PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Informe práctica profesional

Colaboración con:

cienciambiental

Practicante: Diego Aravena Morales
Correo: daravena836@uc.cl

Santiago, Chile, Enero 2024

Índice

1 | Resumen práctica profesional

Cienciambiental es una empresa especializada en ofrecer asesorías para la gestión del medioambiente, centrándose en el estudio, análisis y comprensión de fenómenos ecológicos. Su enfoque inicial se orienta a minimizar conflictos, evaluar la factibilidad ambiental de proyectos y cumplir con la normativa. Actualmente, la empresa busca transformar la perspectiva de sus clientes en la gestión de la biodiversidad, pasando de una posición pasiva a una proactiva. Para lograrlo, utilizan herramientas avanzadas como planes de manejo, servicios de inteligencia ambiental, sistemas de gestión de la biodiversidad y modelamiento meteorológico e hídrico de alta precisión. Esta se posiciona como una opción para una gestión moderna, efectiva y sustentable en proyectos y desafíos relacionados con el medio ambiente.

En particular, el equipo de ciencia de datos se dedica principalmente al estudio y modelamiento utilizando el software R, que desempeña un papel fundamental en el manejo y exploración de los datos. Un aspecto destacado de nuestro equipo es su constante búsqueda de nuevas herramientas de trabajo que permitan obtener resultados de manera más eficiente, sin comprometer la precisión en un campo tan complejo como lo es el medio ambiente. Este enfoque refleja el espíritu innovador de Cienciambiental.

Mi rol principal en el equipo consistió en evaluar el impacto de diversas variables climáticas sobre la biodiversidad marina en el Norte de Chile. Para lograrlo, atravesé varias etapas, las cuales incluyeron la sistematización de diversas fuentes de datos que me fueron proporcionados, seguido de la detección de patrones y eventos de interés. Finalmente, y de manera crucial, me dediqué a modelar una variable de interés a modo de obtener resultados con significado tangible tanto para personas naturales como jurídicas.

Además de cumplir con mis responsabilidades asignadas, el principal desafío surgió en la última etapa, ya que se me solicitó aprender e implementar un nuevo paquete de R, *Tidymodels* [2], el cual permite aplicar diversos métodos de machine learning de forma automática, optimizando y determinando los parámetros más adecuados para los datos. Finalmente, realicé una breve presentación para explicar su funcionamiento, permitiendo así que pueda ser implementado en futuros proyectos de la empresa.

2 | Introducción Proyecto

El cambio climático representa uno de los mayores desafíos ambientales de nuestro tiempo, con consecuencias significativas en la biodiversidad y específicamente sobre los ecosistemas marinos. Es más que evidente que los patrones climáticos están experimentando alteraciones sustanciales, manifestándose en fenómenos como el calentamiento global, cambios en los regímenes de precipitación y eventos climáticos extremos. Estos cambios tienen ramificaciones profundas en los ecosistemas marinos, afectando la distribución, comportamiento y supervivencia de las especies que los habitan.

En particular, las regiones costeras del Norte de Chile se encuentran en una posición geográfica vulnerable, donde la combinación de factores climáticos como el aumento de las temperaturas

y las variaciones en los patrones de corrientes oceánicas podría tener un impacto significativo en la riqueza de especies marinas.

El presente estudio tiene como objetivo explorar la relación entre variables climáticas y la riqueza de especies marinas en el Norte de Chile. La comprensión de cómo estas variables climáticas influyen en la biodiversidad marina es crucial para anticipar y mitigar posibles efectos adversos.

3 | Descripción de los datos

La información climática fue extraída a través de las plataformas de Google Earth Engine (G.E.E) [3] y del Panel de Monitoreo ENSO de la NOAA [4]. La información recopilada a través de G.E.E se origina a partir de una diversidad de fuentes geoespaciales que capturan una amplia gama de fenómenos terrestres, desde cambios en la cobertura vegetal hasta variaciones climáticas a lo largo del tiempo. Por otro lado, el Panel de Monitoreo ENSO de la NOAA recopila información relacionada con El Niño y La Niña, fundamentada en observaciones in situ, mediciones de temperatura del mar, patrones atmosféricos y otros indicadores clave. Estos datos permiten una comprensión profunda de los patrones climáticos del Pacífico ecuatorial y desempeñan un papel crucial en la predicción y monitoreo de eventos climáticos extremos en todo el mundo.

En adición, a través del portal de transparencia [1] se solicitó a SERNAPESCA información sobre la extracción de diversas especies marinas en distintos puntos del Norte de Chile, en particular, de las regiones Arica y Parinacota, Tarapacá, Antofagasta y Atacama. Tal como se observa en la Figura 3.1, cada punto corresponde a una caleta, en el cual se registran mensualmente la cantidad de toneladas de especies marinas distintas que se logran extraer.

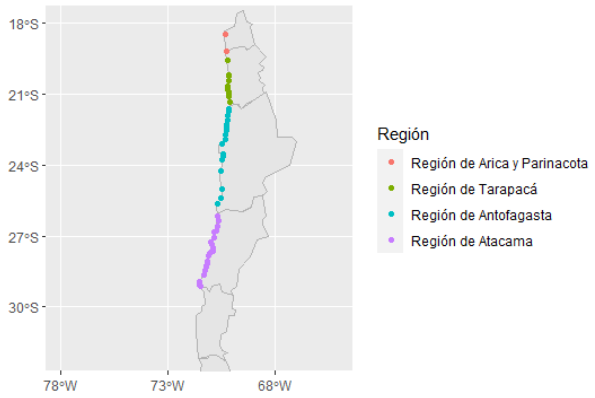

Figura 3.1: Ubicación Caletas en el Norte de Chile

El período de tiempo del cual se tiene información sobre la extracción de las especies data entre los años 2000 y 2022. Luego, con el fin de tener un solo conjunto de datos se cruzó la información geográfica de las caletas con las climáticas extraídas del G.E.E y del panel de monitoreo ENSO para cada mes y año entre dicho período. Además, cabe destacar que la

información obtenida desde G.E.E es para cada caleta en particular, mas no la obtenida a través de la NOAA, pues estas son variables globales y no cambian en cada caleta. Finalmente, se logró armar un set de datos con 187.212 filas y 28 columnas. La descripción de las variables se encuentra detallada en la Tabla 10.1 del Anexo.

4 | Análisis exploratorio

El principal interés está en identificar patrones visuales que nos permita conocer el panorama de la situación. En primer lugar, se busca conocer el estado de la riqueza marina a través de los años y tratar de identificar alguna tendencia. La Figura 4.1 nos presenta la riqueza de cada región mensualmente, siendo posible observar rápidamente que las 4 regiones presentan una disminución a partir del año 2008. Además, se puede concluir que el comportamiento es similar en todo el Norte, por ende, no sería de gran interés distinguir entre regiones.

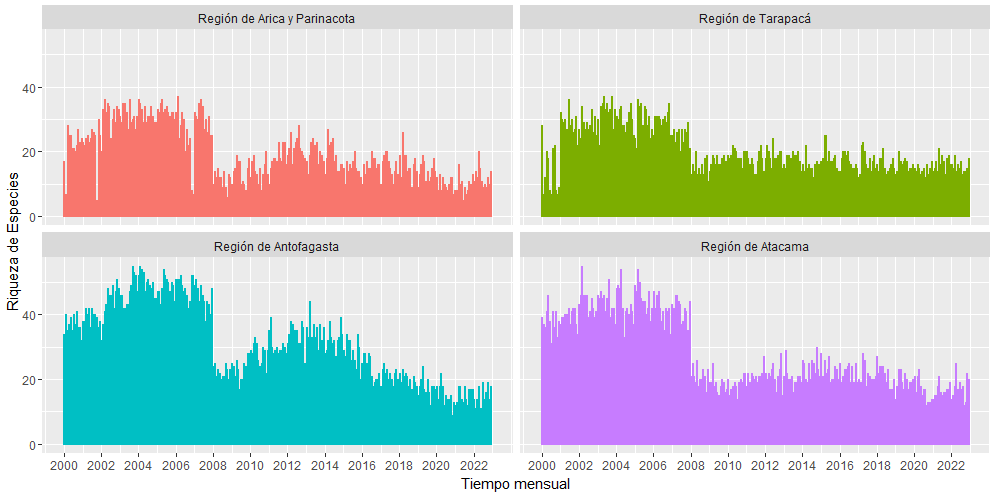

Figura 4.1: Riqueza de especies por región a través del tiempo

Luego, es el turno de revisar el comportamiento de las variables climáticas que se tienen a disposición. En la Figura 4.2 se puede observar que la gran mayoría mantiene un patrón estacional y no se detectan tendencias muy claras a excepción de *pdo.timeseries.sstens* y *vel_v* que tienden a disminuir en los últimos años. Cabe destacar que en algunas variables se pueden observar algunos espacios vacíos, que vendrían siendo datos faltantes y cuyo tratamiento se hablará más adelante en la sección 5.1.

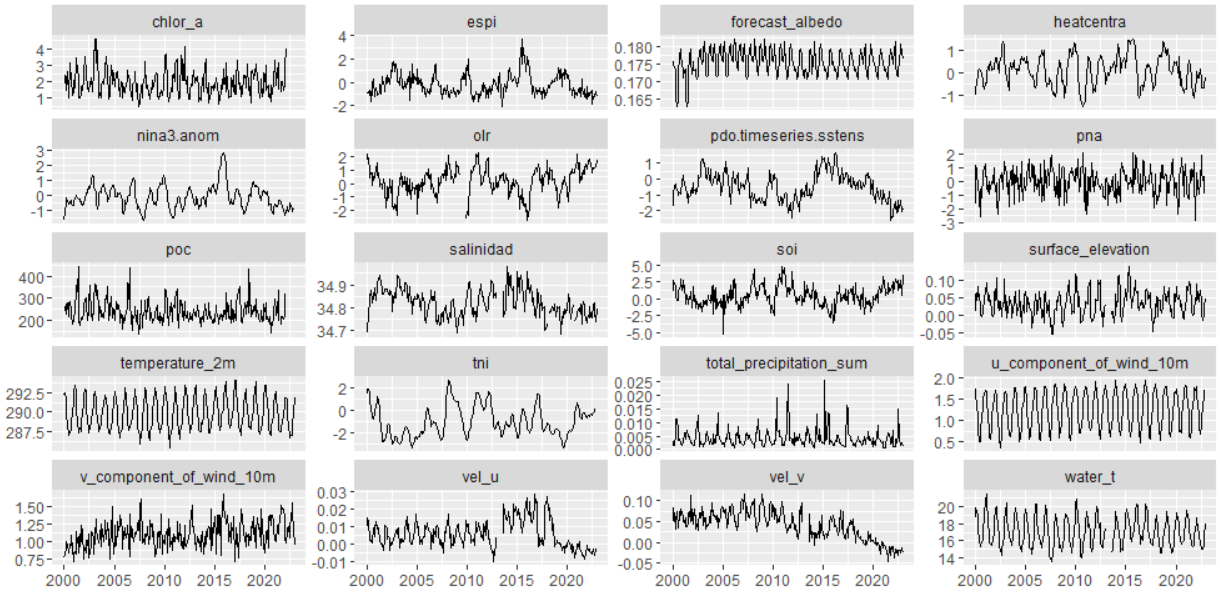

Figura 4.2: Variables climáticas a través del tiempo

Complementando los comentarios realizados acerca de la Figura 4.1, en la Figura 4.2 vemos que las variables no llegan a tener un comportamiento anómalo en el año 2008, lo cual nos lleva a pensar que aquel suceso no es atribuible a los factores climáticos que se tienen a disposición.

5 | Preprocesamiento

5.1 | Análisis de datos faltantes

Tal como se ve en la Figura 4.2, existen datos faltantes en algunas covariables, en particular en 9 de ellas. El máximo de observaciones que se pierde por variable es de a lo más 12, es decir, se perdería información de un año en total, por lo cual se decide imputar aquellos valores utilizando dos métodos. El primero es mediante interpolación lineal y estará enfocado en aquellas observaciones que no estén en los extremos, es decir, que no estén en el primer mes del año 2000 y en el último mes del año 2022. Este hecho sí ocurre en las variables *chlor_a*, *poc* y *tni*, por ende, en ellas se aplica el segundo método que sería imputación por la media de los meses respectivos.

5.2 | Análisis de componentes principales

La Figura 10.1 del Anexo muestra los coeficientes de correlación de Pearson de todas las covariables. En general no se logra establecer un patrón claro de covariables que presenten correlaciones demasiado altas entre si, no obstante, cabe destacar que las variables relacionadas con fenómenos intercontinentales, como el niño o la niña, tienen un grado de asociación en ambos sentidos más fuerte que el resto.

Dado que el enfoque se centra en estudiar el grado de asociación o importancia de los factores climáticos en la presencia de fauna marina, parece razonable incorporar un análisis de componentes principales. De esta manera, se estaría investigando si los patrones climáticos capturan la variabilidad en la riqueza de especies.

6 | Metodología

Para efectos de este estudio, a nivel regional se agruparon las variables climáticas obteniendo la mediana para cada mes y año. La variable respuesta corresponde a la riqueza de especies en cada instante de tiempo, lo cual estaría sugiriendo la posibilidad de ser una respuesta poisson.

Se han considerado dos conjuntos de datos; el primero se construye con base en los predictores originales, ya agrupados, y el segundo utilizando las componentes principales. Sin embargo, al observar la Figura 6.1, se logra identificar que la riqueza de especies presenta dos medias debido a un fenómeno ocurrido en el año 2008. Por lo tanto, se decide ajustar por separado estos dos grupos para evitar discrepancias en los resultados, descartando así un modelo de mezclas debido a su complejidad. Esta resolución nos permite analizar la influencia de las variables climáticas antes y después del año 2008. Para ello, se han considerado diversos algoritmos de métodos supervisados, incluyendo una regresión Poisson regularizada y otros capaces de identificar patrones no lineales.

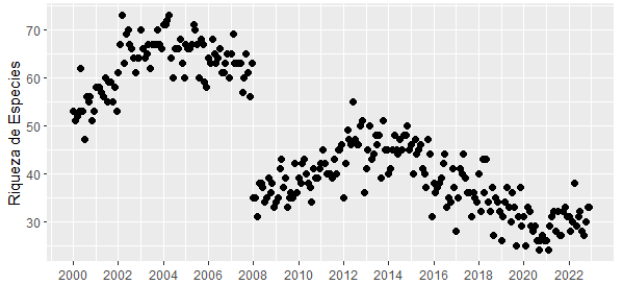

Figura 6.1: Riqueza de especies en el norte de Chile a través del tiempo

Para ajustar simultáneamente una amplia variedad de algoritmos, se empleó el paquete *Tidymodels*, el cual proporciona una estructura consistente para entrenar y evaluar diversos modelos, además facilita la integración de múltiples pasos en un único flujo de trabajo.

Cada conjunto de datos se dividió en un grupo de entrenamiento (80 %) y uno de prueba (20 %). Luego, se aplicó validación cruzada dividiendo el conjunto de entrenamiento en 10 grupos. Los algoritmos considerados para evaluar la importancia de las covariables incluyen la regresión poisson regularizada, árboles de regresión, árboles potenciados por gradiente, bosques aleatorios y regresión spline adaptativo multivariado. Además, se incorporaron otros métodos que podrían mejorar la performance predictiva, como máquinas de vector de soporte polinómico y radial, redes neuronales y k-vecinos más cercanos.

A lo largo de todo el proceso las funciones que carga *Tidymodels* que ejecutan todo el algoritmo consideran diversas combinaciones de parámetros, los cuales le indico que ajuste automáticamente con el objetivo de encontrar aquellos que proporcionen las mejores métricas. En este contexto, la métrica principal a considerar es la raíz del error cuadrático medio presentado en la ecuación (6.1), donde valores cercanos a cero estarán asociados a un mejor rendimiento.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \rightarrow \sqrt{MSE} = RMSE \tag{6.1}$$

La Figura 10.2 del Anexo es una representación del esquema del proceso de modelado, el cual ayuda a entender el proceso que sigue para determinar el método o modelo que logre el mejor rendimiento.

7 | Resultados

De inicio se modelaron los datos originales distinguiendo entre los que fueron tomados antes y desde del año 2008. Gracias al uso de *Tidymodels*, comparar los resultados es un proceso bastante sencillo y rápido debido a cómo se ha desarrollado este paquete.

Partiendo por el conjunto de datos previo al año 2008, en la primera fila de la Figura 10.3 del Anexo se ordenan todos los modelos según un intervalo de confianza para el RMSE, siendo el primer lugar el árbol de regresión potenciado por gradiente. En la segunda fila de la primera columna se muestran los datos observados versus predichos, donde se nota una leve sobre valoración de lo observado, aunque el valor del RMSE luego de testear es de 4.38 (véase Tabla 10.2). Por último, se tienen las variables que más influencia tienen sobre el cambio en la riqueza de especies, cuyos grados más altos son adquiridos por el índice del Niño 3, el gradiente TSM, la velocidad horizontal del aire en dirección norte y la oscilación decadal del pacífico. Vemos que 3 de los 4 factores más importantes son atribuibles a variables globales, mientras que sólo hay una variable local.

Por otro lado, en la Figura 10.4 en el Anexo se muestran los resultados luego de modelar aquellas observaciones que fueron tomadas desde el año 2008. Si bien el modelo con mejor RMSE corresponde al de máquina de vector de soporte polinómico, la mejor performance en los datos de testeo se lo lleva K-vecinos más cercanos con un RMSE de 3.43 (véase Tabla 10.2). El gráfico de lo observado versus predicho explica este valor, vemos que estos tienden a ser bastante similares, por lo que se trataría de un modelo adecuado. Sin embargo, el árbol de regresión potenciado por gradiente está bastante cerca en cuanto a rendimiento a nivel general, además nos interesa este método en particular, ya que, nos permite identificar la importancia de las variables. Al ver el gráfico en la segunda columna de la segunda fila se puede notar claramente que la mayor influencia es propiciada por la salinidad y la velocidad del mar tanto en dirección Este y Norte, que corresponden a variables locales.

Ahora es el turno de analizar los resultados obtenidos al utilizar componentes principales. Para los datos que fueron tomados antes del año 2008 se consideraron las primeras 10 componentes

principales, que abarcan un 90.25 % de la variabilidad total de las variables climáticas. Por otro lado, para aquellos datos registrados desde el año 2008 se tuvieron en cuenta las primeras 11 componentes principales, logrando abarcar un 91.27 % de la variabilidad total de las covariables.

La Figura 10.5 del Anexo presenta los resultados obtenidos luego de modelar el primer conjunto de datos (previo al año 2008) utilizando componentes principales. El modelo con mejor rendimiento corresponde a la regresión poisson regularizada, cuyo RMSE de testeo es igual a 5, más alto que el modelo obtenido con los datos originales. El componente principal más importante es *V8*, es decir, la octava componente, el cual cubre un 3.38 % de la variabilidad total. Si se observa la Figura 10.6 del Anexo, el coeficiente que abarca mayor importancia en la octava componente es la salinidad, seguido por un promedio ponderado mayormente influenciado por aquellos coeficientes con valores cercanos a 0.3, los cuales corresponden a variables globales. Lo anterior significa que aquella combinación lineal, donde salinidad posee mayor peso, tiene mayor influencia en la predicción del modelo.

Por último, el mejor modelo que se obtiene al considerar los datos ingresados desde el año 2008 utilizando componentes principales es el árbol de regresión potenciado por gradiente, cuyo RMSE de testeo es 3.59 (véase Tabla 10.2), levemente más alto que el modelo utilizando los datos originales. Al observar la Figura 10.7 en el Anexo es posible percibir rápidamente que el grado de importancia del cuarto componente principal es notablemente superior al resto. Este componente abarca un 7.28 % de la variabilidad total y está mayormente influenciado por la velocidad del mar en dirección Norte y Este, seguido de la velocidad del aire en dirección Norte; todas variables locales (véase Figura 10.8 del Anexo).

7.1 | Interpretación resultados

Nótese que en los resultados presentados anteriormente tan sólo se señalaron aquellas variables con mayor grado de importancia. Si bien esto carece de un cierto grado de interpretación, en esta sección se busca comprender en qué sentido aquellas covariables logran afectar la variable respuesta.

De aquellas variables que más influencia tienen previo al año 2008, se puede observar en la Figura 7.1 que el gradiente TSM (*tni*) es la única variable con una asociación negativa con la variable respuesta, mientras que el resto sigue una tendencia positiva a excepción de la salinidad, la cual no logra determinar niguna asociación, esto explica aquel valor-p asociado a la regresión lineal ajustada (línea azul) que es altamente no significativo. Cabe destacar que estas regresiones sólo se utilizan con el fin de tener una mejor comprensión de la tendencia.

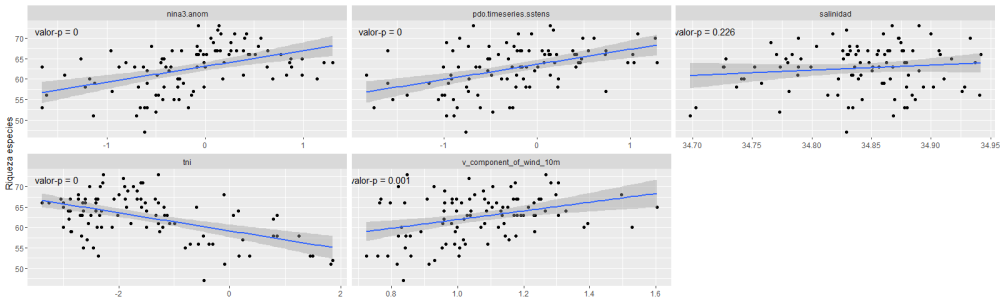

Figura 7.1: Gráfico de asociación entre la riqueza de especies y las variables más influyentes previo al año 2008

Por otro lado, se puede observar en la Figura 7.2 que la velocidad horizontal del aire en dirección norte es la única variable, de aquellas más influyentes, cuya asociación es negativa con la riqueza de especies en aquellos datos que fueron observados desde el año 2008.

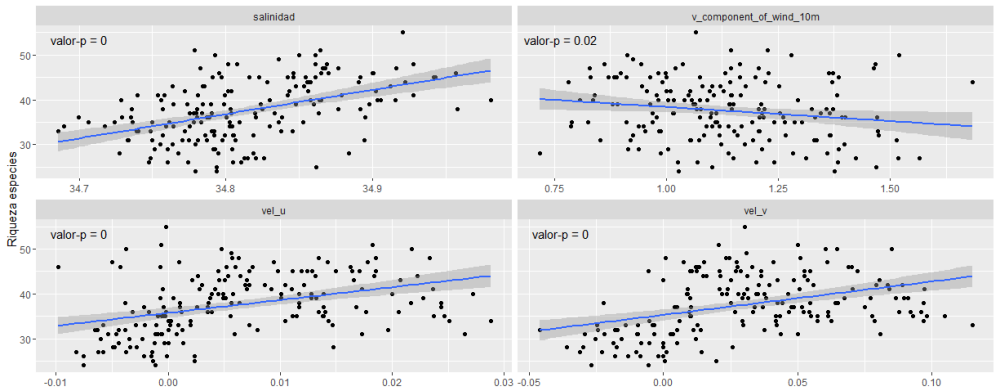

Figura 7.2: Gráfico de asociación entre la riqueza de especies y las variables más influyentes a partir del año 2008

8 | Conclusiones

En primer lugar, se detectó un cambio drástico en la variable de respuesta en el año 2008, del cual no se logró obtener una explicación. Esto generó comportamientos diferentes en la respuesta antes y después de dicho evento, por lo que se optó por llevar a cabo el estudio dividiendo el conjunto de datos en torno a este acontecimiento

Tras la imputación de los datos faltantes en las variables predictoras se facilitó el desarrollo del trabajo. Posteriormente, se dividieron los datos en dos conjuntos: uno centrado en la información registrada antes del año 2008 y otro en la registrada a partir de ese año. Ambos conjuntos se subdividieron en dos grupos, uno para los datos originales y otro para los componentes principales, resultando en un total de cuatro conjuntos de datos con los que se llevó a cabo el análisis.

Luego, gracias al paquete *Tidymodels* se logró plantear una gran cantidad de métodos y modelos con el objetivo de encontrar aquellos que mejor logren discriminar las observaciones, es decir, que sean capaces de identificar tendencias y/o patrones que permitan encontrar las variables climáticas más influyentes. No obstante, un aspecto negativo es que en el ecosistema de *Tidymodels* no existe una función específica para ajustar modelos de mezclas, por lo que se propone encontrar una manera adecuada de incorporar este tipo de modelos utilizando como motor alguna función de R que si lo permita.

La modelación de la riqueza de especies previa al año 2008 fue realizada de manera efectiva, evidenciando un margen de error aproximado de 4 a 5 unidades. Además, se identificó que las variables climáticas que más inciden en el cambio de la riqueza están principalmente vinculadas con eventos a nivel global, siendo el gradiente de la Temperatura Superficial del Mar (TSM) el factor que ejerce la mayor influencia negativa en este aspecto. El resto de variables más influyentes, sorpresivamente, tienden a aumentar la riqueza (a excepción de la salinidad, del cual no se puede concluir nada concreto).

La obtención de modelos adecuados para la riqueza de especies a partir del año 2008 también fue exitosa, revelando errores que oscilan entre 3 y 5 unidades. No obstante, en comparación con el escenario anterior, se observó que las variables climáticas con mayor impacto en la riqueza estaban específicamente asociadas a la región del Norte. La velocidad horizontal del aire a 10 metros sobre el nivel del mar se destacó como la variable que tiende a disminuir la riqueza. Esta observación suscita la interrogante sobre la influencia de las variables globales en las locales, planteándonos la posibilidad de cambios significativos capaces de afectar la biodiversidad marina.

Finalmente, resultaría de gran interés explorar el estado futuro de la riqueza en un lapso de 10 años. Esta posibilidad se torna alcanzable mediante la realización de predicciones para cada covariable durante dicho período y la posterior utilización de los modelos previamente desarrollados para anticipar la evolución de la riqueza. De esta manera, este estudio aportará valiosa información a la creciente base de conocimientos acerca de la vulnerabilidad de los ecosistemas marinos frente al cambio climático, respaldando así los esfuerzos futuros destinados a conservar y proteger la rica biodiversidad característica de esta singular región del mundo.

9 | Referencias

[1] Ley 20.285/2008 sobre acceso a la información pública, 2008. Publicada en el Diario Oficial de la República de Chile el día 20 de Agosto de 2008.

[2] Equipo de Desarrollo de Tidymodels. *Tidymodels: A Collection of Packages for Modeling and Machine Learning with 'tidyverse' Compatibility*. The Comprehensive R Archive Network (CRAN), 2023. Consultado el 10 de Diciembre, 2023.

[3] Google Earth Engine. Google earth engine - datasets, 2023. Consultado el Día 31 de Octubre, 2023.

[4] NOAA - Physical Sciences Laboratory. Enso: Recent evolution, current status and predictions, 2023. Consultado el Día 31 de Octubre, 2023.

10 | Anexo

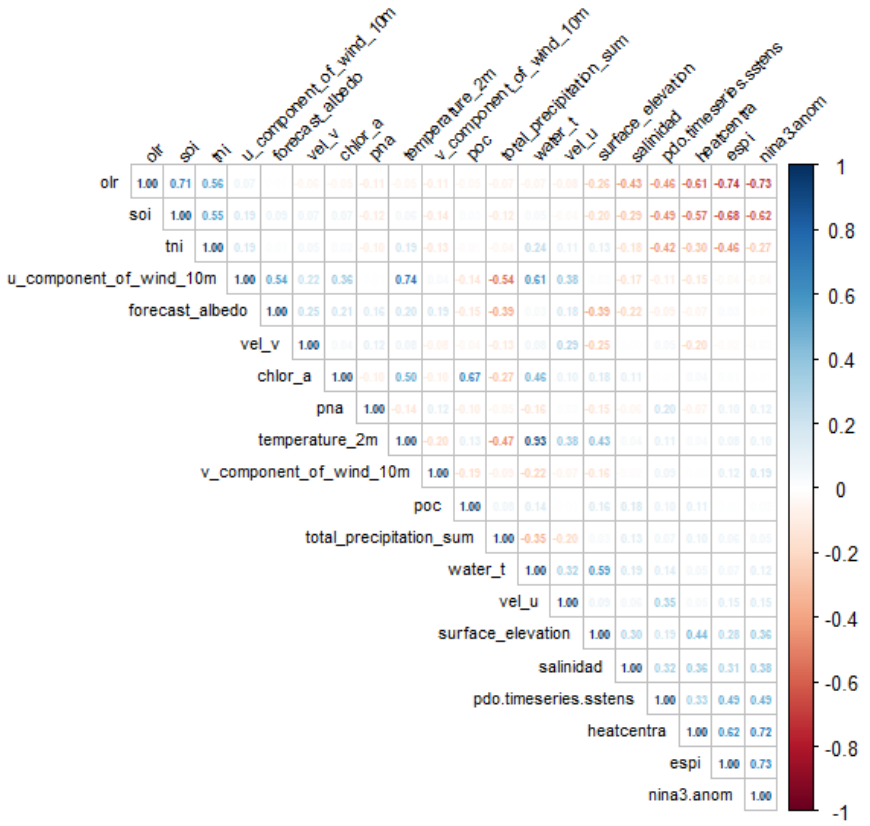

Figura 10.1: Correlación entre variables predictoras

Variable	Descripción
año	Año
mes	Mes
caleta	Nombre de la caleta
c_region	Numero Región
especie	Nombre especie extraída
desembarco	Toneladas de desembarco
lat	Latitud ubicación caleta
lon	Longitud ubicación caleta
forecast_albedo	Es la fracción de la radiación solar (directa o difusa) de onda corta reflejada por la superficie de la tierra
temperature_2m	Temperatura en grados Kelvin del aire a 2 m sobre la superficie
total_precipitacion_sum	Agua líquida y congelada acumulada, incluidas la lluvia y la nieve, que cae a la superficie de la Tierra
u_component_of_wind	Velocidad horizontal en m/s del aire en dirección al este, a una altura de diez metros sobre la superficie
v_component_of_wind	Velocidad horizontal en m/s del aire en dirección al norte, a una altura de diez metros sobre la superficie
chlor_a	Concentración de clorofila-a
poc	Carbono orgánico particulado
surface_elevation	Anomalía de la elevación de la superficie del mar en relación con la elevación media modelada con factor de escala aplicado a los datos
salinidad	Promedio salinidad del mar a diferentes profundidades (0 a 20 metros)
vel_u	Promedio velocidad del mar en dirección al este para diferentes profundidades (0 a 20 metros)
vel_v	Promedio velocidad del mar en dirección al norte para diferentes profundidades (0 a 20 metros)
water_t	Promedio temperatura en grados Celcius del mar para diferentes profundidades (0 a 20 metros)
nina3.anom	Índice del niño 3
soi	Diferencia entre los valores estandarizados de presión superficial de Darwin y Tahití
tni	Gradiente de la TSM en la región ENSO del Pacífico tropical
pdo.timeseries.sstens	Oscilación decadal del pacífico
pna	Variabilidad de baja frecuencia de los extratrópicos del hemisferio norte
olr	Área de radiación de onda larga saliente promediado sobre el Pacífico ecuatorial central
heatcentra	Anomalía media de temperatura ecuatorial sobre los 300 m para 160E-80W
espi	Índice de precipitación ENOS

Cuadro 10.1: Descripción variables

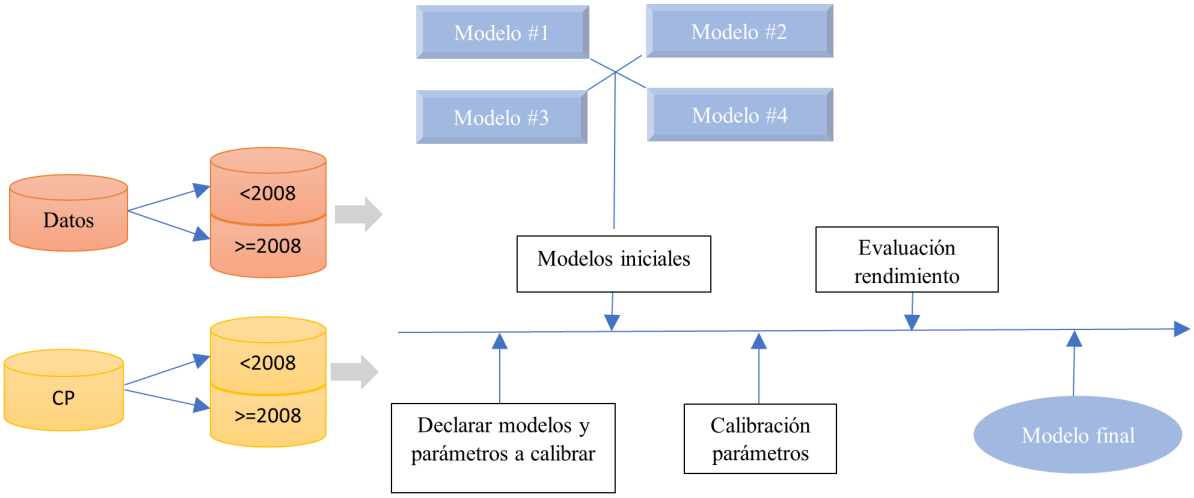

Figura 10.2: Esquema proceso de modelamiento

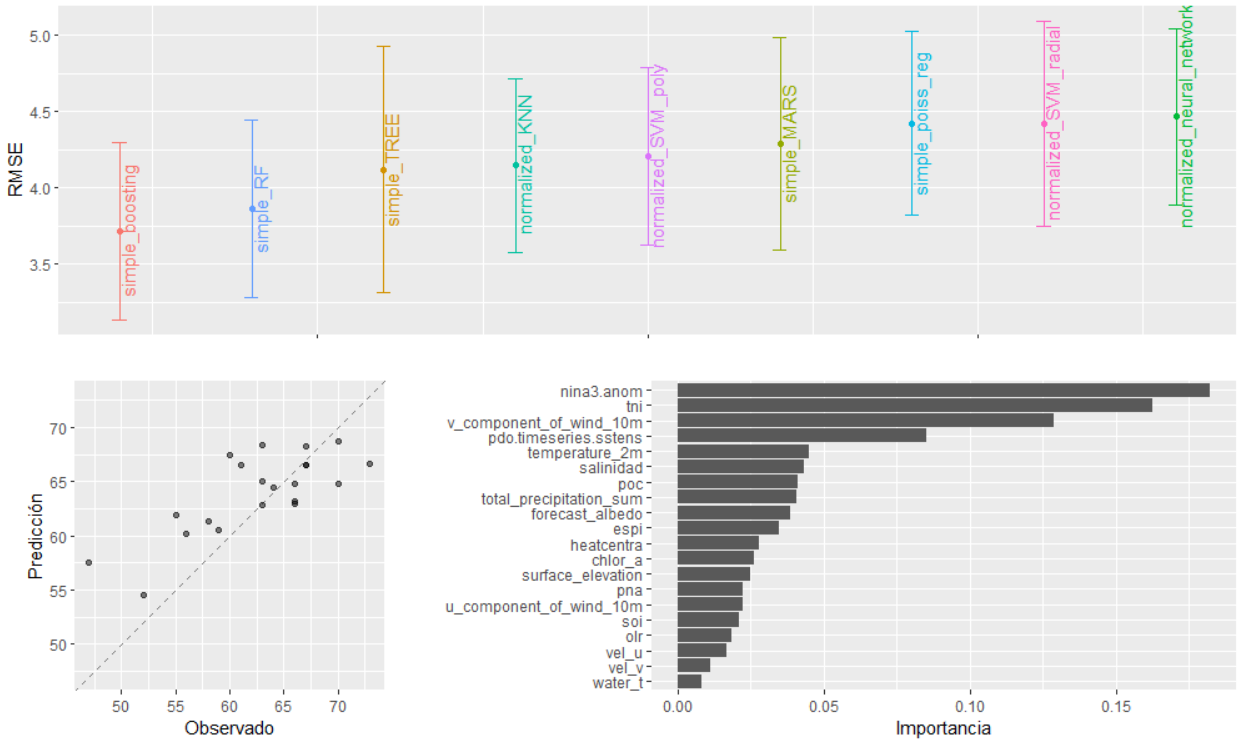

Figura 10.3: Resultados utilizando datos originales previo al año 2008

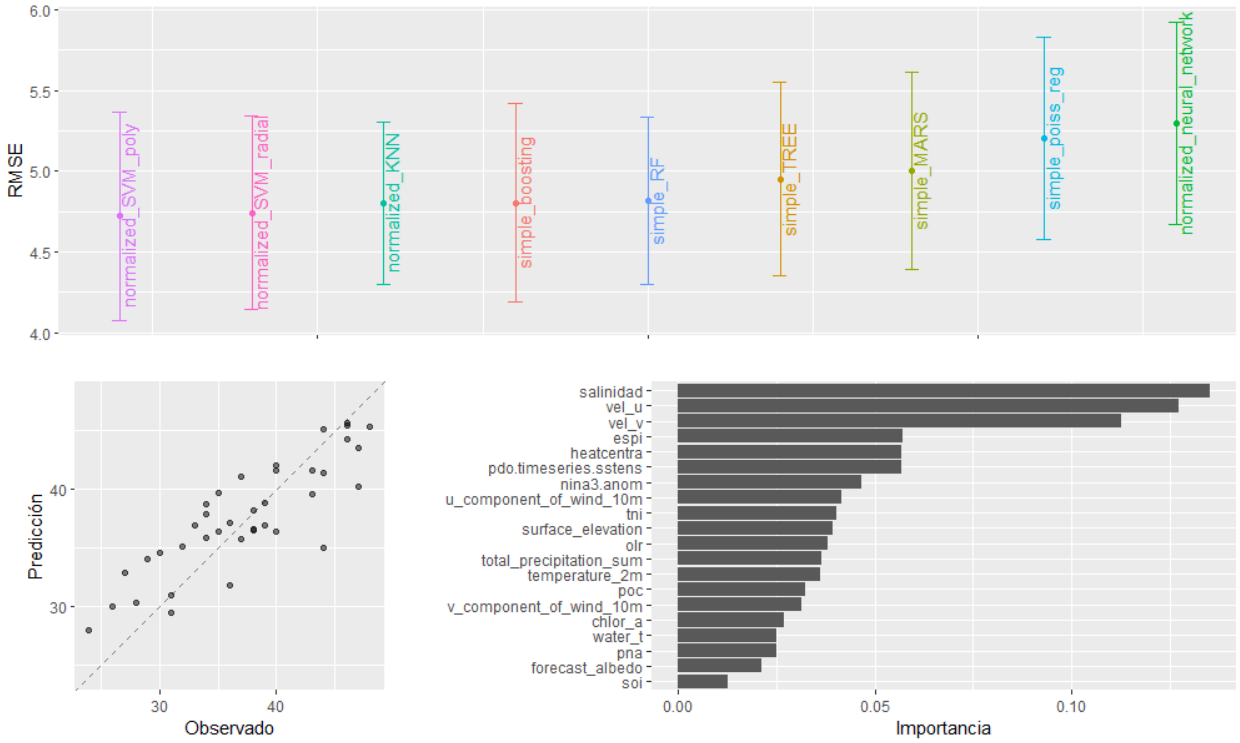

Figura 10.4: Resultados utilizando datos originales desde año 2008

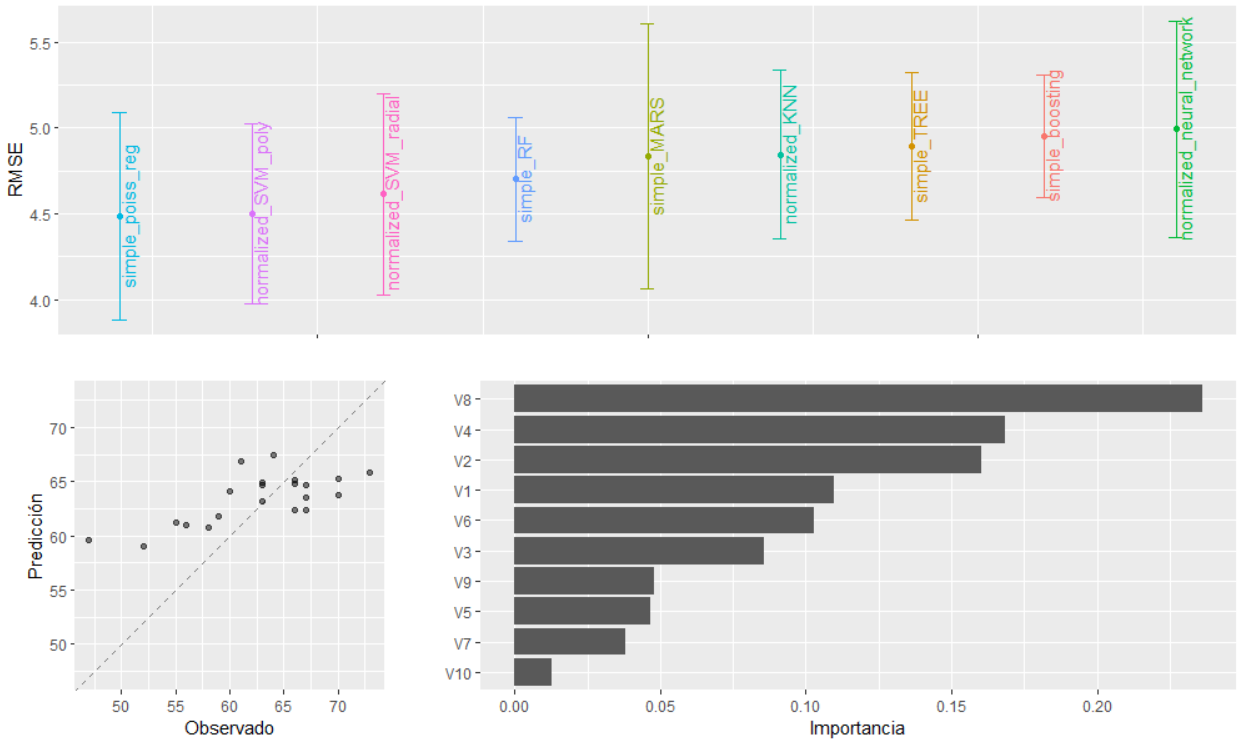

Figura 10.5: Resultados utilizando componentes principales previo año 2008

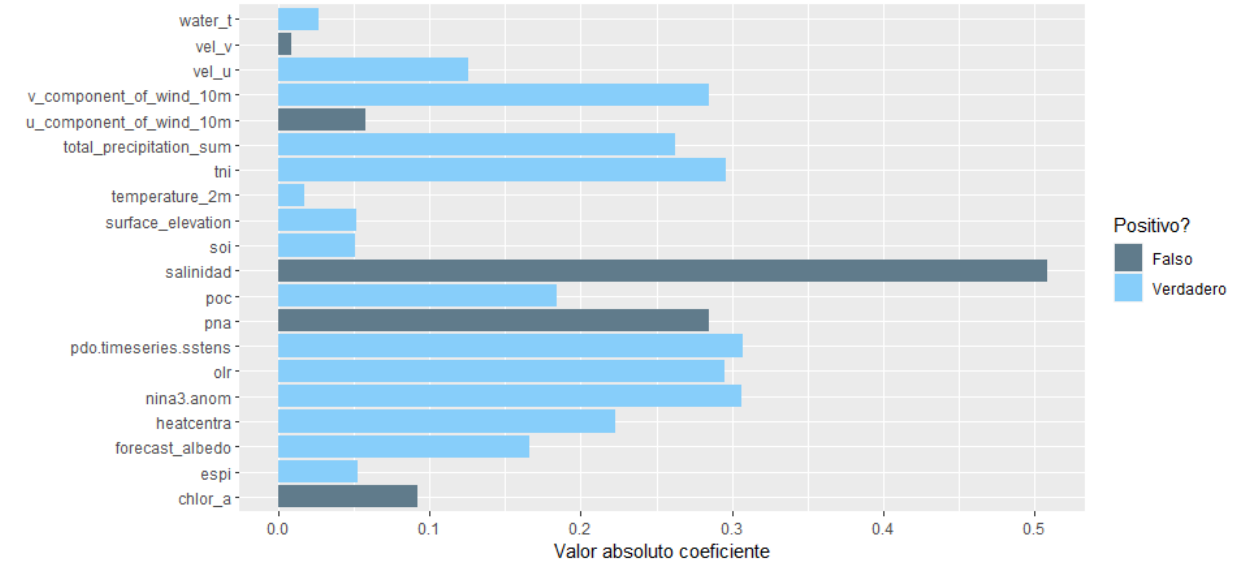

Figura 10.6: Composición octava componente principal para datos previos al año 2008

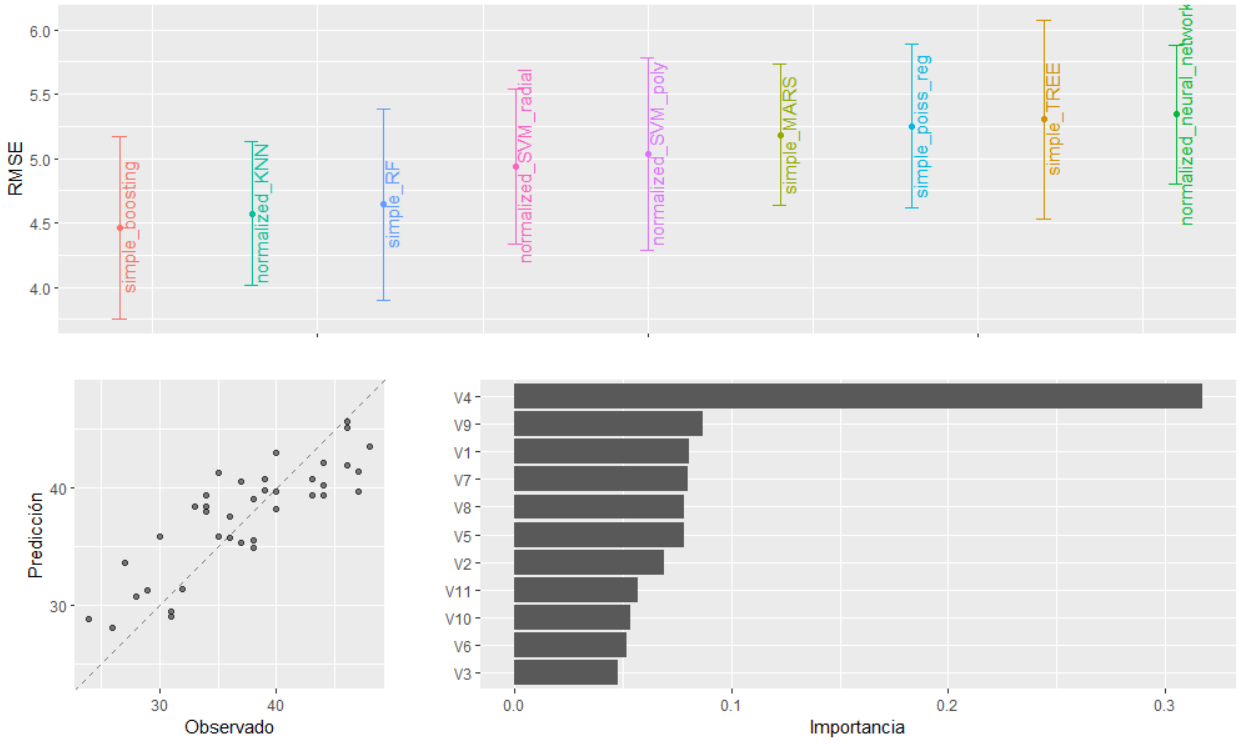

Figura 10.7: Resultados utilizando componentes principales desde año 2008

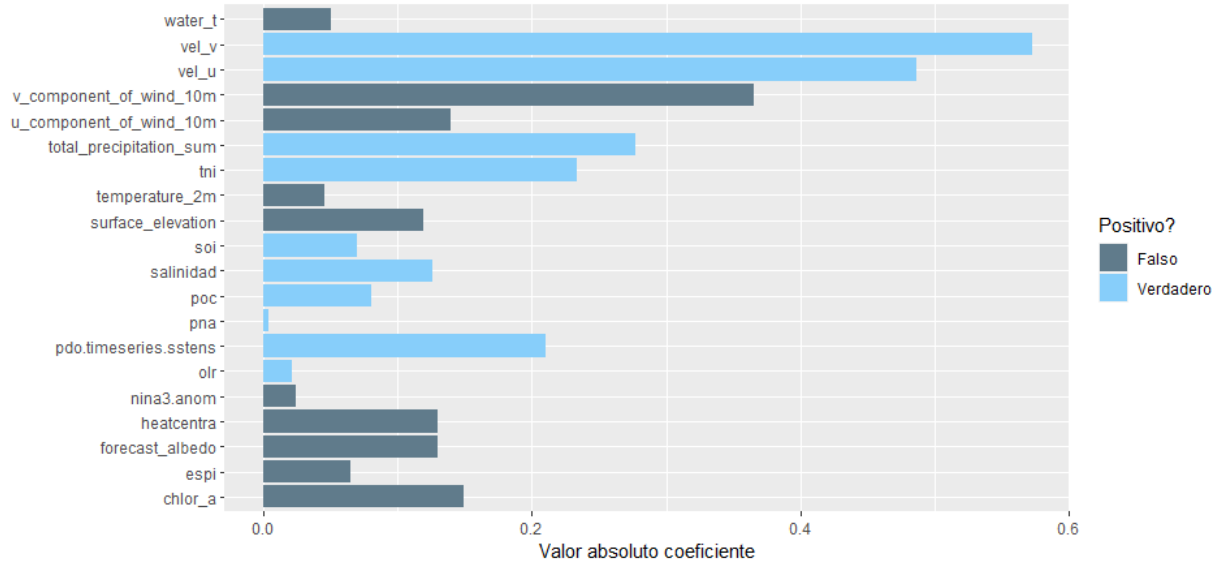

Figura 10.8: Composición cuarta componente principal para datos desde el año 2008

Fuente	**Datos**	**Mejor método**	**Métrica**	**Media entrenamiento**	**Test**
Datos originales	Previo año 2008	Árbol potenciado por gradiente	RMSE	3.72	4.38
	Post año 2008	K-vecinos más cercanos	RMSE	4.8	3.43
Componentes principales	Previo año 2008	Reg. poisson regularizado	RMSE	4.49	5.00
	Post año 2008	Árbol potenciado por gradiente	RMSE	4.46	3.59

Cuadro 10.2: Métricas de los mejores modelos que fueron obtenidos para cada conjunto de datos