profesional mediante la dirección de este trabajo, pero también con su valioso tiempo extra en el cual compartió conmigo pasión por la estadística y su aplicación.

Gracias a la Dra. Tania Hernández por contribuir con el problema que dio lugar a esta investigación.

Gracias a todos mis profesores por formarme con rigor técnico y actitud disciplinada. En especial gracias al Dr. Juan Carlos Pardo por su tutela durante la especialidad y el primer año de maestría.

Agradezco profundamente a mi madre: por su amor incondicional, por siempre apoyar mis decisiones y por todas las veces que viajó sola a Gto. para verme.

Gracias a Diego por su entusiasmo y a Israel por sus recomendaciones y orientación. A mi cuñada Claudia y su familia por apoyarme en todo lo que necesité. A mi abuelito, mi abuelita, mi tías Lupe y Laura, mi tío Jesús y a toda mi familia porque a la distancia fueron siempre mi más grande soporte.

Gracias a Sonny por su compañía y por contenerme en momentos de frustración. A mis amigos Andy, Nigel, Karina, Daniel, Isaías, Juan y Salvador por haber compartido conmigo su conocimiento en tardes de estudio y también por todos los momentos

donde no hubo un libro de por medio.

Al Programa Nacional de Posgrados de Calidad del CONACyT agradezco el apoyo económico recibido durante los dos años de maestría.

Por último, gracias a la hermosa ciudad de Guanajuato y a la cálida gente con la que me crucé por haberme dado tres años llenos de felicidad.

Resumen

En este trabajo se expone el análisis estadístico realizado con base en estimación de traslape de nichos climáticos para elucidar la relación entre suculencia y aridez. Las plantas suculentas han estado comúnmente asociadas a climas áridos, cuyos componentes climáticos son temperatura alta y baja precipitación; sin embargo, la relación entre suculencia y aridez nunca ha sido probada formalmente con técnicas estadísticas. Este trabajo explora mediante reducción de dimensionalidad, ajuste de densidad por kernel y estimación del traslape entre densidades esta relación. Finalmente, se concluye con base en la comparación de intervalos de confianza de la estimación del índice de traslape que no hay evidencia estadística suficiente para afirmar que el nicho climático de las plantas suculentas sea diferente del nicho climático de las plantas no suculentas en niveles altos de aridez.

Palabras Clave

Nicho climático, Componentes principales ponderadas, Traslape entre funciones de densidad, Suculentas, Aridez, Intervalos de confianza, Bootstrap no paramétrico.

Índice

Introducción

El tema de esta tesis está situado en ecología, específicamente en plantas suculentas y su dinámica de supervivencia. Las suculentas son un grupo de plantas que ha desarrollado mecanismos de adaptación diferentes al resto, y que han sido llamados en conjunto como *síndrome de suculencia* (Gibson, 2012). Este síndrome abarca cambios metabológicos, morfológicos y fisiológicos que engloban modificaciones a nivel físico y químico; por ejemplo: fotosíntesis nocturna, reducción de órganos y metabolismo ácido. Estas modificaciones les han permitido a las suculentas ser más resistentes que otras plantas y sobrevivir en condiciones de clima extremos. Lo anterior las ha llevado a ser ejemplo icónico de adaptabilidad en materia de evolución.

La suculencia ha sido principalmente estudiada como mecanismo de adaptación a climas áridos, entendiendo como aridez ambientes con altas temperaturas y bajos niveles de precipitación. Además, existe la creencia de que las plantas suculentas son exclusivas de entornos áridos; es decir, que sus esfuerzos de adaptación tienen como finalidad sobrevivir en climas cálidos y secos. Dicha afirmación carece de evidencia formal, como libros o artículos de divulgación científica con mediciones explícitas. Derivado de lo anterior, se ha limitado el estudio de las plantas suculentas a éste clima, descartado la investigación en otro tipo de ambientes (como tropicales o fríos).

El interés particular de este trabajo es el estudio del nicho climático de las plantas suculentas y su interacción con la aridez, analizando formalmente la relación mediante técnicas estadísticas.

La Dra. Tania Hernández es catedrática del CONACyT[1] asignada al LANGEBIO[2] del IPN[3]. Ella ha integrado y depurado minuciosamente una base de datos de 201,000 registros de ejemplares (plantas) del núcleo *Caryophyllales* distribuidas en todo el mundo. El núcleo de plantas *Caryophyllales* es un grupo heterogéneo en cuanto a variedad de especies y familias. Más aún, éste núcleo cuenta con una amplia variedad de especies suculentas, lo que lo lleva a ser un grupo que replica bien su comportamiento. Por lo anterior, el núcleo *Caryophyllales* es útil para el estudio de la suculencia y sus mecanismos de adaptación. La labor estadística de este trabajo estará basada en los registros observacionales de la base de datos.

Es importante mencionar que las plantas en la base de datos no provienen de un muestreo probabilista, sino de observaciones incidentales. Los ejemplares registrados en la base de la Dra. Hernández integran 28 de las 38 familias del núcleo *Caryophyllales* y 5,070 de las 11,155 especies. Como consecuencia, se tienen problemas de sesgo y desbalance en muestra (ver Subsección 1.2). La base fue construida con información obtenida de dos repositorios. El primero de ellos es el compendio de datos para la biodiversidad, GBIF (Global Biodiversity Information Facility) y del consorcio para información espacial, CGIAR-CSI (Consortium for Spatial Information). El repositorio CGIAR-CSI contiene variables geográficas y climáticas, mientras que GBIF cuenta con variables biológicas. Luego, información disponible consta de 201,000 avistamientos de plantas (suculentas y no suculentas) asociados con variables climáticas, geográficas y biológicas.

En diciembre del año pasado se tuvo un acercamiento con la Dra. Hernández, en el cual planteó su escepticismo sobre la exclusividad de las plantas suculentas en cli-

[1] Consejo Nacional de Ciencia y Tecnología
[2] Laboratorio Nacional de Genómica para la Biodiversidad
[3] Instituto Politécnico Nacional

mas áridos y manifestó su interés en estudiar formalmente la asociación de aridez con suculencia. Con base en su experiencia en estudios filogenéticos de plantas suculentas, la Dra. Hernández ha observado que las éstas tienen un nicho climático variado, abarcando climas que no son necesariamente secos o cálidos. Ésto la ha llevado a pensar que el síndrome de suculencia es más complejo de lo que se ha pensado, lo cual contradice la creencia usual que (como se había mencionado anteriormente) carece de evidencia científica. Es por ésto que la Dra. Hernández quiere esclarecer la interacción entre suculencia y aridez a través de métodos estadísticos, aprovechando los avistamientos de plantas **Caryophyllales**. Poco trabajo se ha hecho en este sentido, lo que lleva a esta tesis a ser parte de una investigación original y progresista.

Derivado de lo anterior, el objetivo principal de la investigación es analizar la suposición de la Dra. Hernández y dotarla de formalidad estadística. Se realiza un trabajo conjunto entre el director de la tesis, el Dr. Miguel Nakamura y la tesista Melina Del Ángel en la parte de estadística y la Dra. Tania Hernández en la parte de ecología en un modelo de trabajo tipo consultoría estadística. El marco de trabajo comprende las etapas: entendimiento, diagnóstico, implementación y explicación de los resultados. Para tener un entendimiento profundo del problema es necesaria la aclimatación en el contexto ecológico, el cual se obtiene por medio de la interacción entre ambas partes. El diagnóstico se traduce en plantear estadísticamente el problema, para lo que se recurrirá a la exploración de datos, representaciones probabilistas e inferencia estadística. La ejecución y explicación de resultados se derivan del diagnóstico. Lo anterior significa que el mayor reto es tener un diagnóstico adecuado, claro y preciso de las herramientas estadísticas necesarias para la solución del problema y en este sentido los pasos descritos anteriormente son indispensables.

Se utilizarán diversas herramientas de análisis de datos. El vector aleatorio de interés será definido en la Subsección 1.3 y se compondrá de variables de diversa índole (ecológicas, geográficas y climáticas), lo que resultará en un vector multivariado. La reducción de dimensión facilitará la interpretación y dará lugar a la discusión.

Además, a lo largo del trabajo se utilizarán herramientas de descripción gráfica, con las cuales se aportarán conclusiones adyacentes que pueden no estar relacionadas directamente con la exploración de la suculencia con aridez, pero que serán de interés ecológico. Lo anterior es parte de las primeras etapas de la tesis y servirán para validar el supuesto de la Dra. Hernández, lo que justificará la continuación de la investigación.

Así mismo, para la asimilación del problema es necesario el concepto de *nicho climático*. En la literatura no existe una única definición de nicho climático (Townsend et al., 2011), para fines de este trabajo se dará una definición de éste como objeto matemático. La formulación involucrará estimación de densidades vía estadística no paramétrica. Se hará un esfuerzo para que el nicho climático recoja la mayor parte de la información disponible, y estará asociado con las variables de temperatura, precipitación y estacionalidad de lluvia. Posteriormente se realizará de manera concreta el análisis del nicho climático y el clima árido, utilizando bases de referencia y medidas estándar de aridez. Lo anterior conducirá de manera natural al planteamiento del problema pero desde un punto de vista estadístico, con el cuál serán claras las etapas para la resolución del mismo.

Al diagnóstico le procede la etapa de inferencia estadística. En esta etapa se ejercitarán habilidades de modelación estadística y análisis multivariado para la inspección de la dinámica entre grupos de plantas y tipos de ambiente; con ésto se obtendrán deducciones sobre el tipo de clima y adaptación, perfilando las primeras conclusiones. También, será necesaria la implementación computacional de métodos estadísticos, para lo cual se utilizará el software estadístico `R` (R Core Team, 2019). Éste software cuenta con una amplia gama de librerías, donde han sido implementados los métodos estadísticos más comunes, por lo que es útil y versátil para análisis estadísticos.

En todo momento la retroalimentación de la Dra. Hernández será fundamental. Ella analizará las gráficas y conclusiones que se vayan obteniendo y realizará la interpretación ecológica. De acuerdo a sus comentarios análisis se perfilará hasta llegar a

la explicación final, de dónde se obtendrán conclusiones sobre la adaptación biológica de las plantas suculentas al ambiente árido. Las conclusiones contradirán la creencia usual y traerán consigo un cambio de paradigma, éstas estarán sustentadas con análisis estadístico, el cuál las dotará de validez y formalidad.

La importancia de la investigación radica en el entendimiento de la complejidad del síndrome de suculencia y biodiversidad en zonas áridas. Rebatir la suposición entre suculencia y aridez representa un aporte científico a través del trabajo multidisciplinario. Además, en términos del trabajo de grado se maduran conceptos estadísticos, solución de problemas, se mejoran de habilidades de comunicación y de modelación.

Todo lo anteriormente dicho se resume en tres capítulos. El capítulo dos empieza dando una descripción detallada de la información disponible (base de datos), y obteniendo información complementaria que fungirá como marco referencial para la ubicar los diferentes climas del mundo. Posteriormente, se definirá nicho ecológico con base en funciones de densidad en dos dimensiones, para lo cual será necesario el uso de reducción de dimensionalidad. El capítulo dos concluye con el diagnóstico de la metodología estadística, el cuál hace uso del índice de traslape entre dos funciones de densidad. En el capítulo tres se define el índice de traslape de Weitzman, y se da un estimador no paramétrico del mismo. Posteriormente, se construyen intervalos de confianza para el índice de traslape y se da una interpretación en términos ecológicos. En el último capítulo se resumen las conclusiones y se da respuesta al cuestionamiento principal, *¿hay evidencia estadística suficiente para pensar que el nicho climático de las plantas suculentas es exclusivamente árido?*. Por último, se dan comentarios finales con base en lo identificado durante la realización de esta tesis. De ser retomados, los puntos propuestos en el último capítulo no sólo se mejoraría esta investigación, también traerían consigo nuevos y atractivos resultados.

CAPÍTULO 1

Análisis exploratorio

Este capítulo abarca la inspección de la base de datos (BD) e interpretación. Se inicia explicando las variables que contiene, posteriormente se habla del trabajo de minería realizado en las primeras etapas de la investigación; también, se habla de la construcción de la base de datos complementaria, la cual incluye información climática y geográfica del mundo. Asimismo, se resume el trabajo de descripción gráfica que fue fundamental para la interacción con la Dra. Hernández. Finalmente, se da una definición de *nicho ecológico* en términos de funciones de densidad, ésta es utilizada en el resto del trabajo. Los pasos anteriores ubican el problema en un contexto estadístico, el cuál es necesario para construir el diagnóstico final, con el cuál se concluye el capítulo.

El trabajo de minería busca atribuir un significado a los datos y explicar su comportamiento en contexto determinado. En las siguientes secciones se efectúa la exploración de la base de datos, la cuál comprende las etapas: descripción de las variables, inspección gráfica, preparación de información complementaria e interpretación del

análisis. Al término de este capítulo se tendrá dominio suficiente del problema para dar la definición de *nicho ecológico* con base en la información disponible.

1.1. Descripción de la base de datos

La base de datos contiene 201,000 registros de plantas (suculentas y no suculentas) del núcleo *Caryophyllales*. Las plantas están agrupadas en 28 familias y 5,670 especies. Estos registros fueron colectados por la Dra. Hernández durante varios años de investigación con plantas suculentas; representan observaciones incidentales de plantas en el mundo. Es decir, no fueron seleccionadas mediante métodos de muestreo, sino que se trata de ejemplares vistos y registrados en el repositorio GBIF. Cada registro está asociado con 38 variables (ver Tabla 1.1), divididas entre variables climáticas, ecológicas, geográficas y otras que no son de interés o son redundantes; por ejemplo, el país donde la planta fue encontrada es redundante con latitud y longitud. Como se mencionó en la introducción, estas plantas son ricas en especies de suculentas, por lo que son útiles para el estudio y son estos datos los que se estarán analizando en el resto de la tesis.

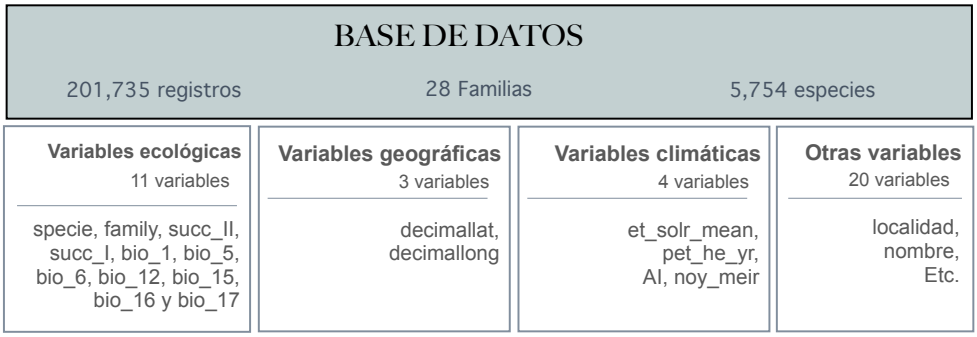


Figura 1.1: Agrupación de variables de la BD.

El grupo que incluye más variables es el ecológico, en él se incluyen siete variables estándar que usualmente se ocupan en el gremio biológico: Bio_1, Bio_5, Bio_6, Bio_12, Bio_15, Bio_16 y Bio_17. Estas variables cobran relevancia en el estudio de

la filogenia de las plantas suculentas, por lo que se incluirán también en el análisis estadístico. De ahora en adelante, cuando se quiera hacer referencia al grupo de las siete variables biológicas se le llamará simplemente *variables Bios*. Las variables Bios son de naturaleza continua, mientras que las variables Specie, Family, Succ_II y Succ_I son categóricas. No sólo en las primeras etapas de la tesis, sino en todas las etapas del proceso es será necesario tener presente el significado e interpretación de cada variable. A continuación se describe cada una:

Species: (Categórica) Especie.
Family: (Categórica) Familia.
Succ_I: (Categórica) Describe si la especie a la que pertenece la planta encontrada es considerada suculenta o no suculenta (succ, non_succ).
Succ_II: (Categórica) Describe si la familia a la que pertenece la planta encontrada es considerada suculenta, no suculenta o extremadamente suculenta (succ, non_succ, x_succ).
Bio_1: (Grados Celsius) Temperatura media anual.
Bio_5: (Grados Celsius) Temperatura máxima del mes más caliente del año en que fue encontrado el ejemplar.
Bio_6: (Grados Celsius) Temperatura mínima del mes más frío del año en que fue encontrado el ejemplar.
Bio_12: (Milímetros) Precipitación anual, es la suma de la precipitación mensual del año en que fue encontrado el ejemplar.
Bio_15: Estacionalidad de la precipitación.
Bio_16: (Milímetros) Precipitación total del trimestre más húmedo.
Bio_17: (Milímetros) Precipitación total del trimestre más seco.
Decimallat: (Grados) Latitud.
Decimallong: (Grados) Longitud.
Ai_yr: (Continua positiva) Índice global de aridez.
AI: (Categórica) Índice global de aridez categorizado.

Noy_meir: (Categórica) Aridez según la clasificación de Noy-Meir (1973).

Et_solr_mean: (Continua) Radiación solar terrestre media anual.

Pet_he_yr: (Continua) Evapotranspiración potencial media anual.

Más puntualmente, la variable Bio_15 hace referencia a qué tan bien distribuida está la lluvia a través del año. Consideremos las precipitaciones mensuales en un año PPT_i, con $i \in \{1, \ldots, 12\}$. Bio_15 se calcula como:

$$\text{Bio\_15} \quad = \quad \frac{\sqrt{\text{Var}(\text{PPT}_1, \ldots, \text{PPT}_{12})}}{1 + \left(\frac{\text{Bio\_12}}{12}\right)}.$$

Luego, Bio_15 será grande cuando la lluvia esté bien distribuida en el año y pequeño cuando esté concentrada en un período. Esta variable juega un papel importante para la adaptación de las suculentas, ya que si un área geográfica tiene menor Bio_15, implica que las plantas que habiten ahí tendrán que ser más eficientes en su consumo de agua y tener mayor capacidad de almacenamiento que en otros climas. Para más detalles sobre el cálculo del resto de las variables Bios consultar O´Donnell and Ignizio (2012).

Se considerarán dos medidas de aridez: AI y Noy_meir_aridity. La primera es una variable categórica de cinco niveles construida a partir de AI; dicha división es propuesta por Trabucco and Zomer (2018) y busca caracterizar el cima de acuerdo al nivel de aridez, valorando la precipitación, temperatura y evapotranspiración. Para el cálculo de explícito de AI, ver Trabucco and Zomer (2018). Por su parte, Noy_meir_aridity segmenta la precipitación total anual (Bio_12) en cuatro clases. Para poder analizar la suculencia en función de la aridez es necesario escoger una medida de la misma (pudiendo ser AI o Noy_meir_aridity). En la Sección 1.3 se realiza una comparación entre ambos índices para la sección del más útil en términos de los objetivos de la tesis. En la Tabla 1.1 se ilustran los criterios con los que son construidas AI y Noy_meir_aridity.

Valor de Ai_yr	Clima	AI
AI < 0.03	Hiper árido	HA
$0.03 \leq$ AI < 0.2	Árido	A
$0.2 \leq$ AI < 0.5	Semi árido	SA
$0.5 \leq$ AI < 0.65	Semi-húmedo seco	DSH
AI ≥ 0.65	Húmedo	H

Valor de Bio_12	Clima	Noy_meir_aridity
Bio_12 < 100mm	Extremadamente árido	extreme_arid
100mm $\leq$ Bio_12 < 250mm	Árido	arid
250mm $\leq$ Bio_12 < 500mm	Semi árido	no_arid
Bio_12 ≥ 500mm	No árido	semiarid

Cuadro 1.1: Segmentación de climas de acuerdo al nivel de aridez.

Descritas las variables en la BD, es necesaria una inspección más minuciosa. Se realizarán una serie de tareas para la limpieza, reducción y enriquecimiento de la BD. En la siguiente sección se explican a detalle dichas modificaciones y se concluye con la obtención de la base de datos depurada.

1.2. Tratamiento de la base de datos

Esta subsección abarca validación y condensación de la base de datos. La validación comprende actualización de la información con fuentes referenciales y supresión de registros no confiables, mientras que la condensación de la BD consiste en la reducción de la misma mediante el cambio de la unidad muestral. Los pasos en esta etapa aseguran la obtención de una base confiable, reduciendo errores de sesgo y tratando valores atípicos.

En la actualización de la base se rectificaron los nombres de especies. El sitio

Royal Botanic Gardens, Kew and Missouri Botanical (2013) provee una lista con el nombre y situación de todas las especies de plantas conocidas y se utilizó como marco de referencia para cotejar la variable Species. Resultó que la variable se encontraba desactualizada, lo que provocaba que algunos ejemplares de plantas se consideraran en diferentes especies, cuando en realidad correspondían a la misma. También, se tenían registros donde el nombre de la especie no estaba registrado. En total se actualizó el nombre de 112 especies, y se eliminaron 295 especies de las cuales no se conocía el nombre. Como se verá más adelante en esta sección, es fundamental que la variable specie sea confiable, pues las especies serán los individuos en el estudio. Después de esta actualización la variable asigna correctamente las plantas a su especie.

Las plantas de interés en el estudio son aquellas que sobreviven de manera natural. En ciertos lugares se inducen condiciones ambientales artificiales (como riego, abono y calor artificial), para mantener vivas a las plantas aunque éstas no estén en un ambiente con condiciones óptimas para sobrevivir. Ejemplos de estos lugares son los jardines botánicos, mercados, invernaderos y lugares de cultivo para comercialización (ver Figura 1.2). En particular, el núcleo *Caryophyllales* contiene a la especie *Amaranthus hybridus* (amaranto) y familia *Portulaca oleracea* (verdolagas), que por ser comestibles existen una gran cantidad de cultivos. La base de datos fue construida desde su extracción de GBIF para no considerar este tipo de plantas, de tal manera que las que se encuentran en la base de datos viven sin intervención humana.

Por otra parte, las observaciones de plantas registradas en GBIF provienen de avistamientos individuales, que pueden provenir de cualquier país del mundo. Recordemos que los registros en la base de datos no provienen de un muestreo probabilista: no se tiene control del número de observaciones que se tienen por país y ésto dependerá de la la administración de cada uno. Algunos países asignan mayor presupuesto que otros a tareas de investigación, y otros tienen centros de investigación especializados en materia ecológica. En consecuencia, en algunos países se tiene una muestra

Figura 1.2: Ejemplos de lugares con clima inducido.
(Figuras: Granma.travel, www.engormix.com)

grande y en otros países la muestra es escasa.

En la Figura 1.3 se muestra la dispersión de las 201,000 plantas. Como podemos ver, los cambios de muestra son drásticos: hay países donde se tiene mayor muestra y cruzando la frontera del país vecino la muestra es prácticamente nula. Como ejemplo podemos ver la frontera México-EUA, donde hay un fenómeno de sobremuestra en EUA y la frontera de Francia con España (ver Figura 1.4). Este no es un fenómeno natural, sino que la muestra está sesgada por la disparidad de los presupuestos asignados a investigación por cada país, no reflejando la dinámica de población de plantas.

El desbalance de las observaciones deriva en conclusiones erróneas. El desbalance de muestra sobrerepresenta la población de plantas en geografías donde se tiene mayor muestra, lo que implica que el clima donde habitan sea también sobre representado. Tomar la unidad muestral como ejemplares de plantas, no es una opción recomendable porque induce problemas de sesgo.

Una solución al problema anterior es cambiar la unidad muestral de plantas a especies. En consecuencia, la base a nivel especie perderá la ubicación geográfica de las plantas, pero mantendrá las variables ambientales. En particular, el nivel de suculencia es un atributo por especie, es decir, una especie pertenece a una y sólo una especie, por lo que esta variable será heredada. Las especies no se ven afectadas por la sobremuestra de plantas, ya que aunque se tenga una gran cantidad de plantas

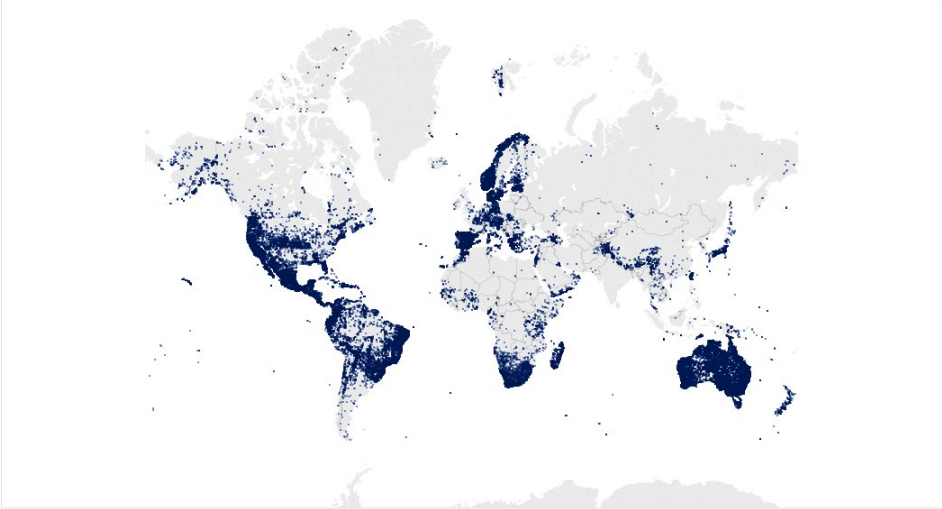


Figura 1.3: Observaciones de plantas en todo el mundo.

colectadas en una región, se estará representando únicamente el número de especies que habiten en esa región.

En esta nueva versión de la base de datos, donde cada renglón corresponde a una especie, se conservarán sólo las variables de importancia estadística. Se recalcula el valor de las variables a nivel especie, con base en la información de la base a nivel ejemplar. Las variables son recalculadas de acuerdo a su naturaleza: categóricas o continuas. Se tienen 5,063 especies diferentes y catorce variables de interés: diez

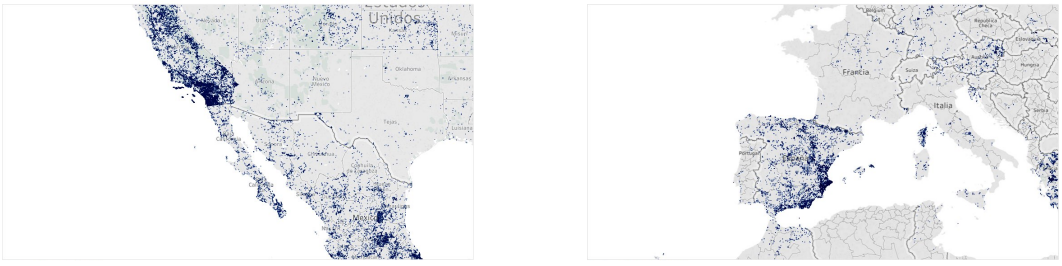


Figura 1.4: Problemas de imbalance en las observaciones, principalmente en las fronteras.

ecológicas, dos clasificaciones de índice de aridez y dos climáticas.

Las variables categóricas en la base de datos de especie se heredaron de la base original. Las variables categóricas en la base de datos a nivel ejemplar son Family, Succ_I y Succ_II. Una especie puede provenir sólo de una familia y tener asociado sólo uno de los tres niveles de suculencia, de tal manera que la variable Family en la base de datos a nivel especie será la familia a la que ésta corresponda y la variable Succ_II será el nivel de suculencia de la especie. Así pues, no hubo dificultad alguna para la asignación de esta variable.

Las variables continuas (Bios y Pet_he_yr) fueron reconstruidas en la base de datos a nivel especie, usando mediana empírica. Se aprovechó la información de los registros de plantas en la base original, obteniendo la mediana empírica basada en los registros de plantas por especie. Se seleccionó la mediana porque es una medida de tendencia central más robusta que la media, de modo que si se diera el caso de que una especie tenga pocas plantas registradas, no afecte si alguna de ellas tiene errores de medición o valores muy contrastantes. Entonces, se construyeron las variables continuas para cada especie tomando en cuenta las información de las plantas. Esta condensación dio lugar a la variable Freq. La variable Freq es un registro de cuántos ejemplares hay en cada especie:

$$\text{Freq}_i := \text{Número de registros de la especie } i \text{ en la base original.}$$

Así, si para la especie $i \in \{1, \ldots, 5063\}$ había originalmente x avistamientos de plantas, la variable de frecuencia será $\text{Freq}_i = x$. Es importante tener registro de con cuántos ejemplares fueron calculadas las variables de cada ejemplar y de esta manera poder asignarle un grado de credibilidad. Las cinco especies con mayor número de registros en la base se muestra en la Tabla 1.2.

Por otro lado, también se tiene una gran cantidad de especies de las cuáles sólo se tenía la observación de una planta en la base de datos original, por lo que sus variables no pueden tener la misma robustez que aquellas especies donde se contaba con la

Especie	Frecuencia	Freq / Total de registros
Chenopodium album	$2,878$	$1.443\,\%$
Atriplex prostrata	$1,625$	$0.815\,\%$
Portulaca oleracea	$1,490$	$0.747\,\%$
Petiveria alliacea	$1,377$	$0.691\,\%$
Stellaria media	$1,331$	$0.668\,\%$

Cuadro 1.2: Top 5 especies por número de ejemplares en la base de datos.

información de 500 o más plantas. Se hablará más de esta sutileza y sus implicaciones en la Sección 2, donde se tratará de manera especial a las unidades muestrales para que no se vean afectadas por lo anteriormente descrito.

1.3. Discriminación de variables de interés estadístico

Para el estudio estadístico se necesita definir el vector aleatorio de interés, dicho de otro modo, qué variables son estadísticamente relevantes de acuerdo a los objetivos del análisis. Las variables climáticas son de interés ecológico (Bio_1, ..., Bio_17), por lo que no se retirarán del estudio, pero se analiza si Pet_he_yr debe o no ser parte del juego de variables de estudio. También, es necesario definir la medida de aridez de referencia que se usará. Entre las variables disponibles se cuenta con dos índices de aridez: AI y Noy_meir_aridity, pero no es claro con cuál es mejor trabajar y se tendrá que seleccionar alguna. Se utilizaran clasificadores basados en la variable Succ_I y se usarán curvas ROC (Receiver Operating Characteristic) para la discriminación de variables. A continuación se describe a la regresión logística como método de clasificación y a las curvas ROC para evaluación de clasificadores.

De manera general, la definición de clasificador es la siguiente: considérese el vector (X, Y), donde X es una variable aleatoria sobre $\mathcal{X}$, y Y es una etiqueta en un

conjunto (discreto) de clases $K \in \mathbb{N}$. Un clasificador es una función

$$h : \mathcal{X} \to K,$$

y decimos que se comete error de clasificación cuando $g(X) \neq Y$. Ahora, en el caso de las plantas sea $Y =$Succ_I. El conjunto de etiquetas es $Y = \{0, 1\}$, donde 0 corresponde a la etiqueta *No suculentas* y 1 a *Suculentas*. El modelo de regresión logística (Hastie et al., 2008) es

$$\text{logit}(p_i) \quad = \quad \beta X^T,$$

donde $\mathbb{E}(y_i) = \mathbb{P}(y_i = 1) = p_i$ y X es una matriz de características. El modelo logístico retorna la probabilidad de pertenecer al grupo de suculentas para cada individuo, para ello se debe de seleccionar un umbral $t \geq 0$ para clasificar entre el grupo 0 y el 1. El clasificador basado en regresión logística para clasificación de suculencia a partir de las características X es

$$h^t(x) = \begin{cases} 1, & \text{si, logit}(p_i) > t \\ 0, & \text{C.O.C.} \end{cases}, \quad x \in X. \tag{1.1}$$

Con base en el modelo 1.1 se pueden definir varios clasificadores. Como el interés de esta sección está en discernir entre los índices de aridez y evaluar la utilidad de Pet_he_yr como predictor de la suculencia se definen los siguientes clasificadores:

$$h_1^t, \quad \text{donde logit}(p_i) = \beta^{(1)} X_1^T, \quad X_1 = \text{(AI)}, \tag{1.2}$$

$$h_2^t, \quad \text{donde logit}(p_i) = \beta^{(2)} X_2^T, \quad X_2 = \text{(Noy\_meir\_aridity)}, \tag{1.3}$$

$$h_3^t, \quad \text{donde logit}(p_i) = \beta^{(3)} X_3^T, \quad X_3 = \text{(Bios)}, \tag{1.4}$$

$$h_4^t, \quad \text{donde logit}(p_i) = \beta^{(4)} X_4^T, \quad X_4 = \text{(Bios, Noy\_meir\_aridity)}, \tag{1.5}$$

$$h_5^t, \quad \text{donde logit}(p_i) = \beta^{(5)} X_5^T, \quad X_5 = \text{(Bios, Pet\_he\_yr)}. \tag{1.6}$$

Una manera de comprar el desempeño de clasificadores es mediante curvas ROC. Las curvas ROC son una herramienta gráfica para evaluar el desempeño de modelos estadísticos que fungen como clasificadores basados en un umbral $t > 0$ (Zou et al., 2016). Han sido usados principalmente para casos donde el conjunto de etiquetas K es un subconjunto de los números naturales, como en este caso. Para la construcción, las curvas ROC hacen uso de la tasa de verdaderos positivos y falsos positivos resultantes de la comparación con el umbral t. Sea $x_1, \ldots, x_n$ una muestra aleatoria[1] de la variable aleatoria X, y $y_1, \ldots, y_n$ los valores de sus respectivas etiquetas. Asumamos que en el caso de la regresión logística, el umbral de clasificación es $t > 0$, y consideremos la variable y etiqueta (x_i, y_i). De acuerdo al clasificador, x_i será clasificada con etiqueta $y_i = 1$ si $p_i > t$. Derivado del procedimiento anterior, se definen las siguientes dos tasas:

$$\begin{aligned} \text{VP}(t) &= \mathbb{P}(p_i > t | y_i = 1) \quad \text{(Verdaderos positivos)}, \\ \text{FP}(t) &= \mathbb{P}(p_i > t | y_i = 0) \quad \text{(Falsos positivos)}, \end{aligned}$$

la tasa de verdaderos positivos es también llamada *Sensibilidad*, mientras que la tasa de verdaderos negativos es llamada *Especificidad*

$$\begin{aligned} \widehat{\text{VP}}(t) &= \frac{\sum_{i=1}^{n} \mathbb{I}_{(x_i=1)}}{\sum_{i=1}^{n} \mathbb{I}_{(y_i=1)}}, \\ \widehat{\text{FP}}(t) &= \frac{\sum_{i=1}^{n} \mathbb{I}_{(x_i=1)}}{\sum_{i=0}^{n} \mathbb{I}_{(y_i=1)}}. \end{aligned}$$

Una manera de evaluar las curvas ROC es mediante el área debajo de la curva (Zou et al., 2016), AUC por sus síglas en inglés[2]. El AUC va de 0.5 a 1, donde mientras más grande sea significa que el desmpeño del clasificador es mejor. Este será el método por el cuál se compararán las curvas ROC de los modelos propuestos.

Usando curvas ROC se evalúa el desempeño de los clasificadores de las Ecuaciones 1.2 a la 1.6. En la Figura 1.5 se muestra la evaluación de los modelos g_3^t y g_4^t. Queda

[1] Independiente e idénticamente distribuidas

[2] Area Under Curve (AUC)

evidenciado que ambos tienen prácticamente el mismo desempeño; es decir, cuando se han incluido las variables Bios resulta irrelevante agregar Noy_meir_aridity. Esto era de esperarse porque la variable de Noy_meir_aridity es construida con Bio_12, lo que lleva a duplicar información ya existente en Bios.
Luego, la variable Noy_meir_aridity es descartada del análisis sin afectar los resultados.

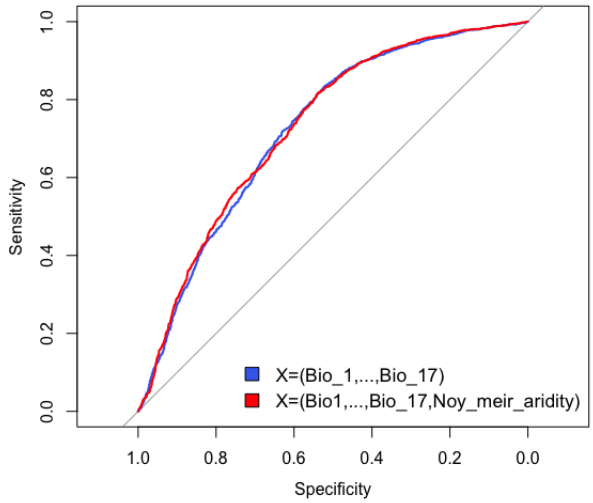


Figura 1.5: Curvas ROC para discriminación del índice de aridez.

Una vez descartada Noy_meir_aridity, se procede a evaluar la utilidad de Pet_he_yr en el estudio. En la Figura 1.6 se muestran las curvas ROC de los clasificadores g_3^t y g_5^t. El AUC resultó aumentar considerando el modelo Bios y Pet_he_yr, representando una mejoría con relación al otro modelo. Ésto significa que la variable Pet_he_yr aporta información valiosa para la clasificación de suculencia que no está contenida en las Bios. Finalmente, con base en el argumento anterior se decide continuar con el análisis considerando que el vector aleatorio de interés es

$$\mathrm{X} = (\text{Bio\_1}, \text{Bio\_5}, \text{Bio\_6}, \text{Bio\_12}, \text{Bio\_15}, \text{Bio\_16}, \text{Bio\_17}, \text{Pet\_he\_yr}), \quad (1.7)$$

y AI como variable climática de referencia para el nivel de aridez.

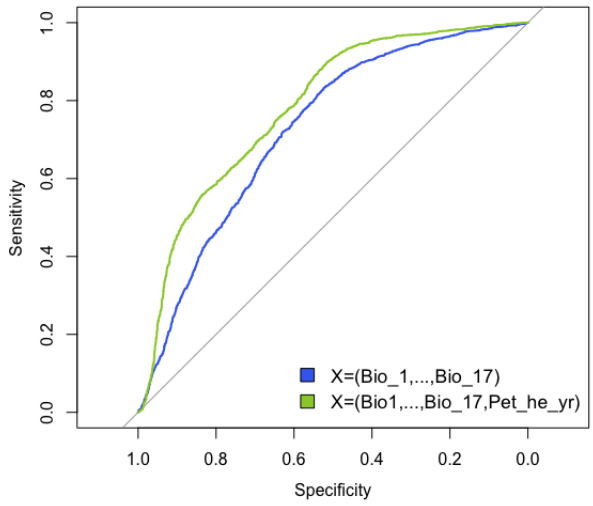


Figura 1.6: Curvas ROC para evaluación de Pet_he_yr.

1.4. Construcción de la base de datos complementaria

En el curso de la historia no se ha llegado a un consenso sobre la definición estricta de *nicho ecológico*, aunque han habido avances importantes en este sentido (Townsend et al., 2011). En términos generales, todas las definiciones apuntan a que el *nicho ecológico* es el subconjunto del espacio ambiente, donde se tienen condiciones óptimas se supervivencia para cierta especie. Dicho esto, el nicho ecológico de las plantas suculentas es un subconjunto de un espacio ambiente más general que en este caso será el espacio ambiente disponible en todo el mundo, al igual que el nicho climático de las plantas suculentas el espacio ambiente puede ser modelado estadísticamente. Tener un modelo del espacio ambiente disponible permitirá colocar en perspectiva el nicho climático de las plantas suculentas. Entonces, es necesaria la obtención de las ocho variables en el vector aleatorio de la Ecuación 1.7 de interés y la medida de referencia de aridez Pet_he_yr, pero esta vez en el espacio climático disponible. En esta subsección se describe el procedimiento para la obtención de datos almacenados en sistemas de información geográfica.

Los valores para todo el mundo son obtenidos de los sitios GBIF y CGIARCSI. Por una parte, de GBIF se obtienen la información de las siete variables climáticas (Bio_1, Bio_5, Bio_6, Bio_12, Bio_15, Bio_16, Bio_17), mientras que de CGIARCSI se obtiene el índice de aridez AI y el índice de evapotranspiración Pet_he_yr. La extracción de información se realiza mediante funciones implementadas en `R`, por medio de la librería `raster` (Hijmans, 2019). Lo anterior requirió habilidades de programación y el entendimiento de funciones para la extracción de información.

Es posible vincular con GBIF y CGIAR-CSI con `R` mediante la librería `raster`, específicamente con la función `getData`. La función `getData` extrae información de GBIF y CGIAR-CSI a nivel pixel. Para ello se crea una rejilla en el espacio de coordenadas de referencia (latitud y longitud) que define rectángulos/pixeles (ver Figura 1.7). `getData` extraerá la información que se solicite para cada uno de los rectángulos; por ejemplo, para un rectángulo extraerá la temperatura y precipitación locales si se solicita extraer Bio_1 y Bio_12. Se define una rejilla sobre todo el mundo y se extrae la información de las nueve variables, de tal manera que para cada rectángulo definido por la partición se tenga la misma información que se tiene para las plantas.

Figura 1.7: Ilustración de rejilla definida por segmentación de latitud y longitud. (https://intercienciassociales.wordpress.com)

Por otra parte, el número de observaciones en la base de datos dependerá de la resolución de la rejilla de latitud y longitud. La función `getData` permite descargar los datos a resolución 0.5min, 2.5min, 5min o 10min; mientras más baja sea la

resolución, mayor número de rectángulos que definirá la partición. Se decide usar resolución 2.5min, lo que genera 9,000,000 de pixeles. Con esta partición se obtienen las variables Bios, AI y Pet_he_yr. Posteriormente, se crea la variable AI a partir de Ai_yr de acuerdo a la clasificación mostrada en las Tablas 1.1, al igual que se hizo con la base de plantas.

Consolidando la información extraída, se obtiene una nueva base de datos de que consta de 8,870,094 observaciones y doce variables: latitud, longitud, Bio_1, Bio_5, Bio_6, Bio_12, Bio_15, Bio_16, Bio_17, Pet_he_yr, Ai_yr y AI. Esta base representa el clima en todos los lugares del mundo (salvo océanos) y será el marco climático de referencia.

CAPÍTULO 2

Definición de nicho ecológico y diagnóstico

Como se mencionó anteriormente, no existe una única definición de nicho ecológico. Sin embargo, para términos de este trabajo se entenderá como el subconjunto del espacio ambiente disponible donde cierto grupo de plantas sobreviven de manera natural. El vector aleatorio presentado en la Ecuación 1.7 vive en un espacio de ocho dimensiones, si estas variables definen el hábitat, ésto significa que el nicho climático es un subconjunto de un espacio de ocho dimensiones. Sin embargo, tener una dimensión de nueve representa pérdida de interpretabilidad, ya que no es ni siquiera posible visualizarlo. Entonces, se trabaja en reducir la dimensión del vector aleatorio para posteriormente dar la definición de nicho ecológico.

2.1. Reducción de dimensionalidad

En la Subsección 1.3 se seleccionaron las variables aleatorias que serán utilizadas para definir el nicho ambiental. El vector aleatorio

$$\text{X} = (\text{Bio\_1, Bio\_5, Bio\_6, Bio\_12, Bio\_15, Bio\_16, Bio\_17, Pet\_he\_yr}) \quad (2.1)$$

está definido sobre $\mathbb{R}^9$, pero trabajar en esta dimensión tiene el inconveniente de que la interpretación es complicada, y al estar trabajando con personas que no son propiamente del área estadística, la facilidad de interpretación es deseable. El análisis multivariado provee técnicas de reducción de dimensionalidad, que en las siguientes subsecciones se aplicarán al vector X para solucionar el problema de alta dimensión, facilitando el entendimiento.

2.1.1. Componentes principales ponderadas (WPCA)

Esta subsección tiene como objetivo dar los elementos del método de componentes principales ponderadas WPCA[1]. Se comienza describiendo de manera general el análisis de componentes principales PCA[2], para después definir WPCA. El método WPCA será utilizado en la siguiente subsubsección para reducción de dimensión del vector de la Ecuación 1.7.

El procedimiento PCA tiene como objetivo la reducción de dimensión de vectores aleatorios (Hastie et al., 2008). PCA es un método que no ocupa supuestos distribucionales y busca explicar la estructura de relación entre un juego de variables de un vector aleatorio $Z = (Z_1, Z_2, \ldots, Z_n)$ a través de combinaciones lineales. La construcción del método depende de la matriz de covarianzas Σ de Z. Algo recomendable en esta técnica, es escalar variables del vector Z cuando tienen diferentes escalas (magnitudes); es decir, si $\mu = \mathbb{E}(Z_i)$ y $\sigma_{ii} = \mathbb{E}\left[(Z_i - \mu_i)^2\right]$ la i-ésima entrada centrada y escalada de Z es

$$X_i = \frac{Z_i - \mu_i}{\sigma_{ii}}, \quad \text{donde } i \in \{1, \ldots, n\},$$

[1]Weighted Principal Component Analysis
[2]Principal Component Analysis

y esto define un nuevo vector aleatorio $X = (X_i, \ldots, X_n)$ tal que su matriz de varianzas y covarianzas es $\Sigma = X^T X$. Sean las parejas de eigenvalores y eigenvectores de Σ, $(\lambda_1, \mathbf{u_1}), \ldots, (\lambda_p, \mathbf{u_p})$, donde $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ y $\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,p})$. Entonces, la i-ésima componente principal es

$$Y_i = \mathbf{u}_i^T X, \quad \text{con } i \in \{1, \ldots, p\}. \tag{2.2}$$

Además, se sabe que la varianza de las componentes principales es igual al eigenvalor que le corresponde:

$$\text{Var}(Y_i) = \lambda_i, \quad \text{con } i \in \{1, \ldots, p\}.$$

Por otro lado, los vectores propios $\mathbf{u}_i$ son dignos también de análisis, ya que por la Ecuación 2.2, los coeficientes de $\mathbf{u}_i$ dictan el grado de importancia de las variables en la $i-$ésima componente.

Otra manera de interpretar las componentes principales, es mediante gráficas biplot. Las gráficas biplot están basadas en la relación 2.3, que se da cuando las variables aleatorias X y Y están centradas. Si el ángulo entre X y Y es cercano a cero, la correlación entre las variables será cercana a uno, mientras que si el ángulo es cercano a $\frac{\pi}{2}$ la correlación entre X y Y es cercana a cero. Para más detalles sobre la construcción e interpretación de las gráficas biplot ver Kroonenberg (2008).

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{< X - \mathbb{E}(X), Y - \mathbb{E}(Y) >}{||X - \mathbb{E}(X)||||Y - \mathbb{E}(Y)||} = \cos(\angle(X,Y))^3.$$

Ahora, en el análisis de componentes principales convencional todos los individuos tienen la misma importancia. Esto no siempre es cierto, ya que en algunas ocasiones se le quiere asignar mayor importancia a un individuo que a otro. Ésto puede deberse muchos factores, como la fiabilidad de las observaciones o que los individuos no tienen la misma trascendencia en el contexto del problema. Una manera

de calibrar la importancia diferenciada por individuo en el análisis es ponderar las observaciones, empleando una matriz diagonal de pesos

$$W = \begin{bmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & w_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & w_n \end{bmatrix}, \quad 0 \leq w_i \leq 1, \quad \sum_{i=1}^{n} w_i = 1,$$

y modificar la matriz de covarianzas $\Sigma^W = X^T W^T W^X$, de tal manera que w_i calibre la importancia del $i-$ésimo individuo. A este ajuste se le ha denominado en la literatura **Componentes Principales Ponderadas** Yue and Tomoyasu (2005), WPCA. Este análisis permite calibrar el método y puede ser útil para reducir errores de sesgo. En la siguiente sección se empleará WPCA para ponderar las especies de plantas de acuerdo al número de ejemplares que se tienen de ella.

2.1.2. Aplicación de WPCA para ambiente de plantas

Las observaciones en la base de datos de plantas no son igualmente confiables. Cada especie cuenta originalmente con un número diferente de observaciones de plantas (variable Freq), lo que provoca que los valores de las variables del vector X sean calculadas sobre una base diferente para cada especie. Ésto conlleva a que las especies que cuentan con mayor número de ejemplares son más confiables que las que constaban de uno o dos ejemplares. El método WPCA, a diferencia de PCA ordinario, protege contra aberrantes por lo cuál en esta tesis se utiliza WPCA.

Se aprovecha la información de la variable Freq para asignar pesos diferentes a cada especie. Cada observación $i \in \{1, \dots, 5063\}$ tiene asignado un valor en la base de datos para Freq_i. La ponderación de la $i-$ésima especie en este caso se calcula de manera uniforme, es decir:

$$w_i \quad = \quad \frac{\text{Freq}_i}{\sum_{i=1}^{n} \text{Freq}_i},$$

Claramente, $0 \leq w_i \leq 1, \quad \sum_{i=1}^{n} w_i = 1$. La asignación de w_i por especie se dio de manera natural por la construcción de la base de datos de plantas y representa una manera razonable de aprovechar la información con WPCA.

Como extensión de este trabajo podría considerarse una manera no uniforme de asignar pesos, incluso que éstos dependieran de otros factores y no sólo del número de ejemplares; sin embargo, para fines de este trabajo y como primer ejercicio se tomarán los pesos de la manera ya mencionada.

Luego, se usa WPCA en el vector aleatorio X previamente centrando y escalando las columnas, ya que las columnas tienen diferentes escalas. La composición de varianza se muestran en la Tabla 2.1. En adelante se denotarán las componentes principales ponderadas proyectadas mostradas en la Ecuación 2.2, construidas con los datos de plantas como $Y_1^{\text{P}}, Y_2^{\text{P}}, Y_3^{\text{P}}, Y_4^{\text{P}}, Y_5^{\text{P}}, Y_6^{\text{P}}, Y_7^{\text{P}}, Y_8^{\text{P}}$.

WPCA$_i$	Varianza	Varianza explicada por la WPCA$_i$	Acumulado
Y_1^{P}	3.919	48.99 %	48.99 %
Y_2^{P}	2.702	33.77 %	83.76 %
Y_3^{P}	0.956	11.95 %	94.72 %
Y_4^{P}	0.229	2.86 %	97.58 %
Y_5^{P}	0.114	1.43 %	99.02 %
Y_6^{P}	0.059	0.74 %	99.76 %
Y_7^{P}	0.011	0.13 %	99.90 %
Y_8^{P}	0.007	0.09 %	100 %

,

Cuadro 2.1: Componentes principales ponderadas para datos de plantas.

En problemas ecológicos típicamente se usan las componentes que acumulen 80 %

de varianza. En la interacción con la Dra. Hernández se acordó usar sólo las primeras dos componentes principales, ya que aunque se tendría mayor varianza explicada usando tres en lugar de dos, usando sólo dos se gana simplicidad en la interpretabilidad. Las primeras dos componentes principales acumulan 83.76 % de la varianza total, que de acuerdo a lo anterior es un porcentaje de variabilidad aceptable. Ambas componentes están definidas sobre los reales, por lo que el plano de proyección es $\mathbb{R}^2$.

Siguiendo con el análisis, se buscará dotar de significado a las componentes de acuerdo a los coeficientes de los vectores propios. Los primeros tres vectores propios de la matriz de covarianzas Σ se anotan en el Cuadro 2.2.

	$\mathbf{u_1}$	$\mathbf{u_2}$	$\mathbf{u_3}$
Bio_1	0.487	0.084	0.197
Bio_5	0.437	−0.169	0.290
Bio_6	0.431	0.226	0.151
Bio_12	0.008	0.602	−0.056
Bio_15	0.314	0.001	−0.779
Bio_16	0.101	0.566	−0.262
Bio_17	−0.207	0.477	0.397
Pet_he_yr	0.484	−0.058	0.136

Cuadro 2.2: Primeros tres vectores propios del análisis WPCA.

Primera componente principal: Y_1^{P} aumenta cuando Bio_1, Bio_5 o Bio_6 aumentan, es decir, cuando la temperatura lo hace. Bio_15 tiene un coeficiente positivo y de 0.31, por lo que Y_1^{P} también aumentará cuando Bio_15 sea grande, es decir, cuando el coeficiente de variación de la estacionalidad de lluvia se agrande, lo cual es equivalente a que la lluvia esté distribuida de manera más homogénea a lo largo del año.

Segunda componente principal: Y_2^{P} aumenta cuando Bio_12, Bio_16 o Bio_17 aumentan, es decir, cuando la precipitación aumenta. Esta componente será en escencia el resumen de la precipitación anual en cantidad.

Tercera componente principal: Y_3^{P} se reduce cuando Bio_15 es grande en valor absoluto, es decir, cuando la lluvia está dispersa a través del año.

La gráfica biplot con las observaciones de plantas se incluye en la Figura 2.1.

Se observa que las Bios relacionadas a temperatura, así como la variable de vapotranspiración tienen ángulos cercanos a cero, por lo que la correlación entre ellas es alta. Éstas se encuentran alineadas con la primera componente principal, mientras que las variables relacionadas con precipitación están alineadas en la segunda

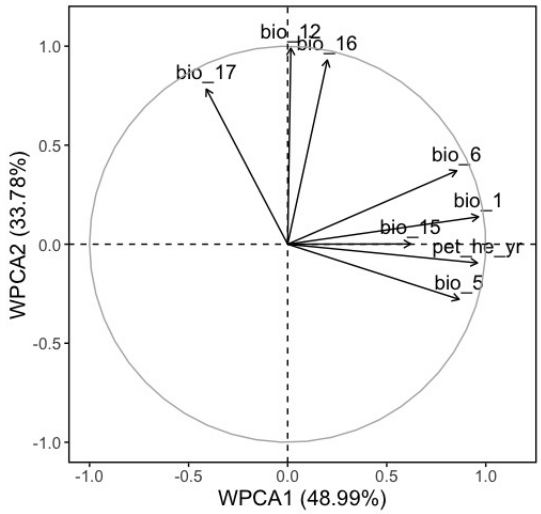


Figura 2.1: Gráfica biplot para plantas.

componente principal y también tienen ángulos pequeños entre ellas, por lo que la correlación es alta.

En dos dimensiones ya se tienen representaciones gráficas (ver Figura 2.2). Debido a la gran cantidad de puntos, en el diagrama de dispersión no se puede apreciar la intensidad de puntos en el plano. Una manera más efectiva de modelar el nicho ecológico de las plantas es a través de regiones de máxima densidad, las cuales serán explicadas más ampliamente en la Subsección 2.2.1. Éstos permitirán ver dónde se encuentran las zonas de mayor densidad de plantas en el plano $\mathbb{R}^2$.

2.1.3. Aplicación de WPCA para ambiente disponible

Es posible proyectar cualquier observación que provenga del vector aleatorio X en el plano $\mathbb{R}^2$ mediante u_1 y u_2. Entonces, se pueden proyectar las observaciones de la base de datos del ambiente disponible y se estaría proyectando los climas ocupados del mundo en el mismo espacio ambiental en el que se proyectan las especies, haciendo ambas proyecciones comparables.

Al proyectar las observaciones del ambiente disponible puede que algunas se tras-

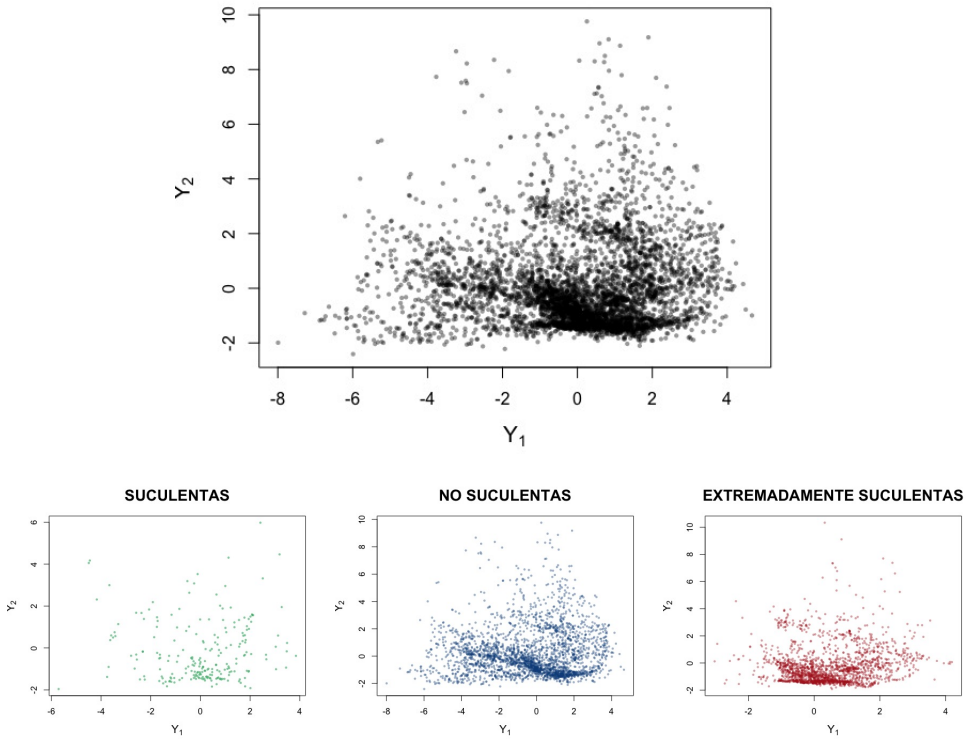


Figura 2.2: Gráfica de dispersión de plantas.

lapen. Las direcciones que provee el análisis de WPCA con las especies no son necesariamente las de mayor variabilidad del ambiente disponible, y en estas nuevas direcciones las observaciones se superponen, se acumulan, se distancian y da lugar al concepto de densidad de aridez sobre el plano $\mathbb{R}^2$. Aunque la aleatoriedad de estos datos es debida principalmente a errores de medición y no a cambios en el clima, porque en este caso se está considerando sólo una ventana de tiempo, insuficiente para cambiar las características del clima.

Dicho lo anterior, sea el vector aleatorio X definido en la Subsección 2.1, y sean $\mathbf{X}^{\text{AD}} = (x_1, \ldots, x_{8{,}870{,}094})$ la muestra aleatoria X que corresponde al ambiente disponible. Esta muestra aleatoria está asociada a los 8,870,094 pixeles que hay en la base de datos del mundo, en la que cada registro representa un rectángulo definido por la partición de latitud y longitud. La proyección por componentes principales de $\mathbf{X}^{\text{AD}}$ en

el plano $\mathbb{R}^2$ usando los vectores propios del análisis de la Subsección 2.1.2 es

$$Y_i^{\text{AD}} = \mathbf{u}_i^T X^{\text{AD}}, \quad i \in \{1, \ldots, 8\}.$$

Las proyecciones de componentes principales ponderados de las especies son denotados por Y_i^{P}. En este caso las proyecciones considerando el ambiente disponible se denotarán Y_i^{AD}.

2.2. El nicho ecológico en dos dimensiones

En esta sección se describirá probabilísticamente el nicho ecológico de las plantas suculentas y ambiente disponible, mediante funciones de densidad. Se pretende que en la descripción intervengan todas las variables que aporten información sobre la preferencia climáticas de las plantas, es decir, las seleccionadas en la Sección 1.2. En la sección anterior se proyectaron tanto las observaciones de especies como los rectángulos del ambiente disponible en el plano $\mathbb{R}^2$ o WPCA1$\times$WPCA2. Por medio de WPCA se extrajo la información más relevante de las ocho variables, convirtiéndose en una buena síntesis en dos dimensiones. Luego, la modelación será sobre WPCA1$\times$WPCA2.

El ajuste de densidad por kernel es un método no paramétrico para aproximar la función de densidad. Los estimadores no paramétricos son útiles cuando no se tiene mayor información de lo que se quiere modelar que una muestra aleatoria. A continuación se define la estimación de densidad por kernel.

Sea un vector aleatorio bivariado $Z = (X, Y)$ cuyo soporte es sop$(Z) = \mathcal{X} \times \mathcal{Y}$. Sean $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ realizaciones de Z, $h > 0$ una constante conocida como *ancho de banda*, y K una función kernel, es decir,

$$\begin{aligned} \int_{\mathcal{X}} \int_{\mathcal{Y}} K(x, y) dx dy &= 1, \\ K(x, y) &\geq 0, \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{aligned}$$

Sea $(x, y) \in \mathcal{X} \times \mathcal{Y}$, el estimador de densidad por kernel de la función de densidad $f_X(x, y)$ es

$$\widehat{f}(x, y) \quad = \quad \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}, \frac{y - y_i}{h}\right),$$

este estimador tiene la propiedad de converger a $f(x, y)$ conforme n tiende a infinito y h se va a cero. Para ello, no es importante la elección particular del kernel. La formulación tiene una extensión natural para el caso de estimar una densidad multivariada. El ancho de banda es más determinante, y existe una varidad de métodos para especificarlo con base en la muestra misma.

Actualmente, casi todos los software estadísticos implementan el método de estimación de densidad por kernel. En R, la librería `ks` provee herramientas para estimar funciones de densidad univariadas o multivariadas (Duong et al., 2019). La función con la que se estiman densidades es `kde`, la cual toma como argumento una muestra aleatoria, un ancho de banda y un kernel. Si el ancho de banda no se especifica, `kde` selecciona uno óptimo empíricamente (Wand and Jones, 1994). En lo siguiente, la elección del ancho de banda se delegará a la librería `ks` y el kernel que se adoptará es gaussiano estándar con entradas independientes: la especificación por default.

El estimador de densidad por kernel representa una técnica adecuada para estimar el nicho de las plantas y ambiente disponible. La muestra aleatoria se tomará en ambos casos como la proyección de las especies y pixeles en el plano WPCA1$\times$WPCA2; de ésta manera, se tiene un resumen probabilista de las condiciones climáticas, que por su naturaleza considera la variabilidad en los datos y permite saber cuáles son las zonas con mayor densidad. En la siguiente subsección se realiza la modelación mediante ajuste de densidad no paramétrico.

2.2.1. Ajuste de densidad por kernel para ambiente de plantas

En esta subsección se realiza aproximación de densidad por kernel para las plantas, el ajuste de densidad será considerado el nicho climático. La estimación de densi-

dad por kernel para las plantas se basa en las primeras dos componentes principales, la cuál se representa mediante el vector $Y^{\mathrm{P}} = (Y_1^{\mathrm{P}}, Y_2^{\mathrm{P}})$ en el caso de las especies. La muestra de Y^{P} son las especies proyectadas por WPCA encontradas en la Subsección 2.1.2, y se denotarán $\mathbf{Y}^{\mathrm{P}} = (y_1, \ldots, y_{5,026})$. El kernel seleccionado es un kernel gaussiano estándar con entradas independientes, es decir:

$$K(x, y) = K(x)K(y) = \frac{1}{\sqrt{2\pi}} e^{\left\{\frac{1}{2}x^2\right\}} \frac{1}{\sqrt{2\pi}} e^{\left\{\frac{1}{2}y^2\right\}}.$$

Luego,

$$\begin{aligned}
\widehat{f}(x, y) &= \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}, \frac{y - y_i}{h}\right) \\
&= \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \frac{1}{h} K\left(\frac{y - y_i}{h}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\sqrt{2\pi}} \exp\left\{\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right\} \frac{1}{h\sqrt{2\pi}} \exp\left\{\frac{1}{2}\left(\frac{y - y_i}{h}\right)^2\right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} Q(x)Q(y),
\end{aligned}$$

donde $Q(x)$ es el kernel de una distribución normal $N(x_i, h^2)$. Se diferenciarán las especies por nivel de suculencia. Considérense las siguientes muestras aleatorias:

$$\begin{aligned}
\mathbf{x}^{\mathrm{XS}} &= (x_1^{\mathrm{XS}}, x_2^{\mathrm{XS}}, \ldots, x_{m_1}^{\mathrm{XS}}), & m_1 &= 2,028, \\
\mathbf{x}^{\mathrm{S}} &= (x_1^{\mathrm{S}}, x_2^{\mathrm{S}}, \ldots, x_{m_1}^{\mathrm{S}}), & m_2 &= 224, \\
\mathbf{x}^{\mathrm{NS}} &= (x_1^{\mathrm{NS}}, x_2^{\mathrm{NS}}, \ldots, x_{m_3}^{\mathrm{NS}}), & m_1 &= 2,811.
\end{aligned}$$

Sea (x, y) un punto en $\mathbb{R}^2$. Las estimaciones de las funciones de densidad plantas suculentas, no suculentas y extremadamente suculentas respectivamente son las siguientes:

$$\begin{aligned}
\widehat{f}_{\text{XS}}(x,y) &= \frac{1}{nh^2}\sum_{i=1}^{n_1} K_2\left(\frac{x-x_i^{\text{XS}}}{h}, \frac{y-y_i^{\text{XS}}}{h}\right), \\
\widehat{f}_{\text{S}}(x,y) &= \frac{1}{nh^2}\sum_{i=1}^{n_1} K_2\left(\frac{x-x_i^{\text{S}}}{h}, \frac{y-y_i^{\text{S}}}{h}\right), \\
\widehat{f}_{\text{NS}}(x,y) &= \frac{1}{nh^2}\sum_{i=1}^{n_1} K_2\left(\frac{x-x_i^{\text{NS}}}{h}, \frac{y-y_i^{\text{NS}}}{h}\right).
\end{aligned}$$

Para facilitar notación se usará

$$\widehat{f}_I, \qquad I \in \{\text{S, NS, XS}\}.$$

Para la definición de nicho ecológico en dos dimensiones, es necesaria la noción de *región de máxima densidad*, HDR por sus siglas en inglés[4]. La definición del $100(1-\alpha)\,\%$ HDR en dos dimensiones, es la siguiente,

Definición 1. *Sea $f_X(x)$ la función de densidad de la variable aleatoria X sobre $\mathbb{R}^2$. Entonces, la $100(1-\alpha)\,\%$ Highest density región es el subconjunto $R(c)$ del espacio muestral de X tal que,*

$$R(c) = \{x \in \mathbb{R}^2 | f(x) \geq c\},$$

donde c es la constante más grande tal que $\mathbb{P}(X \in R(c)) \geq 1-\alpha$. Ver Hyndman (1996).

Nótese que de la definición se sigue que de entre todas las regiones con probabilidad de cobertura $(1-\alpha)\,\%$ la HDR es la región con menor área sobre $\mathbb{R}^2$ Hyndman (1996). Por consiguiente, la amplitud del HDR constituye un buen indicador de la variabilidad de X.

[4]Highest Density Region (HDR)

Luego, la región de máxima densidad es una subregión de $\mathbb{R}^2$. El $100(1-\alpha)\,\%$ HDR de f_I sobre el plano definido por las dos primeras componentes principales, $Y_1 \times Y_2$ representa el ambiente que ocupan el $100(1-\alpha)\,\%$ de las plantas del tipo $I = \{\text{XS, X, NS}\}$. Es decir, la caracterización del clima que habitan, que por construcción representa un buen resumen de todas las características biológicas y climáticas en dos dimensiones. De hecho, esto proporciona una definición operativa de nicho ecológico para nuestros propósitos. La interpretación es que el nicho es un conjunto donde se concentra una cantidad especificada y grande (por ejemplo, 95 %) de ocurrencias.

La 95 % HDR de $f_{\text{XS}}, f_S, f_{\text{NS}}$ sobre el plano $Y_1 \times Y_2$ es útil para la interpretación ecológica. La HDR representa la región donde habita el 95 % de las plantas, y es una representación del hábitat de las plantas en términos de precipitación, temperatura, estacionalidad y evapotranspiración. Ésto facilita la comparación entre los tres nichos: plantas suculentas, extremadamente suculentas y no suculentas. Además, permite visualizar la ubicación y ocupación de los nichos en todo el plano $Y_1 \times Y_2$. Lo anterior es una forma sencilla de visualizar f_I y recoge en esencia el trabajo hecho en cuanto reducción de dimensionalidad y modelación del nicho ecológico como función de densidad f_I.

Sean $\text{HDR}_{95}(\widehat{f}_{\text{XS}}), \text{HDR}_{95}(\widehat{f}_{\text{S}}), \text{HDR}_{95}(\widehat{f}_{\text{NS}})$ las regiones de densidad máxima basadas en las estimaciones de densidad por kernel para plantas extremadamente suculentas, suculentas y no suculentas respectivamente. En la Figura 2.3 se muestra el perímetro de $\text{HDR}_{95}(\widehat{f}_I)$.

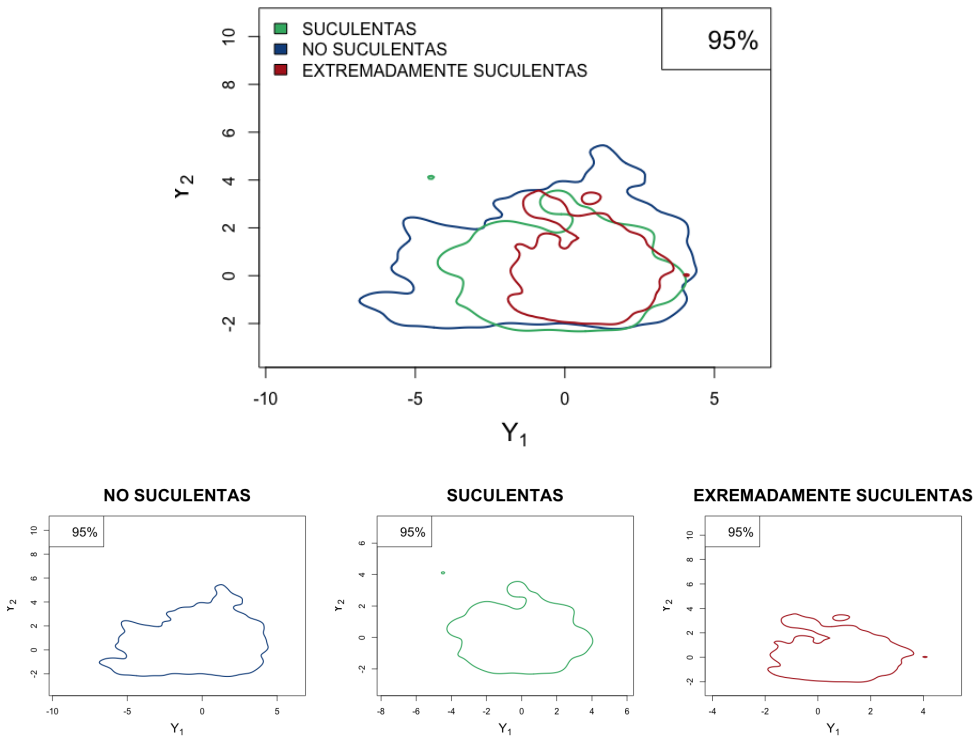


Figura 2.3: Perímetro de región de máxima densidad para plantas.

De acuerdo a los comentarios hechos sobre Y_1 y Y_2 en la Subsección 2.1.2, se puede interpretar $\text{HDR}_{95}(\widehat{f_I})$ como función de las condiciones climáticas. La región de mayor aridez (precipitación baja y temperatura alta), se encuentra en el cuarto cuadrante[5] (ver Figura 2.4). La zona de mayor humedad está en los cuadrantes uno y dos; la más seca, en los cuadrantes tres y cuatro, *etc*. De acuerdo a lo que se piensa sobre las plantas suculentas y su tipo de hábitat se espera que las plantas suculentas dominen el área de mayor aridez, mientras que las plantas no suculentas no tienen oportunidad de supervivencia en esa zona. Viendo los contornos de $\text{HDR}_{95}(\widehat{f_I})$ se aprecia que la densidad de las plantas extremadamente suculentas es casi nula en esta zona, mientras que las plantas no suculentas (de hecho) pueblan esta zona con mayor intensidad. Lo anterior apunta a que las plantas suculentas no son en realidad tan

[5] Esquina inferior derecha

áridas como se piensa, apoyando lo que la Dra. Hernández ha encontrado mediante análisis evolutivos.

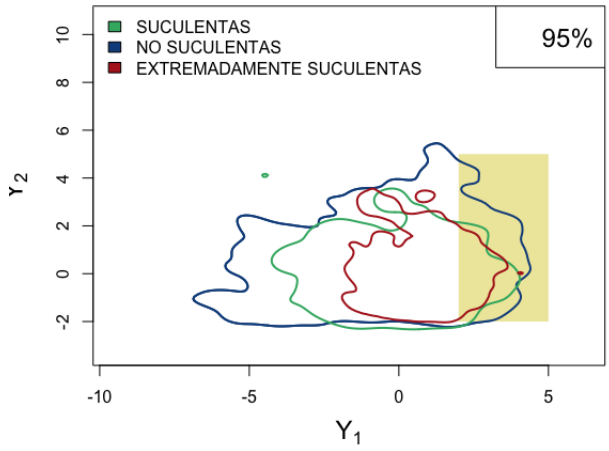


Figura 2.4: HDR 95 %.

2.2.2. Ajuste de densidad por kernel para ambiente disponible

Se modela el ambiente disponible con funciones de densidad de acuerdo al nivel de aridez. Sean las muestras aleatorias,

$$\begin{aligned}
\mathbf{x}^{\text{HA}} &= (x_2^{\text{HA}}, x_2^{\text{HA}}, \ldots, x_{n_1}^{\text{HA}}), n_1 &&= 570,066,\\
\mathbf{x}^{\text{A}} &= (x_2^{\text{A}}, x_2^{\text{A}}, \ldots, x_{n_1}^{\text{A}}), n_2 &&= 1,379,359,\\
\mathbf{x}^{\text{SA}} &= (x_2^{\text{SA}}, x_2^{\text{SA}}, \ldots, x_{n_1}^{\text{SA}}), n_3 &&= 1,738,467,\\
\mathbf{x}^{\text{DSH}} &= (x_2^{\text{DSH}}, x_2^{\text{DSH}}, \ldots, x_{n_1}^{\text{DSH}}), n_4 &&= 1,068,789,\\
\mathbf{x}^{\text{H}} &= (x_2^{\text{H}}, x_2^{\text{H}}, \ldots, x_{n_1}^{\text{H}}), n_5 &&= 4,113,413,
\end{aligned}$$

donde $\mathbf{x}^{\text{HA}}$ es la muestra aleatoria correspondiente a las observaciones del ambiente disponible con clima hiper árido (HA), $\mathbf{x}^{\text{A}}$ son las observaciones con ambiente árido (A), $\mathbf{x}^{\text{SA}}$ las del ambiente semi árido (SA), $\mathbf{x}^{\text{DSH}}$ las de ambiente seco semi-húmedo(DSH) y $\mathbf{x}^{\text{H}}$ las del ambiente húmedo (H).

Las estimaciones de densidad por kernel para cada ambiente son

$$\begin{aligned}
\widehat{g}_{\text{HA}}(x,y) &= \frac{1}{n_1 h^2}\sum_{i=1}^{n_1} K_2\left(\frac{x-x_i^{\text{HA}}}{h}, \frac{y-y_i^{\text{HA}}}{h}\right),\\
\widehat{g}_{\text{A}}(x,y) &= \frac{1}{n_2 h^2}\sum_{i=1}^{n_2} K_2\left(\frac{x-x_i^{\text{A}}}{h}, \frac{y-y_i^{\text{A}}}{h}\right),\\
\widehat{g}_{\text{SA}}(x,y) &= \frac{1}{n_3 h^2}\sum_{i=1}^{n_3} K_2\left(\frac{x-x_i^{\text{SA}}}{h}, \frac{y-y_i^{\text{SA}}}{h}\right),\\
\widehat{g}_{\text{DSH}}(x,y) &= \frac{1}{n_4 h^2}\sum_{i=1}^{n_4} K_2\left(\frac{x-x_i^{\text{DSH}}}{h}, \frac{y-y_i^{\text{DSH}}}{h}\right),\\
\widehat{g}_{\text{H}}(x,y) &= \frac{1}{n_5 h^2}\sum_{i=1}^{n_5} K_2\left(\frac{x-x_i^{\text{H}}}{h}, \frac{y-y_i^{\text{H}}}{h}\right),
\end{aligned}$$

y al igual que las densidades estimadas por nivel de suculencia, en éstas se usará la notación,

$$\widehat{g}_J, \qquad J \in \{\text{HA, A, SA, DSH, H}\}.$$

Dicho lo anterior, se procede a realizar la estimación de de densidad no paramétrica vía kernel con la librería `ks`. En este caso también se usará un kernel normal bivariado estándar. La Figura 2.5 muestra los perímetros de las regiones de máxima densidad $\text{HDR}_{95}(\widehat{g}_{\text{HA}}), \text{HDR}_{95}(\widehat{g}_{\text{A}}), \text{HDR}_{95}(\widehat{g}_{\text{SA}}), \text{HDR}_{95}(\widehat{g}_{\text{DSH}}), \text{HDR}_{95}(\widehat{g}_{\text{H}})$.

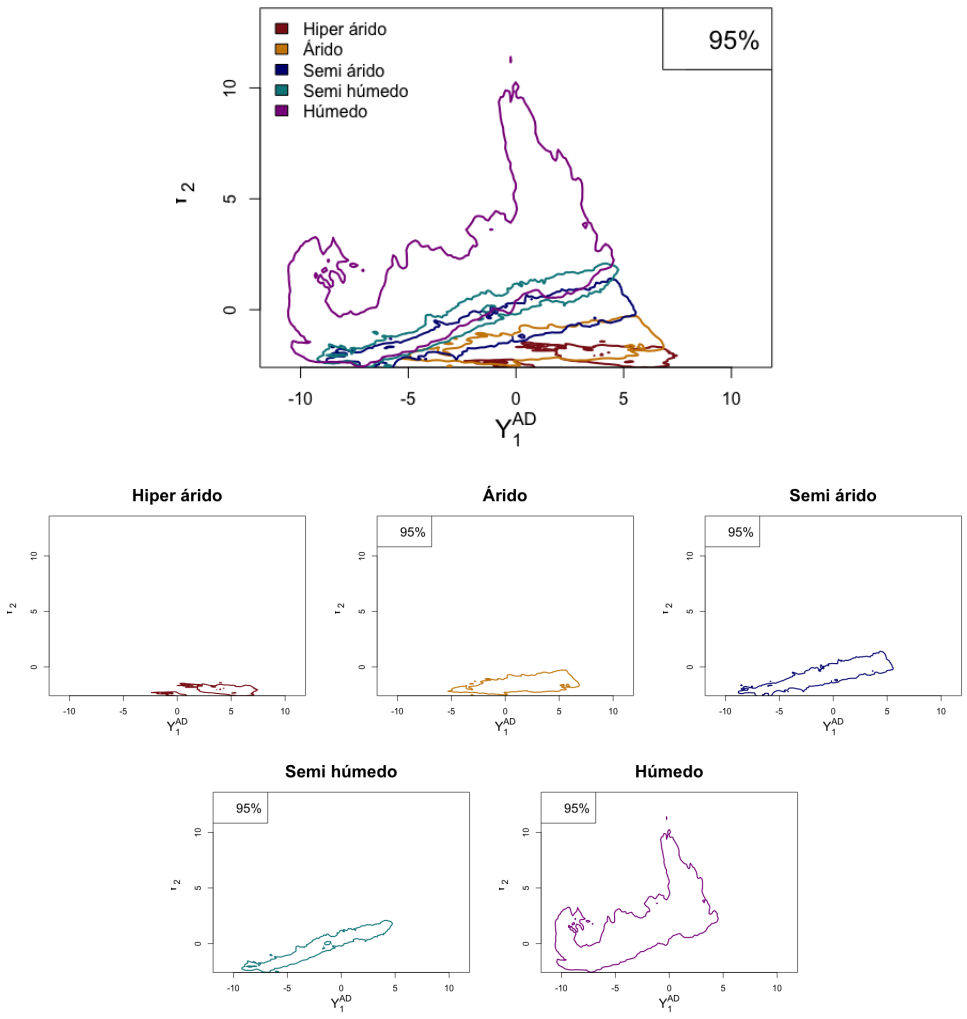


Figura 2.5: HDR de la segmentación del ambiente disponible por nivel de aridez.

Las regiones de máxima densidad basadas en las estimaciones $\widehat{g}_J$ son lineales y conexas. La clasificación estándar de aridez es burda en el sentido de que engloba una gran cantidad de climas en un mismo nivel. Por ejemplo, en el clima húmedo se agrupan climas y no hace distinción entre el clima con Y_1 alto y Y_2 alto que con el clima con Y_1 pequeño y Y_2 pequeño, pero esto último ocurre para todas las clasificaciones de clima. Una mejor clasificación de climas definiría regiones no lineales, y habría una clara zona media que podría ser un disco en el centro de todo el ambiente dis-

ponible, pero escapa a los objetivos de esta tesis hacer clasificación climatológica. Se continuará con el análisis usando esta clasificación, pero se retomará esta observación en la sección de comentarios finales.

2.3. Diagnóstico

En las subsecciones anteriores se propuso una representación probabilista para el nicho climático de las plantas *Caryophyllales*. La modelación se hizo a través de estimaciones no paramétricas de las funciones de densidad de las plantas suculentas, no suculentas y extremadamente suculentas sobre el plano $\mathbb{R}^2$ por medio de WPCA. Ésto permite relacionar la densidad de plantas (en cualquiera de sus niveles de suculencia) con los climas. La Figura 2.4 sugiere que existe un mayor traslape entre las plantas no suculentas y el segmento del plano que corresponde a clima árido, que el traslape que hay entre las plantas extremadamente suculentas y el segmento que corresponde al clima árido. Medir la zona árida con un rectángulo es algo burdo y usado sólo para fines ilustrativos, pero es necesario para un análisis formal una referencia oficial de aridez como la provista por CGIAR-CSI.

La clasificación de aridez de CGIAR-CSI es la más aceptada actualmente. En la Subsección 2.2.2 se modeló, usando estimación de densidad no paramétrica, la densidad de los niveles de aridez en el plano $\mathbb{R}^2$.

Como consecuencia de lo anterior, se tienen suficientes elementos para comparar la densidad de suculencia con aridez. Por un lado, se tienen la función de densidad estimada de las plantas suculentas, extremadamente suculentas y no suculentas. Por otra parte, se tiene la densidad del ambiente árido, hiper árido, seco sub-húmedo, húmedo y semi árido sobre el mismo soporte que las densidades de suculencia. De manera natural se piensa en comparar la magnitud del traslape entre suculencia y aridez.

En algún punto se puede pensar que es conveniente comparar simplemente el

nicho climático de las plantas suculentas con el de las plantas no suculentas por medio de la comparación de las regiones de máxima densidad. Si bien es cierto que la comparación de los nichos climáticos podría darnos buena idea de la riqueza del nicho de las suculentas con relación a las no suculentas, el cuestionamiento principal hace referencia no sólo a la amplitud del nicho climático sino a la interacción de éste con los distintos niveles de aridez. Además, no sería informativo medir el traslape entre $\text{HDR}_{95}(\widehat{f}_{\text{XS}})$ o $\text{HDR}_{95}(\widehat{f}_{\text{S}})$ con $\text{HDR}_{95}(\widehat{f}_{\text{NS}})$ porque al estar superpuestos el traslape sería siempre total, lo cuál no es muy informativo. Luego, se considerará otra forma de analizar el nicho climático.

Medir el nivel de traslape entre un nivel de suculencia y uno de aridez se reduce a calcular el traslape entre dos densidades; por ejemplo, para conocer el nivel de traslape entre las plantas extremadamente suculentas y ambiente hiper árido, bastará con medir el traslape entre la función de densidad de las plantas extremadamente suculentas y la función de densidad del ambiente hiper árido. Lo anterior da paso a la comparación del traslape con el ambiente árido de las plantas extremadamente suculentas, suculentas y las no suculentas. Si es verdad que las plantas suculentas tienen un nicho climático exclusivamente árido, entonces las densidades de las plantas suculentas y extremadamente suculentas traslaparían con las densidades de ambientes de mayor aridez en mayor medida que la densidad de plantas no suculentas. Con lo anterior se diagnostica que una forma natural y de índole estadística para medir relación entre suculencia y aridez es cuantificar el traslape de densidades. En la Figura 2.6 se ilustran los cinco niveles de aridez superpuestos al nicho climático de las plantas separado por nivel de aridez.

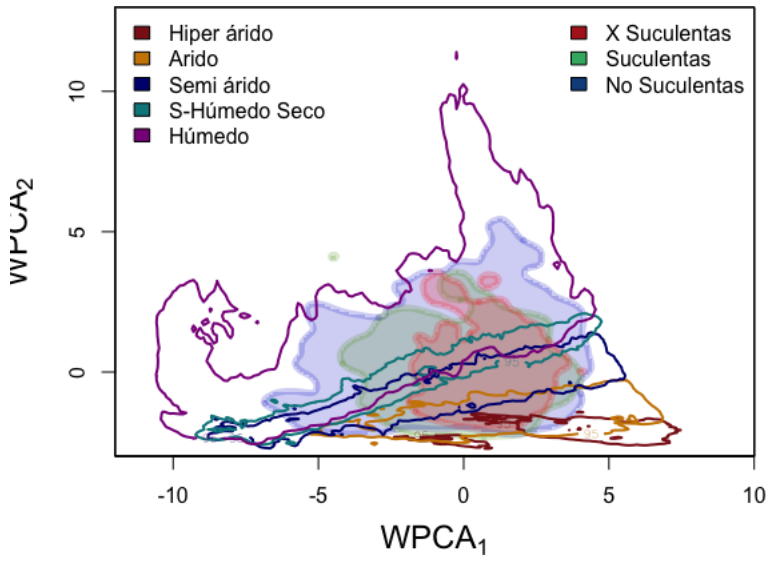


Figura 2.6: Traslape entre niveles de aridez y niveles de suculencia.

CAPÍTULO 3

Inferencia estadística

El diagnóstico apunta a la cuantificación del traslape entre suculencia de las plantas *Caryophyllales* y los niveles de aridez ambiente disponible, en este capítulo se efectúa esta labor. En la primera sección se define el índice de traslape de Weitzman y se da una estimación no paramétrica de él. El índice de Weitzman se usa para medir el traslape entre f_I y g_J, no obstante, se debe de tomar en cuenta que al no conocer la distribución real de los datos es necesario recurrir a las estimaciones $\widehat{f_I}$ y $\widehat{g}_J$. Así mismo, el índice de traslape será estimado mediante un estimador no paramétrico, el cuál es definido en la Subsección 3.1. Estimar el índice de traslape da lugar a la construcción de intervalos de confianza, los cuáles serán estimados mediante el método bootstrap. El resultado de este capítulo es la estimación de intervalos de confianza para el índice de traslape, y son la herramienta estadística sustancial que respalda las conclusiones ecológicas de la tesis, las cuáles se darán en el siguiente y último capítulo.

3.1. Índice de traslape de Weitzman

En la literatura hay diferentes propuestas para cuantificar el traslape entre densidades. Una de las más utilizadas es el índice de traslape de Weitzman (Ridout and Linkie, 2009), el cuál se define a continuación para dos dimensiones.

Definición 2. *Sean dos funciones funciones de densidad continuas* $f : \mathbb{R}^2 \to \mathbb{R}$, $g : \mathbb{R}^2 \to \mathbb{R}$. *El índice de traslape de Weitzman es:*

$$\Delta(f, g) = \int_{\mathbb{R}} \int_{\mathbb{R}} \text{mín}\{f(x, y), g(x, y)\}\, dxdy. \tag{3.1}$$

Este índice considera varianza, superposición y al mismo tiempo comparación de soportes. El índice de Weitzman es una medida confiable y rigurosa para medir traslape entre funciones de densidad. En el contexto de relación entre suculencia y aridez, cuantificar la superposición permitirá la exploración de la convivencia entre plantas suculentas y no suculentas y del nexo de aridez con suculencia a través de la comparación de funciones de densidad.

Otra particularidad de $\Delta(f, g)$, es que cuando $f = g$ el traslape es total y al ser funciones de densidad $\Delta(f, g) = 1$. En efecto,

$$\begin{aligned}
\Delta(f, g) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \text{mín}\{f(x, y), g(x, y)\}\, dxdy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \text{mín}\{g(x, y), g(x, y)\}\, dxdy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} g(x, y) dxdy \\
&= 1,
\end{aligned}$$

y el soporte de f es totalmente ajeno al de g, sop$(f)\cap$sop$(g) = \emptyset$, entonces $\Delta(f, g) = 0$, ya que

$$\begin{aligned}\Delta(f,g) &= \int_{\mathbb{R}}\int_{\mathbb{R}} \text{mín}\left\{f(x,y), g(x,y)\right\} dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} \text{mín}\left\{0\right\} dxdy \\ &= 0.\end{aligned}$$

Por consiguiente Δ es acotado: $0 \leq \Delta \leq 1$. Es decir, mientras más cerca esté Δ de uno, más grande será el traslape entre ambas densidades y mientras más cerca esté de cero, las densidades traslapan en menor medida.

Si las plantas suculentas tienen mecanismos de adaptación a climas áridos que las plantas suculentas no poseen, el índice de traslape de éstas sería mayor que el de las no suculentas en regiones áridas. El traslape entre suculentas y regiones áridas es medido mediante $\Delta(f_{\text{XS}}, g_{\text{HA}}), \Delta(f_{\text{XS}}, g_{\text{A}}), \Delta(f_{\text{XS}}, g_{\text{SA}})$ y $\Delta(f_{\text{S}}, g_{\text{HA}}), \Delta(f_{\text{S}}, g_{\text{A}}), \Delta(f_{\text{S}}, g_{\text{SA}})$ también pueden aportar información valiosa, ya que aunque no se consideren extremadamente suculentas, las plantas simplemente suculentas también han desarrollado el síndrome de suculencia. Más aún, si las plantas suculentas son hegemónicas de regiones áridas, a menor nivel de aridez el índice de traslape debería de ser menor, es decir,

$$\Delta(f_{\text{XS}}, g_{\text{HA}}) \geq \Delta(f_{\text{XS}}, g_{\text{A}}) \geq \Delta(f_{\text{XS}}, g_{\text{SA}}) \geq \Delta(f_{\text{XS}}, g_{\text{DSH}}) \geq \Delta(f_{\text{XS}}, g_{\text{H}}).$$

Resumiendo, el índice de traslape es útil para la exploración y comparación de nichos ecológicos con niveles de aridez si se considera el traslape entre las funciones de densidad f_I y g_J.

Las funciones de densidad f_I y g_J no se conocen explícitamente, pero se cuenta con las estimaciones obtenidas en las Subsecciones 2.1.2 y 2.2.2. Aunque el interés primordial de la investigación radica en las regiones áridas, para obtener una descripción más completa se estimará el índice de traslape entre cada combinación posible de densidad de suculencia $\widehat{f_I}$ y de aridez $\widehat{g}_J$ mediante la ecuación

$$\Delta(\widehat{f}_I, \widehat{g}_J) \quad = \quad \int_{\mathbb{R}} \int_{\mathbb{R}} \text{mín} \left\{ \widehat{f}_I(x,y), \widehat{g}_J(x,y) \right\} dxdy, \tag{3.2}$$

la cuál es una estimación *Plug-in* basada en la Definición 1. En la Subsubsección 3.1 se referenciará un método para aproximar $\Delta(\widehat{f}_I, \widehat{g}_J)$ utilizando integración numérica.

El cálculo teórico de Δ puede ser complejo, pues no se cuenta con la forma paramétrica de f_I ni de f_J. En la siguiente subsección se aborda una alternativa, que consiste en estimar Δ con una técnica no paramétrica.

Estimación del índice de traslape

Algunos estimadores de Δ son abordados en Schmid and Schmidt (2006), mientras que Ridout and Linkie (2009) hace un estudio comparativo de ellos. Los estimadores fueron comparados en términos de varianza y precisión. Por la parte de los estimadores no paramétricos, el desempeño fue medido en función del tamaño de muestra. En concreto, se usará el estimador recomendado para tamaños de muestra mayores a setenta y cinco. Los detalles de la deducción del estimador pueden ser consultados en Schmid and Schmidt (2006).

Sea$(x_1, \ldots, x_n), \quad x_i \in \mathbb{R}^2$ una muestra aleatoria de Y^{P}, y $(y_1, \ldots, y_m), \quad y_i \in \mathbb{R}^2$ una muestra aleatoria de Y^{AD}. Un estimador no paramétrico de $\Delta(\widehat{f}, \widehat{g})$ es

$$\widehat{\Delta}(\widehat{f}, \widehat{g}) \quad = \quad \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^{n} \text{mín} \left\{ 1, \frac{\widehat{g}(x_i)}{\widehat{f}(x_i)} \right\} + \frac{1}{m} \sum_{j=1}^{m} \text{mín} \left\{ 1, \frac{\widehat{f}(y_j)}{\widehat{g}(y_j)} \right\} \right). \tag{3.3}$$

Los estimadores de interés se obtienen de combinar tres niveles de suculencia con cinco de aridez y son las siguientes:

$$
\begin{aligned}
\widehat{\Delta}_1 &= \widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{HA}}), & \widehat{\Delta}_6 &= \widehat{\Delta}(\widehat{f}_{\text{S}}, \widehat{g}_{\text{HA}}), & \widehat{\Delta}_{11} &= \widehat{\Delta}(\widehat{f}_{\text{NS}}, \widehat{g}_{\text{HA}}),\\
\widehat{\Delta}_2 &= \widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{A}}), & \widehat{\Delta}_7 &= \widehat{\Delta}(\widehat{f}_{\text{S}}, \widehat{g}_{\text{A}}), & \widehat{\Delta}_{12} &= \widehat{\Delta}(\widehat{f}_{\text{NS}}, \widehat{g}_{\text{A}}),\\
\widehat{\Delta}_3 &= \widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{SA}}), & \widehat{\Delta}_8 &= \widehat{\Delta}(\widehat{f}_{\text{S}}, \widehat{g}_{\text{SA}}), & \widehat{\Delta}_{13} &= \widehat{\Delta}(\widehat{f}_{\text{NS}}, \widehat{g}_{\text{SA}}),\\
\widehat{\Delta}_4 &= \widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{DSH}}), & \widehat{\Delta}_9 &= \widehat{\Delta}(\widehat{f}_{\text{S}}, \widehat{g}_{\text{DSH}}), & \widehat{\Delta}_{14} &= \widehat{\Delta}(\widehat{f}_{\text{NS}}, \widehat{g}_{\text{DSH}}),\\
\widehat{\Delta}_5 &= \widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{H}}), & \widehat{\Delta}_{10} &= \widehat{\Delta}(\widehat{f}_{\text{S}}, \widehat{g}_{\text{H}}), & \widehat{\Delta}_{15} &= \widehat{\Delta}(\widehat{f}_{\text{NS}}, \widehat{g}_{\text{H}}).
\end{aligned}
$$

Estos quince índices se abreviarán de la siguiente manera:

$$\widehat{\Delta}_\iota, \quad \iota \in \{1, \ldots, 15\}.$$

Nótese que $\widehat{\Delta}_1, \ldots, \widehat{\Delta}_5$ están asociados con plantas extremadamente suculentas, $\widehat{\Delta}_6$, $\ldots, \widehat{\Delta}_{10}$ con plantas suculentas y $\widehat{\Delta}_{11}, \ldots, \widehat{\Delta}_{15}$ con plantas no suculentas.

Estimar todas las combinaciones entre suculencia y nivel de aridez permitirá comparar el nicho climático y ambiente disponible a través de un número entre cero y uno. De acuerdo a la interpretación de Δ, si el índice de traslape correspondiente a plantas no suculentas y ambiente hiper árido es mayor que el que corresponde a plantas suculentas y ambiente hiper árido, será evidencia suficiente para contradecir la creencia popular. Lo anterior es expresado como:

$$\widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{HA}}) < \widehat{\Delta}(\widehat{f}_{\text{NS}}, \widehat{g}_{\text{HA}}).$$

Los estimadores tienen asociada una función de distribución, a partir de la cuál se pueden construir intervalos de confianza para el parámetro de interés. La distribución de los estadísticos no siempre es inmediata, en este caso la distribución de 3.1 no es clara; no obstante, ésta se puede aproximar con técnicas estadísticas. En la siguiente sección se estima la distribución del índice de traslape mediante el método bootstrap, logrando con ello construir intervalos de confianza para Δ.

3.2. Intervalos de confianza para la estimación del índice de traslape

El método boostrap es útil para encontrar la distribución de un estimador. En este caso, para estimar intervalos de confianza para Δ es necesario contar con la función de distribución de 3.1, la cuál no es clara. Se cuenta con tres muestras: $\mathbf{x}^{\text{S}}$, $\mathbf{x}^{\text{XS}}$ y $\mathbf{x}^{\text{NS}}$. Sobre dichas muestras se realizará bootstrap no paramétrico (Davison and Hinkley, 1997) para conocer la distribución de los estimadores $\widehat{\Delta}_\iota$. Posteriormente, se usará esta aproximación para construir intervalos de confianza y de esta manera comparar los nichos climáticos de las plantas en función de la aridez.

El clima en el mundo no cambia a la escala de tiempo fijo que se está considerando. En cambio, las estimaciones de las funciones de densidad asociadas al ambiente $\widehat{g}_J$ fueron construidas con base en las observaciones $\mathbf{x}^{\text{AD}}$, que deben su aleatoriedad factores como precisión de instrumentos de medición, errores de captura, entre otros. En cambio, las densidades estimadas de las plantas, $\widehat{f}_I$ fueron construidas con la muestra observada $\mathbf{x}^{\text{P}}$ que debe su aleatoriedad a los avistamientos de plantas. El método bootstrap será aplicado considerando a f_I como componente aleatorio y fijando g_J. De esta manera la incertidumbre recae en los avistamientos de plantas. En consecuencia, sólo se tienen que realizar tres simulaciones bootstrap, una para cada nivel de suculencia, para estimar las quince densidades de los índices de traslape $\widehat{\Delta}_\iota$.

Bootstrap es un método asintótico, que requiere un número de repeticiones *grande* para alcanzar la convergencia. se realizaron $N = 2,500$ remuestreos usando bootstrap no paramétrico. El método bootstrap conlleva remuestrear de $\mathbf{x}^{\text{XS}}, \mathbf{x}^{\text{S}}$ y $\mathbf{x}^{\text{NS}}$ de manera independiente y sin remplazo. El i-ésimo remuestreo se denotará por

$$\begin{aligned}
\mathbf{x}^{\text{XS},(k)} &= (x_1^{\text{XS},(k)}, \ldots, x_{2,028}^{\text{XS},(k)}), \\
\mathbf{x}^{\text{S},(k)} &= (x_1^{\text{S},(k)}, \ldots, x_{224}^{\text{S},(k)}), \\
\mathbf{x}^{\text{NS},(k)} &= (x_1^{\text{NS},(k)}, \ldots, x_{2,811}^{\text{NS},(k)}),
\end{aligned}$$

donde $i \in \{1, \ldots, N\}$ y $x_i^{I,(k)} \in \mathbb{R}^2$, $x_i^{I,(k)} = (x_{(1,i)}^{I,(k)}, x_{(2,i)}^{I,(k)})$.

Cada remuestreo da lugar a nuevas estimaciones de f_I, que se estimarán nuevamente usando estimación de densidad por kernel no paramétrica. Nuevamente se usa un kernel gaussiano con entradas independientes. La estimación es

$$\widehat{f}_I^{(k)}(x, y) = \frac{1}{n_I h^2} \sum_{i=1}^{n_I} K\left(\frac{x - x_{1,i}^{I,(k)}}{h}, \frac{y - x_{2,i}^{I,(k)}}{h}\right).$$

Luego, el estimador del índice de traslape de la k-ésima simulación bootstrap entre nivel de suculencia I y nivel de aridez J es

$$\widehat{\Delta}^{(k)}(\widehat{f}_I^{(k)}, \widehat{g}_J) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \text{mín}\left\{1, \frac{\widehat{g}_J(x_i^{I,(k)})}{\widehat{f}_I^{(k)}(x_i^{I,(k)})}\right\} + \frac{1}{m}\sum_{j=1}^{m} \text{mín}\left\{1, \frac{\widehat{f}_I^{(k)}(y_j^J)}{\widehat{g}_J(y_j^J)}\right\}\right).$$

Por ejemplo, la simulación del índice de traslape del ambiente hiper árido y plantas extremadamente suculentas es

$$\begin{aligned}
\widehat{\Delta}_1^{*,1} &= \widehat{\Delta}^{(1)}(\widehat{f}_{\text{XS}}^{(1)}, \widehat{g}_{\text{HA}}) \\
\widehat{\Delta}_1^{*,2} &= \widehat{\Delta}^{(2)}(\widehat{f}_{\text{XS}}^{(2)}, \widehat{g}_{\text{HA}}) \\
&\vdots \\
\widehat{\Delta}_1^{*,\text{N}} &= \widehat{\Delta}^{(N)}(\widehat{f}_{\text{XS}}^{(N)}, \widehat{g}_{\text{HA}})
\end{aligned}$$

y el índice de traslape observado es simplemente $\widehat{\Delta}_1 = \widehat{\Delta}(\widehat{f}_{\text{XS}}, \widehat{g}_{\text{HA}})$.

Los histogramas de $\widehat{\Delta}_\iota$ ilustrados en la Figura 3.1 son acampanados, es decir, no se encuentran concentrados en un solo valor, son simétricos y su varianza no es relativamente pequeña.

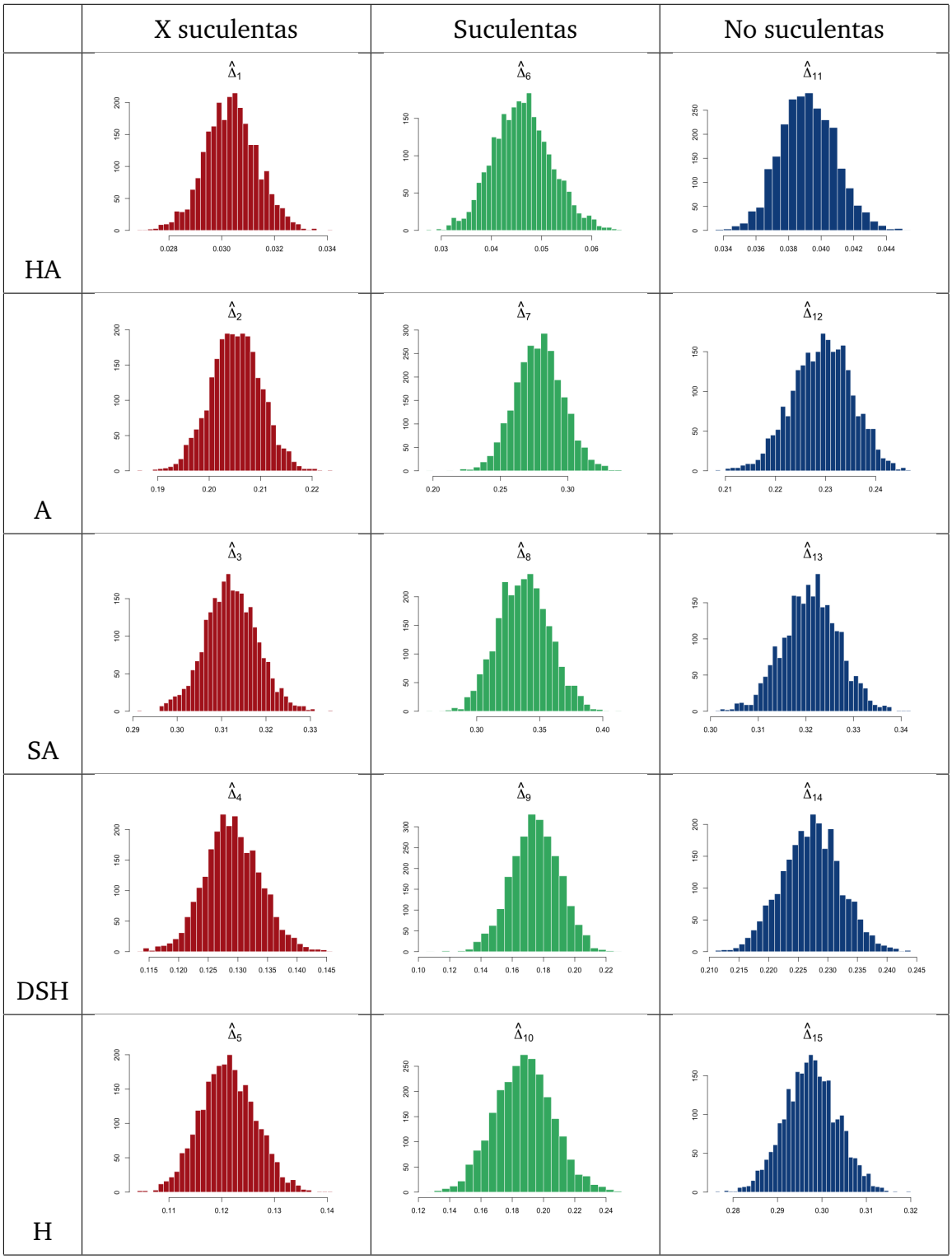


Figura 3.1: Histogramas de $\hat{\Delta}_\iota$.

En la teoría de bootstrap hay varios métodos para la construcción de intervalos de confianza. Davison and Hinkley (1997) propone el método básico para estimación de éstos, el cuál hace uso del cuantil empírico. El cuantil empírico $\widehat{Q}^*_{\iota,\alpha}$ de nivel $\alpha \in [0,1]$ de la distribución estimada del índice de traslape $\widehat{\Delta}_\iota$ es

$$Q^*_\alpha = \text{ínf}\left\{\hat{\Delta}^{*,i} \mid F^*_N(\widehat{\Delta}^{*,i}) \geq \alpha\right\}.$$

Definición 3. *El intervalo de confianza de nivel* $\alpha \in [0,1]$ *para* Δ *es*

$$(2\widehat{\Delta}_\iota - \widehat{Q}^*_{\iota,1-\alpha}, 2\widehat{\Delta}_\iota - \widehat{Q}^*_{\iota,\alpha}), \quad \iota \in \{1,\ldots,15\},$$

donde $F^*_N(x)$ *es la función de distribución empírica de la muestra*

$$\widehat{\boldsymbol{\Delta}}^*_\iota = \left(\widehat{\Delta}^{*,1}_\iota, \widehat{\Delta}^{*,2}_\iota, \ldots, \widehat{\Delta}^{*,N}_\iota\right).$$

Los intervalos de 95 % de confianza[1] para $\Delta_1,\ldots,\Delta_{15}$ se muestran en la Tabla 3.1, y serán dotados de un significado ecológico en el Capítulo 4.

	X Suculentas		Suculentas		No suculentas	
HA	Δ_1	$(0.0290, 0.0330)$	Δ_6	$(0.0362, 0.0591)$	Δ_{11}	$(0.0365, 0.0432)$
A	Δ_2	$(0.1985, 0.2176)$	Δ_7	$(0.2577, 0.3275)$	Δ_{12}	$(0.2196, 0.2422)$
SA	Δ_3	$(0.3073, 0.3308)$	Δ_8	$(0.3147, 0.3943)$	Δ_{13}	$(0.3126, 0.3354)$
DSH	Δ_4	$(0.1236, 0.1421)$	Δ_9	$(0.1602, 0.2208)$	Δ_{14}	$(0.2195, 0.2393)$
H	Δ_5	$(0.1157, 0.1358)$	Δ_{10}	$(0.1691, 0.2424)$	Δ_{15}	$(0.2895, 0.3129)$

Cuadro 3.1: Intervalos de 95 % de confianza para Δ_ι por método básico.

La Figura 3.2 se usa como una herramienta para la visualización de los intervalos mostrados en la Tabla 3.1. De esta forma es fácil tener una perspectiva general de cómo se comporta Δ de cada nivel de suculencia, respecto a los cinco niveles de aridez.

[1] En lo siguiente, cuando se hable de intervalos de confianza para Δ se estará haciendo referencia al intervalo básico de $95\,\%$ de confianza.

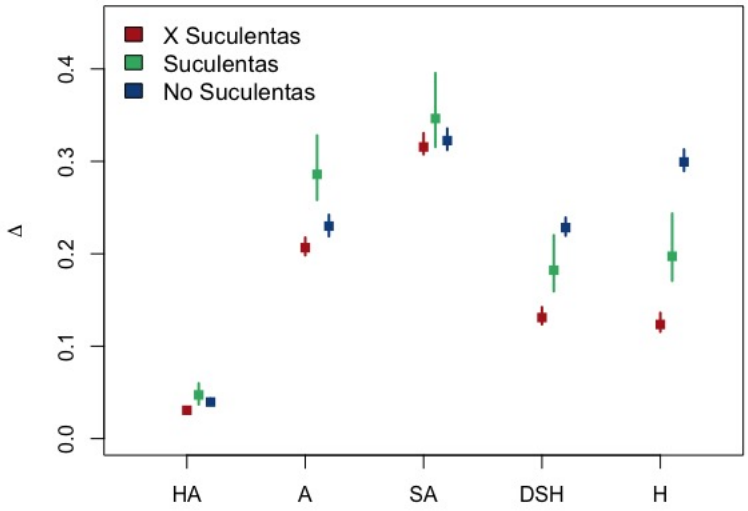


Figura 3.2: Intervalos de confianza de 95 % de confianza por el método básico.

Los intervalos de confianza para el traslape son un resumen de la ocupación del nicho ecológico para cada grupo de plantas en cada tipo de hábitat, por lo que si el traslape de f_I con g_J es cercano a uno, esto sugerirá que el nicho ecológico del grupo de plantas I ocupa en gran medida hábitat J; dicho de otro modo, que hay una gran variedad de especies que habitan éste ambiente. Por lo tanto, que han desarrollado mecanismos que les permiten vivir con ese entorno. En el caso contrario, si el índice de traslape entre f_I y g_J es cercano a cero significa que no han habido especies del grupo I que hayan sido capaces de adaptarse al clima J. Para la interpretación de los intervalos de confianza, se analizará el empalme entre ellos y el típo de hábitat donde ésto sucede. Además, se hablará de la ubicación en la recta $[0, 1]$ de los intervalos y lo que esto significa en términos de ocupación de hábitat.

CAPÍTULO 4

Discusión y conclusiones

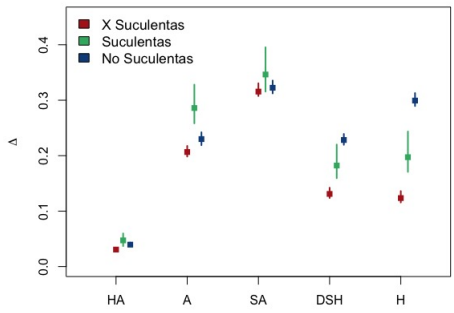
En esta sección se discuten los resultados obtenidos. En el Capítulo 3 efectuó el diagnóstico, y por ello la mayor parte de los resultados están consolidados en esta sección de la tesis. El énfasis aquí será el razonamiento de los intervalos de confianza en términos de ecología. También se darán algunos comentarios finales que fueron recabados a lo largo de la realización del estudio. Los comentarios finales van en sentido de tareas que es posible realizar para mejorar o ampliar las conclusiones. Con estos dos elementos se concluye la tesis.

Las conclusiones aquí serán parafraseadas en términos de intervalos de confianza, los cuales han sido los instrumentos técnicos dispuestos en capítulos anteriores. Sin embargo, como nota aclaratoria se mencionará que en efecto hay hipótesis implícitamente probadas cuando se comparan entre sí dos intervalos para un parámetro, correspondientes a dos poblaciones distintas. El concepto que establece esta equivalencia se conoce como inversión de pruebas (Casella and Berger, 2002), y significa que la hipótesis nula de igualdad entre parámetros se rechaza al nivel α si y sólo si

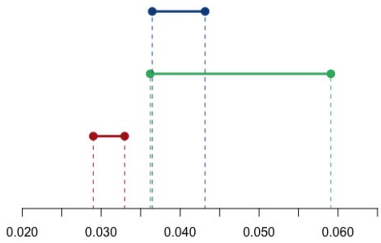
los dos intervalos de confianza $(1-\alpha)\times 100\,\%$ no se intersectan. Por ello en lo que sigue, será utilizado indistintamente un discurso afín a pruebas de hipótesis que versan sobre comparación entre poblaciones.

4.1. Conclusiones ecológicas

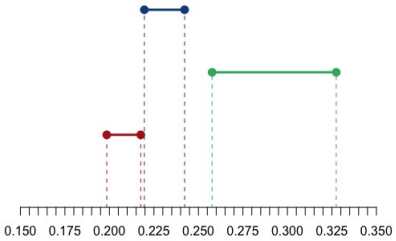
En la Figura 4.1 podemos observar que en los tres climas áridos (Hiper árido, Árido y Semi árido) el intervalo de confianza correspondiente a las plantas extremadamente suculentas toma valores más pequeños que el de las plantas no suculentas (está corrido a la izquierda); esto significa que el nicho climático de las plantas extremadamente suculentas está más limitado, mientras que el de las no suculentas es más versátil y cuenta con mayor diversidad de climas. Por otro lado, en climas húmedos (Seco semi-húmedo y húmedo) los intervalos no se traslapan; es más, en la figura (f) los intervalos son totalmente ajenos y están muy separados. Esto significa que en humedad, cada grupo de plantas se adjudica un comportamiento totalmente diferente. En este caso son las plantas no suculentas las que el intervalo de confianza sugiere mayor sobreposición. Se puede advertir que a mayor humedad los intervalos de confianza tienden a separarse más.



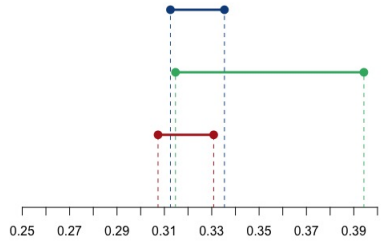
(a) Intervalos 95 % de confianza por método básico



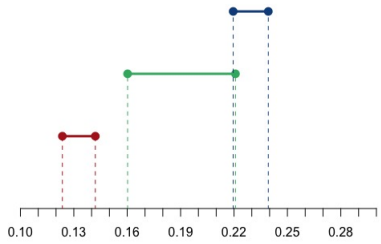
(b) Clima Hiper árido



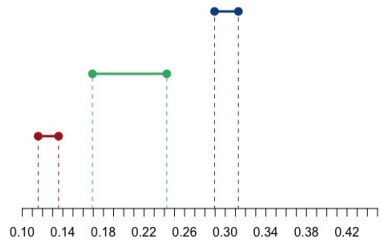
(c) Clima Árido



(d) Clima Semi árido



(e) Clima Seco semi-húmedo



(f) Clima Húmedo

Figura 4.1: Intervalos de confianza por bootstrap no paramétrico para Δ.

Primero, nótese en el clima hiper árido (ver Figura 4.1 (b)) los intervalos de confianza correspondientes al grupo de plantas suculentas y no suculentas se sobreponen, lo cuál significa que no se descarta que el índice de traslape de los estos dos tipos de plantas sea el mismo. Es decir, las plantas suculentas no son exclusivas de esta región sino que existen otras especies de plantas que habitan en este clima en la misma medida. También, nótese que el intervalo de las plantas extremadamente suculentas toma valores muy pequeños (menores a 0.1), lo cuál evidencia que estas plantas en realidad no habitan en regiones hiper áridas. Lo hacen más las plantas suculentas y las no suculentas. Ésto contradice la presunción que se tiene sobre que las plantas suculentas habitan las zonas más áridas, ya que (de acuerdo a los intervalos de confianza del clima hiper árido) las especies de plantas no suculentas habitan más éste clima.

Nótese que los tres intervalos de confianza están contenidos en el intervalo (0, 0.06). Es decir, pareciera que las condiciones en este entorno son tan extremas que ninguna especie (ni suculentas ni no suculentas) ha logrado habituarse para sobrevivir.

En el ambiente árido (Figura (c)) los intervalos de confianza ya no se traslapan, lo que significa que los grupos tienen comportamientos diferentes. Nuevamente el intervalo correspondiente a plantas extremadamente suculentas es el que toma valores más chicos, significando que éstas plantas tienen menor capacidad de adaptación en este clima.

En el clima semi árido es el único donde los tres intervalos se traslapan (cada uno con los otros dos). En otras palabras, no se puede concluir que las especies de cada grupo han desarrollado mecanismos de adaptación diferentes que las lleven a tener una dinámica de supervivencia distinta en el clima semi árido.

En el clima húmedo los intervalos no se sobreponen. La humedad puede ser una característica que ayuda a discernir las plantas suculentas de las no suculentas. Es un hecho verificado entonces, que las plantas suculentas sucumben antes grandes

cantidades de humedad.

Se observa que las plantas extremadamente suculentas tienen un nicho muy concentrado, ya que el intervalo del índice de traslape es mayor en clima semi árido y luego en los demás climas es cercano a cero. Esto significa que tienen un nicho concentrado y fuera del cima semi árido les cuesta trabajo sobrevivir.

Por su parte, las plantas no suculentas tienen índices de traslape altos en todos los climas, con excepción del ambiente hiper árido. Como se percibía también en los contornos de máxima densidad, las plantas no suculentas poseen un nicho climático amplio, esto es, llegan a poblar una amplia gama de climas. Finalmente, las plantas suculentas parecen atesorar un comportamiento diferente, al igual que las otras dos, en el ambiente húmedo.

4.2. Comentarios finales

Durante el desarrollo del trabajo se detectaron mejoras que pueden ser incluidas en este trabajo. Estas mejoras incluyen: enriquecimiento de covariables, clasificación de climas disponibles en el mundo, simulación bootstrap semi-paramétrico y estimación de regiones de confianza en lugar de intervalos. A continuación se detallarán cada uno de los puntos.

Enriquecimiento de covariables

Enriquecimiento de covariables se refiere a ampliar el vector aleatorio

$$X = (\text{Bio\_1}, \text{Bio\_5}, \text{Bio\_6}, \text{Bio\_12}, \text{Bio\_15}, \text{Bio\_16}, \text{Bio\_17}, \text{Pet\_he\_yr}),$$

agregando variables que pueden ayudar a describir mejor el hábitat de las plantas suculentas. Por ejemplo, algunas variables que pueden ser incluidas en el análisis son:

a) Índice topográfico (Compound Topographic Index, o CTI), el cual describe la propensión de un lugar para acumular agua precipitada,

b) el aspecto, que indica el punto cardinal hacia el que se orienta la superficie tangencial del terreno, y

c) la pendiente del terreno medida en grados.

Estas variables se encuentran disponibles en repositorios tales como usgs.gov (US. Geological Survey), aunque posiblemente en distinta resolución y distinto sistema de proyección del que es utilizado para datos climáticos (Bioclim). Lo anterior sugiere que sería necesaria una cuidadosa labor para integrar estas variables.

También, se puede analizar el efecto de incluir el resto de las variables biológicas que no están siendo incluidas, como Bio_4 que es estacionalidad de la temperatura. El estudio de variables adicionales pueden llevar a conclusiones de interés ecológico. El método de reducción de dimensionalidad para caracterizar el nicho climático propuesto en el Capítulo 1 traza el camino para la discriminación e interpretaciones de variables nuevas en el estudio.

Clasificación de climas con sentido biológico

La clasificación del índice de aridez basada en umbrales puede no representar de manera totalmente útil la variedad de climas en el mundo, por lo menos para fines de botánica. Una mejora que puede derivarse como consecuencia secundaria de este trabajo es es construir una clasificación climatológica que sea más realista sobre los niveles de aridez. Es decir, que diferencie los climas extremos con mayor grado de detalle. Una opción sería encontrar una clasificación que en lugar de responder a variables meteorológicas, represente características *biológicas*. Esto es, que el grado y tipo de suculencia sea un indicador biológico más flexible y preciso de aridez.

El instrumento interactivo mencionado en el Capítulo 4 ha sido efectivo para detectar lo siguiente, lo cual es una primera ilustración de una clasificación indirecta

de aridez basada en parámetros biológicos más que climáticos. Se ha descubierto que entre las familias más numerosamente representadas en la zona 5 (ver Figura 4.2) se encuentran *Caryophyllaceae*, *Amaranthaceae* y *Montiaceae*, mientras que en la zona 7 se encuentran Cactaceae, Amaranthaceae y Nyctaginaceae. En la clasificación usual, tanto la zona 5 como la 7 se encuentran indistintamente clasificadas como húmedas. Sin embargo, en términos biológicos las familias que predominan en cada una de esas zonas tienen composiciones muy diferentes. Por ello, la zona denominada *húmeda* realmente tiene subdivisiones más finas que las plantas mismas están delineando.

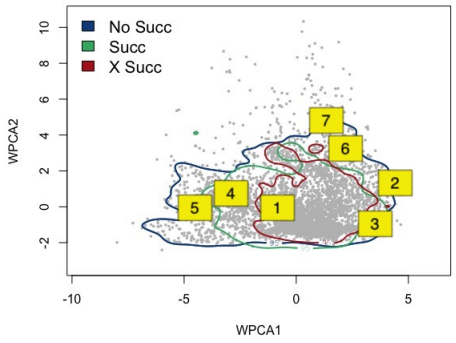


Figura 4.2: Zonas climáticas de acuerdo a población de plantas. Los puntos grises representan las especies, mientras que los contornos son los correspondientes a la HDR de 95 % de probabilidad.

Simulación con bootstrap semi-paramétrico

El método por el cual fueron construidos los intervalos de confianza en la Subsubsección 3 fue basada en bootstrap no paramétrico el cual se basa en la premisa de aproximar la distribución empírica F_n. Esto produce el inconveniente de que F_n es discreta, no obstante el plano (WPCA1, WPCA2) alberga valores en un continuo. Surge la inquietud de si un estimador no discreto de F proporcionaría mejores resultados (ver Davison and Hinkley (1997)). En la Subsubsección 2.2.1 se realizó el ajuste de densidad por kernel para las plantas, y se llegó a que la estimación de f_I se puede

ver como una mezcla de kernel gaussianos K_i, con matriz de covarianzas $\Sigma = h\mathrm{I}_2$, y media $\mu = \left[x_i^I, y_i^I\right]^T$

$$\widehat{f}_I(x,y) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} K_i\left(x,y\right), \tag{4.1}$$

de modo que en lugar de simular de F_n se puede simular valores de la mezcla formulada en la Ecuación 4.1, y así suprimir el problema de discretización. La librería `ks` cuenta con un generador de números aleatorios para objetos de la clase `kde`, y en este sentido resulta útil para la simulación de la mezcla.

Este trabajo pretende contribuir progreso del esclarecimiento de la dinámica de adaptación de las plantas suculentas, mediante el análisis del grupo *Caryophyllales*. La adecuación de las plantas suculentas a climas áridos ha sido poco explorada, y en este sentido el producto final de esta tesis (conclusiones, base complementaria, comentarios finales, *etc*.) representan un precedente para futuros trabajos de ésta línea de investigación. El trabajo futuro sugerido en ésta última sección (enriquecimiento de covariables, clasificación de climas con sentido biológico y simulación bootstrap semiparamétrica) sugiere tres nuevos estudios, que para su realización requieren el ejercicio de habilidades estadísticas y computacionales. Además, invocan por si mismas la combinación de las áreas estadística y botánica.

Referencias

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury Advanced Series, 2nd edition.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge Series on Statistical and Probabilistic Mathematics.

Duong, T., Wand, M., Chacon, J., and Gramacki, J. (2019). *ks: Kernel Smoothing*. Version 1.11.5. Available online: www.mvstat.net/mvksa (accessed on 10 July 2019).

Gibson, A. (2012). *Structure-Function Relations of Warm Desert Plants*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer Series in Statistics, 2nd edition.

Hijmans, R. (2019). *raster: Geographic Data Analysis and Modeling* . Version 2.9–5. Available online: www.rspatial.org (accessed on 10 July 2019).

Hyndman, R. (1996). Computing and Graphing Highest Density Regions. *The American Statistician*, 50(2):120–126.

Kroonenberg, P. (2008). *Applied Multiway Data Analysis*. Wiley.

Noy-Meir, I. (1973). Desert ecosystems: Environment and Producers. *Annual Review of Ecology and Systematics*, 4(1):25–51.

O´Donnell, M. and Ignizio, D. (2012). Bioclimatic Predictors for Supporting Ecological Applications in the Conterminous United States. *US Geological Survey Data Series*, 691(10):4–9.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Available online: http://www.R-project.org/ (accessed on 10 July 2019).

Ridout, M. and Linkie, M. (2009). Estimating Overlap of Daily Activity Patterns From Camera Trap Data. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(3):322–337.

Royal Botanic Gardens, Kew and Missouri Botanical (2013). *The Plant List*. Published online: http://www.theplantlist.org/ (accessed on 10 July 2019).

Schmid, F. and Schmidt, A. (2006). Nonparametric estimation of the coefficient of overlapping–theory and empirical application. *Computational Statistics and Data Analysis*, 50(6):1583–1596.

Townsend, A., Soberón, J., Pearson, R., Anderson, R., Martínez-Meyer, E., Nakamura, M., and Bastos, M. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press.

Trabucco, A. and Zomer, R. (2018). Global Aridity Index and Potential Evapo-Transpiration (ET0) Climate Database v2. *CGIAR Consortium for Spatial Information*, pages 1–6. Published online: https://cgiarcsi.community/data/global-aridity-and-pet-database/ (accessed on 10 July 2019).

Wand, M. and Jones, M. (1994). *Kernel Smoothing*. Chapman and Hall.

Yue, H. and Tomoyasu, M. (2005). Weighted principal component analysis and its applications to improve FDC performance. In *Conference Paper in Proceedings of the IEEE Conference on Decision and Control*, volume 4, pages 4262–4267.

Zou, K., Liu, A., Bandos, A., Ohno-Machado, L., and Rockette, H. (2016). *Statistical Evaluation of Diagnostic Performance*. Chapman and Hall.