Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería
Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Caracterización y modelado de eventos extremos en estaciones centenarias de las cuencas del Paraná, Uruguay y Limay

Tesis presentada para optar al título de Magíster en Explotación de Datos y Descubrimiento de Conocimiento

Luciana Quarracino

Directora: Dra. Melanie Meis

Codirector: Dr. Maximiliano Camporino

Buenos Aires, febrero 2025

RESUMEN

La presente tesis de maestría se centró en el estudio de los ríos Paraná, Limay y Uruguay analizando los patrones temporales de las series de caudal con el objetivo de agrupar las crónicas hidrológicas y pronosticar el comportamiento futuro de las cuencas.

En una primera instancia se realizó un análisis exploratorio de datos para cada río y se analizaron los cambios en las tendencias históricas. Asimismo, se encontraron correlaciones entre los meses, trimestres y medidas descriptivas de las crónicas hidrológicas.

Además, se identificaron patrones comunes en los regímenes de los ríos y se utilizó la técnica de agrupamiento K-means que permitió clasificar las crónicas en cuatro clústeres de acuerdo a características comunes relacionadas con la humedad de las crónicas y la temporalidad de esa humedad dentro del año.

Finalmente, los modelos de pronóstico aplicados mostraron que los árboles de decisión son efectivos para predecir a partir de variables del primer trimestre, mientras que la regresión logística con penalización Ridge es más adecuada para los casos en los que se cuenta también con información del segundo y tercer trimestre. Además, se evaluó la influencia del ENSO en el comportamiento de las crónicas, concluyendo que en este caso no aporta información significativa para mejorar los modelos de pronóstico.

Estos resultados podrían replicarse en otras cuencas del país, permitiendo un uso más eficiente de los recursos hídricos y un mejor manejo anticipatorio de los mismos. Asimismo, abre la posibilidad de investigar otras oscilaciones climáticas y períodos de tiempo para mejorar la precisión de los modelos de pronóstico.

Palabras claves: Agrupamiento, pronóstico, caudal, componentes principales, ENSO.

ABSTRACT

Title: Characterization and Modeling of Extreme Events in Centennial Stations of the Paraná, Uruguay, and Limay River Basins

This master's thesis focused on the study of the Paraná, Limay, and Uruguay rivers, analyzing the temporal patterns of flow series with the aim of grouping hydrological chronicles and forecasting future basin behavior.

Initially, an exploratory data analysis was conducted for each river, and changes in historical trends were examined. Additionally, correlations were found between months, quarters, and descriptive measures of the hydrological chronicles.

Furthermore, common patterns were identified in the river regimes, and the K-means clustering technique was used to classify the chronicles into four clusters based on common characteristics related to the moisture of the chronicles and the seasonality of that moisture within the year.

Finally, the applied forecasting models showed that decision trees are effective for predicting from first-quarter variables, while Ridge-penalized logistic regression is more suitable for cases where information from the second and third quarters is also available. Additionally, the influence of ENSO on the behavior of the chronicles was evaluated, concluding that in this case, it does not provide significant information to improve the forecasting models.

These results could be replicated in other basins in the country, allowing for more efficient use of water resources and better anticipatory management. It also opens the possibility of investigating other climatic oscillations and time periods to improve the accuracy of the forecasting models.

Keywords: Clustering, forecast, inflows, principal components, ENSO.

Índice general

INTRODUCCIÓN GENERAL

Descripción del problema y motivación

Los recursos hídricos resultan ser fundamentales para el progreso económico y social de un país. En particular, en la Argentina ciertas cuencas son esenciales para el desarrollo de la agricultura, la generación de energía eléctrica y el turismo así como también para la navegabilidad asociada tanto a las importaciones como a las exportaciones. De esta manera, es fundamental el continuo estudio hidrológico para la generación de sistemas de caracterización, prevención y alerta en las distintas cuencas de nuestro país procurando cooperar hacia un manejo adecuado de los distintos cuerpos de agua.

Más aún, la generación de herramientas e información para contribuir con los tomadores de decisiones resulta ser un desafío de largo proceso entre la academia y las instituciones operativas (Sillitoe, 2021). Sin embargo, la colaboración entre ambas partes es de carácter necesario y urgente (Liu et al., 2008). En ese sentido, de acuerdo con el informe del Grupo Intergubernamental de Expertos sobre el Cambio Climático (IPCC, por sus siglas en inglés) de 2021 (Seneviratne et al., 2021), está previsto que los eventos extremos como sequías e inundaciones aumenten en el futuro cercano. Por lo tanto, la aplicación de distintas técnicas y herramientas estadísticas es necesaria para colaborar con los tomadores de decisiones para la generación de modelos que permitan realizar una prevención y un manejo propicio de cuencas y ríos (Vanelli y Kobiyama, 2021).

En este trabajo de tesis se han considerado tres ríos fundamentales: el río Paraná y el río Uruguay, ambos pertenecientes a la Cuenca del Plata, y el río Limay, pertenenciente a la cuenca del río Negro. Los primeros dos ubicados en gran parte en la región este y noreste de la Argentina mientras que el tercero se ubica en la región Patagónica. La selección de estos ríos no ha sido arbitraria, los mismos representan un accionar fundamental en distintos sectores económicos de la Argentina (agricultura, energía, exportaciones, entre otras). En tal sentido, como ejemplo se puede mencionar que el 26,1 % de la generación de energía eléctrica del sistema interconectado argentino en la década 2013-2022 fue de origen hídrico. De dicha porción, el 86,0 % ha sido generado por represas hidroeléctricas ubicadas en los cauces de los ríos Paraná, Uruguay y Limay obteniendo un 47, 13 y 26 %

de la generación hidroeléctrica total, respectivamente [1].

A partir del estudio de los tres ríos, este trabajo de tesis busca obtener, en primer lugar, una caracterización básica de las series temporales de caudal para luego realizar un análisis predictivo de las mismas.

En particular, mediante la caracterización de los ríos se intenta identificar, clasificar y agrupar, para cada río, patrones temporales anuales que presenten ciertas similitudes explorando sus posibles causas. Además, se enfoca en poder determinar si ese agrupamiento resulta ser consistente en los tres ríos de interés. De esta manera, se pretende explorar distintos modelos predictivos que permitan determinar una posible planificación futura del uso de los sistemas hídricos, considerando las características propias de cada año hidrológico así como también su relación con las distintas oscilaciones climáticas tales como el ENSO (El Niño-Southern Oscillation).

Asimismo, la planificación del manejo de los recursos hídricos es esencial debido a su dependencia con los distintos sectores socio-económicos de un país. En particular, las sequías y las inundaciones resultan de gran interés debido a sus grandes impactos negativos. De esta manera, la generación de información y resultados que permitan contrarrestar los mismos es fundamental.

Más aún, identificar patrones temporales del caudal asociados a eventos que no representan la variabilidad media dentro de cada río y entre ríos es una tarea necesaria para la planificación de la disponibilidad de los cuerpos de agua en el mediano y largo plazo.

Finalmente, a pesar de que el plan de estudio del presente trabajo consideró el análisis de la estacionariedad de las series de caudal, los resultados presentados en el Apéndice A no han mostrado que exista una tendencia en las series temporales por lo que el estudio no continuó por dicha senda.

Trabajos previos

La continuidad del estudio hidrológico y su nexo climático es un área que ha sido desarrollada por varios autores dentro de la región. Sin embargo, la inclusión de nuevos registros temporales, así como también el uso de técnicas estadísticas no empleadas previamente pueden resultar de gran utilidad para colaborar con la generación de información

[1] Fuente: CAMMESA (Compañía Administradora del Mercado Mayorista Eléctrico). [Fecha de consulta: 21/06/2022]. `https://cammesaweb.cammesa.com/`

de los recursos hídricos.

Distintas investigaciones en la Cuenca del Plata, en particular en los ríos Paraná y Uruguay, han sido realizadas a lo largo de las últimas décadas. Los enfoques han sido variados, desde una caracterización de estadística básica de las cuencas ((Berbery y Barros, 2002); consideración de métodos espectrales en el caudal para obtener ciclos temporales y su relación con oscilaciones climáticas (Krepper et al., 2008); (Gulizia y Camilloni, 2020)); (Meis y Llano, 2019); hasta pronósticos estadísticos del caudal ((Meis y Llano, 2018); (Meis et al., 2022)), entre otros. Más aún, un gran número de trabajos han estudiado la correlación de las cuencas anteriores con las distintas oscilaciones climáticas ((Grimm et al., 2020); (Camilloni y Barros, 2000); (Camilloni y Barros, 2003)) estableciendo distintos nexos entre ciertas etapas del año en la señal del caudal y la variabilidad climática. Asimismo, el estudio de dichas relaciones en el tiempo es una tarea primordial. Antecedentes como el de (Meis et al., 2021) han mostrado que no sólo alcanza con establecer una relación entre el caudal y el ENSO, sino que también es posible cuantificar la misma bajo condiciones extremas de las variables. De esta manera, las autoras lograron obtener períodos de retorno conjunto y un eventual modelo para el pronóstico hidro-climático.

Los estudios realizados en la cuenca del río Limay han sido en número inferiores en comparación con las investigaciones en la cuenca del Paraná y la del Uruguay. De todas maneras se puede destacar el trabajo de (González et al., 2015) donde las autoras han estudiado de manera indirecta la asociación entre el caudal y la variabilidad climática mediante la precipitación estableciendo el desfasaje temporal entre los mismos. Más aún, en (Lauro et al., 2019) se estudió la relación entre distintas cuencas de la Argentina como la del río Limay y ciertos índices climáticos para un cierto período temporal afirmando aún más la existente relación entre la variabilidad climática y el regimen hidrológico. Incluso, parte de los autores anteriores han analizado la tendencia del caudal diario en algunas estaciones del río Limay siendo la misma negativa ((Lauro et al., 2018)). Por otro lado, en (Romero y González, 2016) han estudiado la relación entre el ciclo medio anual del caudal y la precipitación determinando el desfasaje temporal entre los mismos para el río Limay, resultando en un estudio útil para los sistemas de planificación hidroeléctricos.

Las sequías e inundaciones en las cuencas del Paraná, Uruguay y Limay no han sido hechos aislados y es por ello que distintos antecedentes han analizado estos eventos tan extraordinarios ((Abelen et al., 2015); (Antico et al., 2016); (González et al., 2017b);

(Barra, 2019); (Moraes et al., 2021)). En particular, determinar similitudes temporales dentro de cada río y entre los ríos resulta ser una herramienta útil de planificación para los distintos sectores como el de demanda hidroeléctrica (González et al., 2021).

Por otra parte, ciertos métodos de clasificación y reducción de dimensionalidad son comúnmente considerados en hidrología para encontrar patrones temporales. La utilidad de estos métodos radica en la posiblidad de simplificar el tamaño de los datos que se consideran así como también identificar patrones que resultan similares para luego intentar determinar posibles generadores de los mismos, puesto que una cuenca hídrica resulta ser una clara síntesis de la variabilidad climática en la región de estudio que se considere (Meis y Llano, 2019).

Para citar como ejemplo, en el trabajo de (Mansor et al., 2019) los autores estudiaron la regionalización de los ríos de Johor, Malasia, a partir de sus caudales y realizaron una agrupación de los mismos basados en patrones y cambios en series de tiempo utilizando la técnica de deformación dinámica del tiempo (DTW, por sus siglas en inglés). A su vez, en el estudio de (Brunner et al., 2020) también fueron aplicadas técnicas de agrupamiento (jerárquico y k-means) y se mostró una similitud de los ríos en cuanto a sus regímenes medios y también de excesos de caudal y sequías. En esta misma línea, (Mihailović et al., 2019) estudiaron una medida de disimilitud para aplicar un agrupamiento de estaciones de medición en el río Brazos, Estados Unidos, a partir de la información de caudal medio diario.

En cuanto al estudio de la regionalización de los ríos en Argentina, el mismo presenta un número reducido de trabajos realizados entre los que se pueden mencionar a (del Carmen Paris y Zucarelli, 2004) y (Zucarelli, 2017). En el primero de ellos fueron aplicadas las técnicas de componentes principales para reducir la dimensionalidad, el agrupamiento de manera jerárquica y la realización de gráficas de Andrews con el fin de identificar regiones hidrológicamente homogéneas en el Noroeste Argentino (NOA) a partir de los ríos de esa región. Asimismo, en el segundo estudio se analizó la cuenca del río Uruguay para así establecer las variables apropiadas a considerar para realizar un agrupamiento y definir regiones hidrológicas homogéneas en el mismo río. En el caso de la aplicación del algoritmo de agrupamiento jerárquico, tal como es explicado por (Tan y Kumar, 2014), se comienza con cada punto como un clúster único y luego se fusionan repetidamente los dos clústeres más cercanos hasta que quede un único clúster que abarque todo. Algunas de estas técnicas

tienen una interpretación natural en términos de agrupamiento basado en grafos, mientras que otras tienen una interpretación en términos de un enfoque basado en prototipos. Por otro lado, las gráficas de Andrews utilizadas por las autoras fueron propuestas por el mismo George Andrews en 1972 (Andrews, 1972). Las mismas son curvas que permiten una forma de visualización de datos multidimensionales. Este tipo de gráficos representa cada observación como una curva en un espacio multidimensional, donde cada dimensión corresponde a una variable en los datos. Las curvas se generan a partir de funciones trigonométricas que dependen de los valores de las variables. Estas curvas se superponen en un solo gráfico, lo que permite visualizar patrones y relaciones entre las observaciones de manera intuitiva. Las gráficas de Andrews son útiles para visualizar la estructura de los datos multivariados, identificar agrupaciones o patrones, y detectar posibles anomalías o valores atípicos.

Finalmente, algunos autores han efectuado análisis predictivos vinculados con el comportamiento de cuencas ubicadas en el territorio de la República Argentina. Entre ellos se encuentra el de (Lauro et al., 2021) cuyo objetivo es el de realizar un estudio de regionalización de las cuencas pertenecientes al sistema hidrográfico del Río Colorado con el objetivo de obtener medidas de caudal máximo anual de los ríos. Para ello, los autores realizaron un modelo de regresión simple luego de un análisis de correlación. Los mejores resultados se encontraron al utilizar variables relacionadas con el área y el perímetro de la cuenca.

Otro de los estudios vinculados con la predicción de caudales hidrológicos es el de (Korsic et al., 2023) mediante el cual los autores analizan el rendimiento de la técnica de Regresión de Vectores de Soporte (SVR) en la predicción de caudales mensuales con un adelanto de un mes en la cuenca del río Tupungato en los Andes Centrales de Argentina. En este caso, las variables utilizadas como entrada del modelo por los autores han sido los datos meteorológicos y estimaciones del área de cobertura de nieve.

Datos

Los datos utilizados en el presente estudio corresponden a la información histórica del caudal medio diario de los tres ríos de interés: Paraná, Uruguay y Limay.

Estos datos fueron obtenidos del Sistema Nacional de Información Hídrica (SNIH), bajo

la jurisdicción de la Secretaría de Infraestructura y Política Hídrica de la Nación, y de la Autoridad Interjurisdiccional de las Cuencas de los ríos Limay, Neuquén y Negro (AIC). El conjunto de datos contiene registros históricos diarios de caudal promedio medidos en m^3/s desde el 31 de marzo de 1921 hasta el 1 de enero de 2021.

En el presente trabajo de tesis se utiliza el término crónica hidrológica o simplemente crónica para hacer referencia a un período específico de tiempo en el que fueron recopilados y registrados dichos datos a lo largo de diferentes estaciones del año. En este caso, cada crónica tiene una duración de 12 meses.

Asimismo, los datos fueron sometidos a un control de calidad que involucró la identificación de datos erróneos o atípicos y la aplicación de técnicas apropiadas de tratamiento de datos faltantes.

Objetivos

El objetivo general de la siguiente tesis es obtener patrones temporales de las series de caudal en cada uno de los ríos a estudiar junto con evaluar la interrelación entre los mismos. Además, se busca desarrollar un modelo probabilístico que posibilite la predicción de las características futuras del año hidrológico en curso, basándose en el comportamiento de los ríos involucrados y en la influencia de oscilaciones climáticas como el ENSO.

A partir del objetivo general, se desprenden los siguientes objetivos específicos de la tesis:

- Realizar una síntesis descriptiva de las series de caudal a lo largo de series centenarias.
- Aplicar técnicas de agrupamiento temporal dentro de cada río de interés. Analizar los agrupamientos obtenidos entre las cuencas.
- Desarrollar un modelo predictivo que habilite la clasificación de crónicas en grupos específicos basados en los datos de los primeros meses de cada año.
- Evaluar la conexión entre los patrones temporales en las cuencas y los fenómenos de variabilidad climática, como el ENSO, para determinar su potencial contribución en la mejora de los pronósticos.

El presente trabajo de tesis está organizado en cuatro capítulos. El primero de ellos

está dedicado a la exploración y caracterización de las series de tiempo de caudal correspondientes a los ríos bajo estudio. En el segundo capítulo se aborda la tarea de agrupar las crónicas hidrológicas de estos ríos en función de sus características compartidas. El tercer capítulo se enfoca en la exploración de diversos modelos predictivos con el objetivo de pronosticar el comportamiento futuro del caudal de los ríos, incorporando el análisis de oscilaciones climáticas. Finalmente, el trabajo culmina con un capítulo conclusivo, en el cual se presentan las consideraciones generales y resultados obtenidos.

1. CARACTERIZACIÓN DE LOS RÍOS PARANÁ, URUGUAY Y LIMAY

1.1. Resumen

En el presente capítulo se ha realizado un análisis exploratorio de los datos provenientes de las series temporales de las estaciones de medición localizadas en las cuencas de los ríos Paraná, Uruguay y Limay.

Con el objetivo de conocer sus características hidrológicas, luego de realizado un tratamiento de los datos, se obtuvieron medidas estadísticas que permitieron caracterizar el comportamiento del caudal de cada uno de los ríos en cuestión.

1.2. Metodología

A continuación se explican las técnicas aplicadas sobre los datos que sirvieron para caracterizarlos.

1.2.1. Análisis de correlación

Con el objetivo de medir la relación entre las variables numéricas creadas para el análisis, se calculó la asociación entre las mismas utilizando el cálculo de correlación de Pearson. Esencialmente, el mismo evalúa si existe una relación lineal entre estas variables y hasta qué punto las mismas están asociadas (Weisstein, 2006).

La fórmula para su cálculo se presenta a continuación:

$$r = \frac{\sum((X - \bar{X})(Y - \bar{Y}))}{\sqrt{\sum(X - \bar{X})^2 \cdot \sum(Y - \bar{Y})^2}}$$

Donde:

- X e Y son los valores individuales de las dos variables a comparar
- $\bar{X}$ e $\bar{Y}$ son las medias de las variables X e Y, respectivamente
- $\sum$ indica la suma de los valores a lo largo de todas las observaciones

El coeficiente de correlación de Pearson varía entre -1 y 1. Un valor positivo cercano a 1 indica una asociación positiva fuerte, lo que significa que cuando una variable aumenta, la otra también tiende a aumentar. Por otro lado, un valor negativo cercano a -1 indica una correspondencia negativa fuerte donde, cuando una variable aumenta, la otra tiende a disminuir. Si el valor está cerca de 0, no hay una relación lineal significativa entre las variables.

1.2.2. Tratamiento de datos faltantes

En el marco de este trabajo, se aplicó el método de interpolación para abordar la tarea de imputación de los datos faltantes. La interpolación consiste en estimar valores desconocidos dentro de un conjunto de datos conocidos, utilizando una función que se ajusta suavemente a los puntos disponibles. Particularmente, en este caso fue realizada una interpolación basada en *splines* que son funciones suaves que se ajustan a los datos de manera que minimizan la curvatura, proporcionando una representación continua de la variable en cuestión (De Boor y De Boor, 1978). Para el análisis de los datos se empleó el software estadístico R (versión 4.1.1) y, en particular, se utilizó el paquete *zoo*, bajo la función *na.spline*, para realizar el tratamiento de datos faltantesZeileis et al. (2022).

Sin embargo, en aquellos casos en que los datos faltantes consecutivos fueran mayores a 60 días, se utilizaron otras fuentes de información tales como datos mensuales o semanales de promedio de caudal. Este proceso se llevó a cabo respetando las siguientes etapas:

- **Análisis de correlación de Pearson:** Se realizó un análisis de correlación de Pearson entre las bases de datos. Este análisis permitió evaluar la relación lineal entre las variables y fue realizada de acuerdo a la explicación mencionada en la Sección 1.2.1.

- **Promedio temporal:** Para facilitar la comparación y el análisis conjunto, se promediaron los datos en función de su escala temporal. En el caso de las bases de datos mensuales, se calculó el promedio mensual a partir de los datos originales correspondientes a cada mes. Para las bases de datos semanales, se determinó el promedio semanal. Este cálculo temporal garantizó la homogeneidad de la información.

- **Modelización con regresión lineal:** Se dividió la base de datos en dos conjuntos: el conjunto de entrenamiento, que comprendía el 70 % de los datos, y el conjunto

de testeo, que contenía el 30 % restante. Esta división se llevó a cabo de manera aleatoria. En el conjunto de entrenamiento se aplicó el modelo de regresión lineal utilizando la función *lm* de la librería *stats* de R R Core Team (2021a). Este modelo se desarrolló con el propósito de estimar las relaciones entre la base original con datos faltantes y las bases correlacionadas con ésta. Finalmente, se evaluó la eficacia del modelo sobre el conjunto de testeo.

- **Cálculo del R^2 Ajustado:** Una vez creado el modelo de regresión lineal, se calculó el coeficiente de determinación ajustado (R^2 ajustado). Este valor proporcionó una medida de la capacidad del modelo para explicar la variabilidad en los datos.

- **Evaluación con RMSE en el conjunto de testeo:** Para evaluar la capacidad predictiva del modelo, se calculó el Error Cuadrático Medio de la Raíz (RMSE) para medir la precisión de las predicciones en relación con los valores reales en el conjunto de testeo. Un RMSE bajo indicó un buen ajuste del modelo a los datos de prueba.

- **Estimación de datos faltantes:** Finalmente, el modelo de regresión lineal resultante se aplicó al conjunto de datos que permitieran estimar los valores ausentes.

1.2.3. Medidas de estadística básica y generación de nuevas variables

Adicionalmente, se han utilizado medidas descriptivas de posición tales como media, mediana y cuartiles así como también medidas de dispersión como varianza, para caracterizar los datos de caudal y crear nuevas variables que describieran cada crónica hidrológica. A esta última técnica se la denomina ingeniería de atributos (del inglés *feature engineering*).

1.2.4. Test de Chow

Para el caso en el que se observasen variaciones en la tendencia de las series temporales, fue aplicado el Test estadístico de Chow (Chow, 1960).

La Prueba de Chow compara dos modelos de regresión lineal estimados en diferentes períodos de tiempo o en diferentes subconjuntos de datos. Supongamos que tenemos un conjunto de datos que se puede dividir en dos grupos o períodos de tiempo: T_1 y T_2. La regresión lineal para cada grupo se puede expresar como:

Para el primer período (T_1):

$$Y_{T_1} = \beta_{0T_1} + \beta_{1T_1} X_{T_1} + \epsilon_{T_1}$$

Para el segundo período (T_2):

$$Y_{T_2} = \beta_{0T_2} + \beta_{1T_2} X_{T_2} + \epsilon_{T_2}$$

La hipótesis nula de la Prueba de Chow es que los coeficientes β_{1T_1} y β_{1T_2} son iguales, lo que indica que no hay cambios estructurales en la relación entre X e Y entre los dos períodos de tiempo. La hipótesis alternativa es que los coeficientes son diferentes, lo que sugiere un cambio estructural.

El estadístico de la Prueba de Chow se calcula como:

$$F = \frac{(SSE_R - (SSE_{T_1} + SSE_{T_2}))/k}{(SSE_{T_1} + SSE_{T_2})/(n_T - 2k)}$$

Donde:

- SSE_R es el error cuadrático residual de la regresión combinada,
- SSE_{T_1} y SSE_{T_2} son los errores cuadráticos residuales de las regresiones individuales para T_1 y T_2 respectivamente,
- k es el número de coeficientes en cada regresión (incluyendo el término de intersección),
- n_T es el número total de observaciones en ambos períodos de tiempo.

El estadístico F sigue una distribución F con k y $n_T - 2k$ grados de libertad bajo la hipótesis nula. Por consiguiente, un valor alto de F con un p valor bajo indicaría que los coeficientes son significativamente diferentes entre los dos períodos de tiempo, lo que sugiere un cambio estructural en la relación entre las variables explicativas y la variable de respuesta.

1.2.5. Estandarización

En el proceso de análisis de datos, fue empleada la técnica de estandarización con el objetivo de homogeneizar las diferentes variables que componen el conjunto de datos.

La estandarización implica transformar las variables originales en una escala común, donde la media de cada variable es igual a cero y el desvío estándar es igual a uno. Esta transformación es esencial para eliminar las disparidades inherentes a las unidades de medida y las magnitudes de las variables.

Al estandarizar los datos, se logra que todas las variables contribuyan equitativamente al análisis. En consecuencia, esta técnica promueve la robustez y fiabilidad de los resultados al minimizar el efecto de las diferencias en escalas.

La fórmula de estandarización se presenta a continuación:

$$z = \frac{x - \mu}{\sigma} \tag{1.1}$$

Donde:

- z representa el valor estandarizado
- x es el valor original de la variable
- μ es la media de la variable
- σ es la desviación estándar de la variable

1.3. Cuenca del Plata

La cuenca hidrográfica del Plata se encuentra ubicada entre 14°– 37° S y 43°–67° O, en el sur de Sudamérica y es drenada por el Río de la Plata y sus afluentes. La misma cubre una superficie de 3,6 millones de kilómetros cuadrados extendiéndose por el territorio de los siguientes países: Brasil, Uruguay, Bolivia, Paraguay y Argentina. Al cubrir una extensión tan amplia de territorio, en la Cuenca del Plata se pueden identificar múltiples regímenes climáticos, de acuerdo a las regiones desarrolladas por (Barros et al., 2006).

En cuanto a sus características, la misma está compuesta por cuatro subcuencas principales que son las de los ríos Uruguay, Paraná, Paraguay y la del Río de la Plata. Debido a su vasta extensión y al ser una importante fuente de agua dulce, los ríos de la Cuenca

del Plata son aprovechados por los países mencionados para sus actividades económicas y socioculturales.

En el presente trabajo se hace énfasis en el estudio de los ríos Paraná y Uruguay los cuales se describen y estudian en las siguientes subsecciones.

1.3.1. Río Paraná

1.3.1.1. Datos

Para el caso del río Paraná, los datos con los que se trabajó para el análisis provienen de los registros de la estación número 3805, Corrientes, ubicada en la latitud 27° 27' 35" y longitud 58° 49' 60" que se puede observar en el mapa de la Figura 1.1, en la provincia homónima, obtenidos del Sistema Nacional de Información Hídrica (SNIH) a través de su portal web en el siguiente url: `https://snih.hidricosargentina.gob.ar/Inicio.aspx`. En cuanto a su estructura, se trata de registros históricos medios diarios de caudal, medidos en m^3/s desde el día 1° de enero de 1904 hasta el día 31 de agosto de 2021. Por lo tanto, esta base contiene más de 42.000 registros con dos variables: la fecha en la que se tomó el registro y el caudal promedio de ese día. Sin embargo, se han utilizado los datos de los años calendario completos por lo que no se utilizó la información del año 2021.

Una vez obtenidos los datos, se realizó un análisis exploratorio de los mismos con el objetivo de conocer su estructura y su distribución para poder así describirlos. Las principales características de los mismos se presentan a continuación.

1.3.1.2. Datos faltantes

En primer lugar, se estudiaron sus datos faltantes y, del total de días del período bajo análisis, los registros detallados en la Tabla 1.1 no contaban con el dato de caudal.

Año	Mes	Registros con datos faltantes	Fechas con datos faltantes
1984	Noviembre	7	17 al 22/11 y 30/11
1986	Enero	1	24/01
1987	Febrero	7	21-27/02
1988	Diciembre	2	17-18/12

Tabla 1.1: Cantidad de datos faltantes por año y mes de la base de datos utilizada, correspondiente al caudal medio diario del río Paraná medido en la estación número 3805, Corrientes.

En el marco de este trabajo, se aplicó el método de interpolación basada en *splines* logrando así una representación continua de la variable en cuestión.

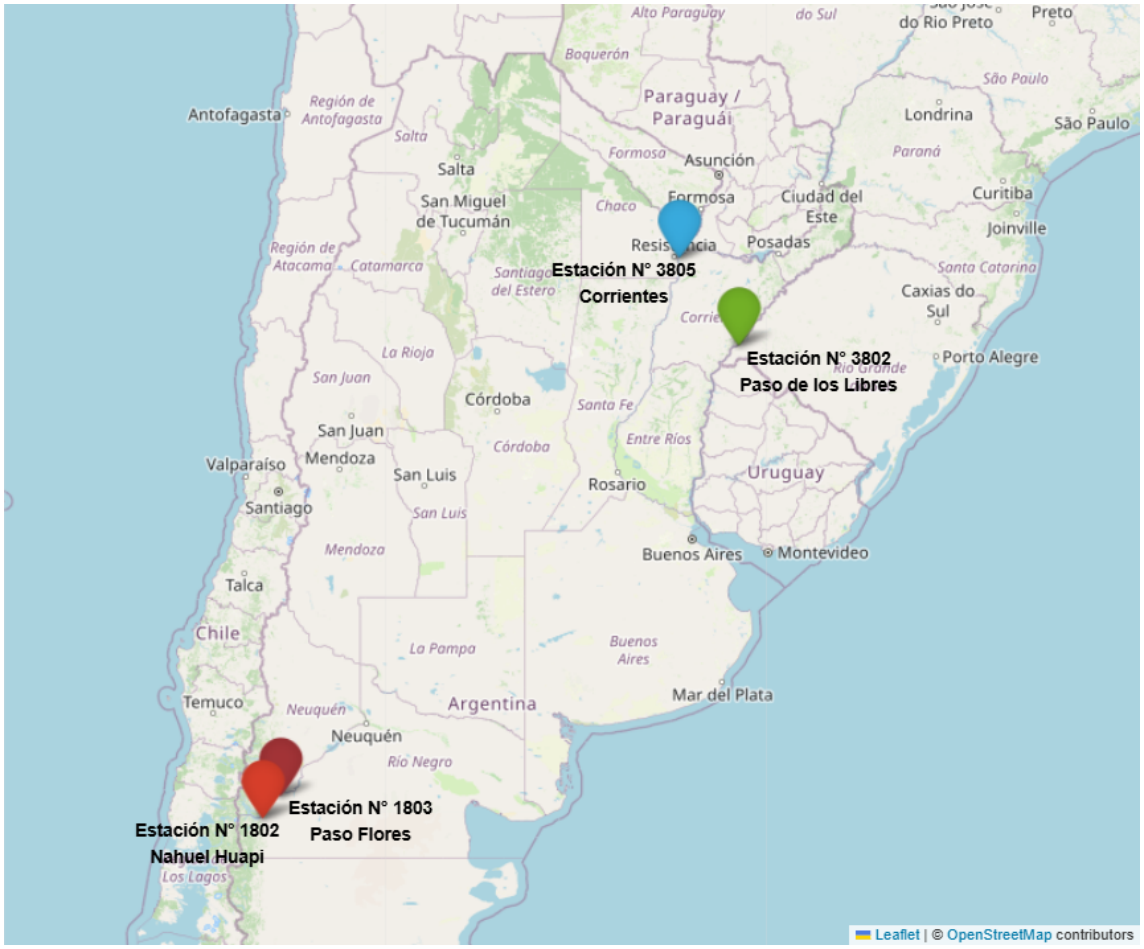

Figura 1.1: Mapa creado con la biblioteca Folium de Python que muestra las estaciones de medición señaladas en sus respectivas ubicaciones de la República Argentina. Los puntos corresponden a las siguientes estaciones: Estación N° 3805 (Corrientes), Estación N° 3802 (Paso de los Libres), Estación N° 1802 (Nahuel Huapi) y Estación N° 1803 (Paso Flores).

1.3.1.3. Medidas estadísticas descriptivas

Seguidamente, se calcularon medidas estadísticas descriptivas de los datos con el objetivo de caracterizarlos. En una primera instancia se obtuvo aquel registro que correspondiera con el valor máximo, mínimo, la media y la mediana de cada año calendario. Esta información se muestra en los gráficos de la Figura 1.2 donde se puede ver, por cada gráfico, las medidas por año a lo largo del período.

Al realizar un análisis de los gráficos se puede notar un cambio en la tendencia de los datos a partir de 1970. Desde esa década se observan valores más altos de las medidas mínimas y medias anuales.

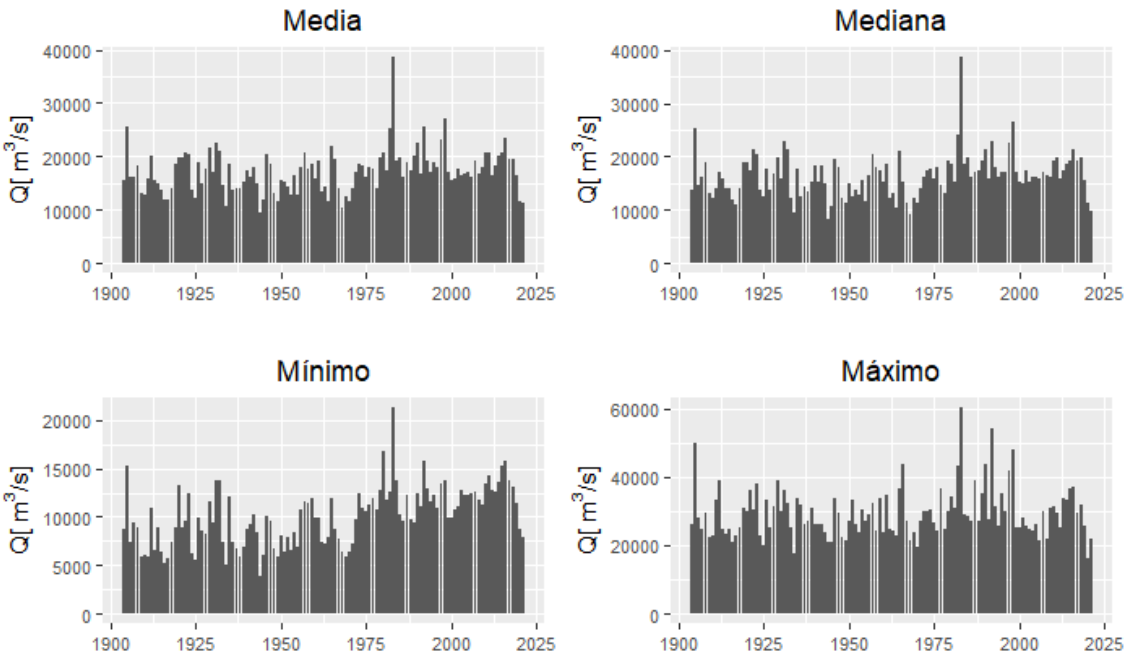

Figura 1.2: Gráficos de barras de medidas estadísticas descriptivas anuales de las crónicas hidrológicas del período 1904-2020 del caudal del río Paraná, medido en la estación 3805, Corrientes.

Esta observación se refuerza en la Figura 1.3 en la que se pueden ver las mismas cuatro medidas por cada uno de los años bajo análisis en el mismo gráfico. En esta Imagen se aprecia con mayor claridad que a partir de la década de 1970 existe un salto notorio en los valores mínimos de las crónicas hidrológicas. Más precisamente, el cambio de tendencia se hace notorio entre los años 1974 y 1975. Dicha observación se encuentra en concordancia con la bibliografía consultada (Meis, 2019) de acuerdo a la cual se destacan cambios abruptos en el valor medio del caudal durante la década del setenta a partir de la que se observa un incremento en la onda media anual entre el 20 % y el 22 % para la estación de Corrientes.

Por tal motivo, se procedió a realizar un análisis de la serie temporal de los mínimos así como también de la mediana, con el objetivo de investigar la presencia de un punto de quiebre estructural. Los resultados de la prueba de Chow revelaron hallazgos significativos tanto para la serie de mínimos como para la serie de medianas.

Para la serie de mínimos, el p valor obtenido fue de $p = 2,55 \times 10^{-4}$, mientras que para la serie de medianas, el p valor fue de $p = 1,12 \times 10^{-3}$. Ambos valores p resultaron

ser significativamente menores que el nivel de significancia de 0.05.

En consecuencia, fue rechazada la hipótesis nula de que no hay un cambio estructural en la serie temporal, lo que sugiere que hay suficiente evidencia para afirmar que existe un punto de quiebre significativo en los datos. Estos resultados respaldan la existencia de un cambio estructural en la serie temporal analizada en el año 1974 que fue el año con el cual se realizaron los tests.

A su vez, de todas las medidas obtenidas, los valores máximos son los que se encuentran dentro de un rango más amplio ya que se encuentran entre 16.333 y 60.215 m^3/s, valores obtenidos de los años 2020 y 1983, respectivamente. Este último dato extremo de 1983 se corresponde con el año de la inundación más severa del siglo XX en la sección argentina del río Paraná durante el fuerte evento de El Niño (Camilloni y Barros, 2000).

Por el contrario, los valores mínimos se encuentran dentro de un rango menor al de los máximos y fueron observados en el año 1944 con 3.946 y el año 1983 con 21.281 m^3/s. El año 1944 corresponde, según la bibliografía consultada (Blanco, 2022), a uno de los tres años severamente secos del río Paraná junto con los años 1934 y 1968.

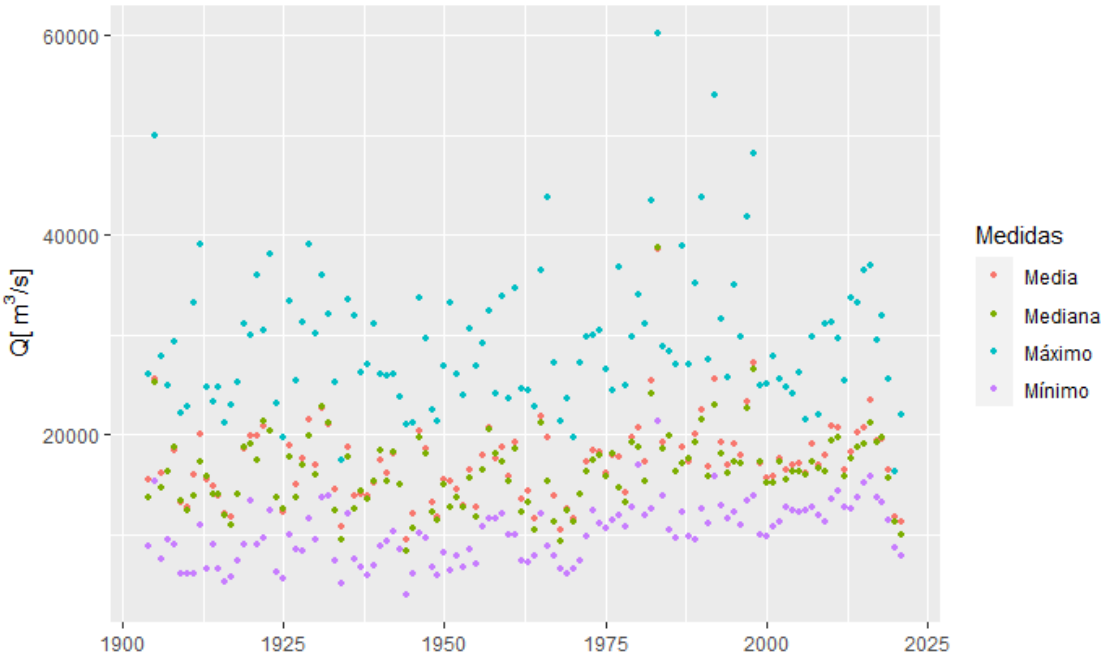

Figura 1.3: Gráfico de medidas estadísticas descriptivas anuales de las crónicas hidrológicas del período 1904-2020 del caudal del río Paraná, medido en la estación 3805, Corrientes.

A continuación, con el objetivo de caracterizar el ciclo anual que tiene el caudal del

río, se tomaron las mismas medidas mencionadas anteriormente pero de cada uno de los días del año. Este cálculo concluyó en la Figura 1.4 en la cual se puede observar que el río Paraná se caracteriza por contar con una estación húmeda que comienza en los últimos meses del año y que se acentúa en los meses de febrero y marzo del año siguiente para finalizar en mayo. Asimismo, la estación más seca se ubica en el mes de septiembre.

En relación a los valores mínimos, medios y medianos, se puede observar que los mismos presentan un comportamiento suavizado siguiendo una tendencia sin sobresaltos. Sin embargo, no ocurre lo mismo en el caso de los valores máximos.

Con respecto a estos últimos, se puede observar que los casos más extremos han ocurrido entre los meses de junio y agosto, meses que no coinciden con los de la época húmeda. Esto se debe a que estos valores máximos están relacionados con influencias climáticas tales como la del Niño en la cual suceden eventos que generan más precipitación y su consecuente mayor caudal (Berbery y Barros, 2002).

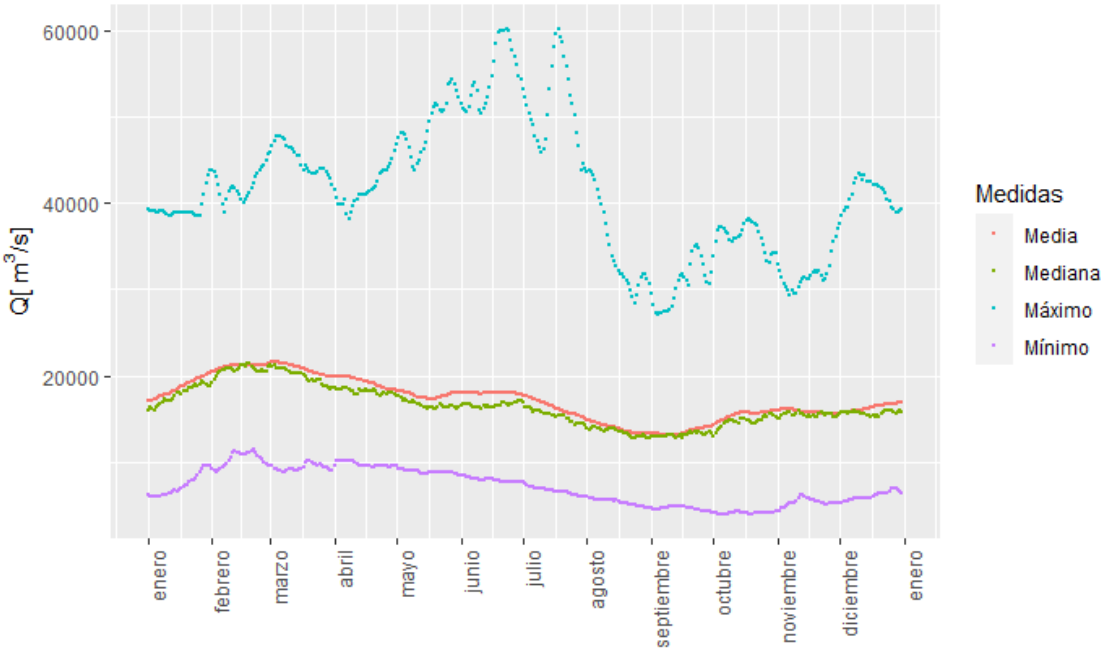

Figura 1.4: Gráfico acumulado de medidas estadísticas descriptivas mensuales de las crónicas hidrológicas del período 1904-2020 del caudal del río Paraná, medido en la estación 3805, Corrientes.

1.3.1.4. *Feature engineering*

Una vez observado el comportamiento de los datos, se procedió a utilizar la técnica de *feature engineering* con el objetivo de descubrir y extraer ciertas características de la base de datos sin procesar para lo cual se aprovechó el conocimiento del dominio anteriormente descripto. Como resultado, surgieron nuevas variables que resumieron o extrajeron nuevas propiedades de la información con la que se trabajó.

Previo a la descripción de las nuevas variables, es relevante aclarar que cada crónica fue analizada en un intervalo temporal que abarca desde diciembre de un año calendario hasta noviembre del año siguiente. Como fue mencionado en la sección previa, ello se debe al comienzo de la estación húmeda a fines del año calendario anterior. En este sentido, cada período comprendido entre diciembre de un año y noviembre del año siguiente, al cual se le denomina crónica, fue considerado como un registro.

Al observarse un comportamiento cíclico, se decidió crear doce variables nuevas que fueran el cálculo de la media mensual de caudal. Seguidamente, se identificaron los trimestres del año, comenzando desde el mes de diciembre, y, en base a estos, se crearon variables que representaran la media, la mediana, el máximo y el mínimo de caudal de cada uno de ellos. Finalmente, se obtuvieron los valores mínimos, máximos, la mediana y la media de cada una de las crónicas en su totalidad. Por lo tanto, en resumen, un nuevo dataset fue creado con 117 observaciones, una por cada año, y con las siguientes variables:

- Un valor promedio por cada mes del año (12 variables):
  - *promedio 1*: promedio de caudal del mes de enero
  - *promedio 2*: promedio de caudal del mes de febrero
  - *promedio 3*: promedio de caudal del mes de marzo
  - *promedio 4*: promedio de caudal del mes de abril
  - *promedio 5*: promedio de caudal del mes de mayo
  - *promedio 6*: promedio de caudal del mes de junio
  - *promedio 7*: promedio de caudal del mes de julio
  - *promedio 8*: promedio de caudal del mes de agosto
  - *promedio 9*: promedio de caudal del mes de septiembre

  - *promedio 10*: promedio de caudal del mes de octubre
  - *promedio 11*: promedio de caudal del mes de noviembre
  - *promedio 12*: promedio de caudal del mes de diciembre

- Un valor promedio por cada trimestre del año (4 variables):
  - *promedio 1Trim*: promedio de caudal del primer trimestre
  - *promedio 2Trim*: promedio de caudal del segundo trimestre
  - *promedio 3Trim*: promedio de caudal del tercer trimestre
  - *promedio 4Trim*: promedio de caudal del cuarto trimestre

- La mediana de cada trimestre del año (4 variables):
  - *mediana 1Trim*: mediana de caudal del primer trimestre
  - *mediana 2Trim*: mediana de caudal del segundo trimestre
  - *mediana 3Trim*: mediana de caudal del tercer trimestre
  - *mediana 4Trim*: mediana de caudal del cuarto trimestre

- El máximo de cada trimestre del año (4 variables):
  - *max 1Trim*: máximo caudal del primer trimestre
  - *max 2Trim*: máximo caudal del segundo trimestre
  - *max 3Trim*: máximo caudal del tercer trimestre
  - *max 4Trim*: máximo caudal del cuarto trimestre

- El mínimo de cada trimestre del año (4 variables):
  - *min 1Trim*: mínimo caudal del primer trimestre
  - *min 2Trim*: mínimo caudal del segundo trimestre
  - *min 3Trim*: mínimo caudal del tercer trimestre
  - *min 4Trim*: mínimo caudal del cuarto trimestre

- La media, mediana, máximo y mínimo de cada crónica (4 variables):
  - *promedio*: caudal medio de la crónica

- *mediana*: mediana de caudal de la crónica
- *max*: máximo caudal de la crónica
- *min*: mínimo caudal de la crónica

En resumen, un total de 32 variables que describen a cada una de las crónicas hidrológicas.

1.3.1.5. Análisis de correlación

Seguidamente, fue realizado un análisis de correlación entre las 32 variables para analizar la relación recíproca entre las variables creadas. Previo a su realización se podía suponer que existiría una alta correlación entre al menos algunas de ellas ya que podía tratarse de las mismas medidas pero con una diferente granularidad como por el ejemplo sería el caso del promedio del mes de enero y del primer trimestre del año. Así fue que se creó un gráfico de correlación que puede ser observado en la Figura 1.5.

Este gráfico muestra, a través de una matriz, la correlación entre las diferentes variables. En ella se observa que, cuanto más alto es el valor de correlación entre dos *features*, más oscuro es el punto en la intersección. De esta manera, se pueden ver áreas donde se identifican grupos de variables más correlacionadas entre sí y grupos donde sucede lo contrario.

Algunas de las conclusiones que se pueden obtener del gráfico se describen a continuación. En primer lugar, la correlación es muy baja entre las medidas del primer trimestre y las del tercero. Eso quiere decir que se espera que el comportamiento de uno sea diferente al del otro. Esto no significa que tiene una correlación negativa porque, en ese caso, luego de una temporada muy húmeda se podría esperar una época de sequía de similar intensidad sino que no existe una relación recíproca entre esas épocas. Lo mismo sucede entre el primero y el cuarto trimestre.

En contraste, se observa una significativa correlación entre los valores correspondientes al segundo y tercer trimestre. Un aspecto notable de esta relación es que el segundo trimestre se caracteriza por ser una temporada húmeda, mientras que el tercero se considera seco. No obstante, esta marcada correlación podría proporcionar la capacidad de determinar la magnitud de la estación seca, dado que está correlacionada con la temporada húmeda del trimestre anterior.

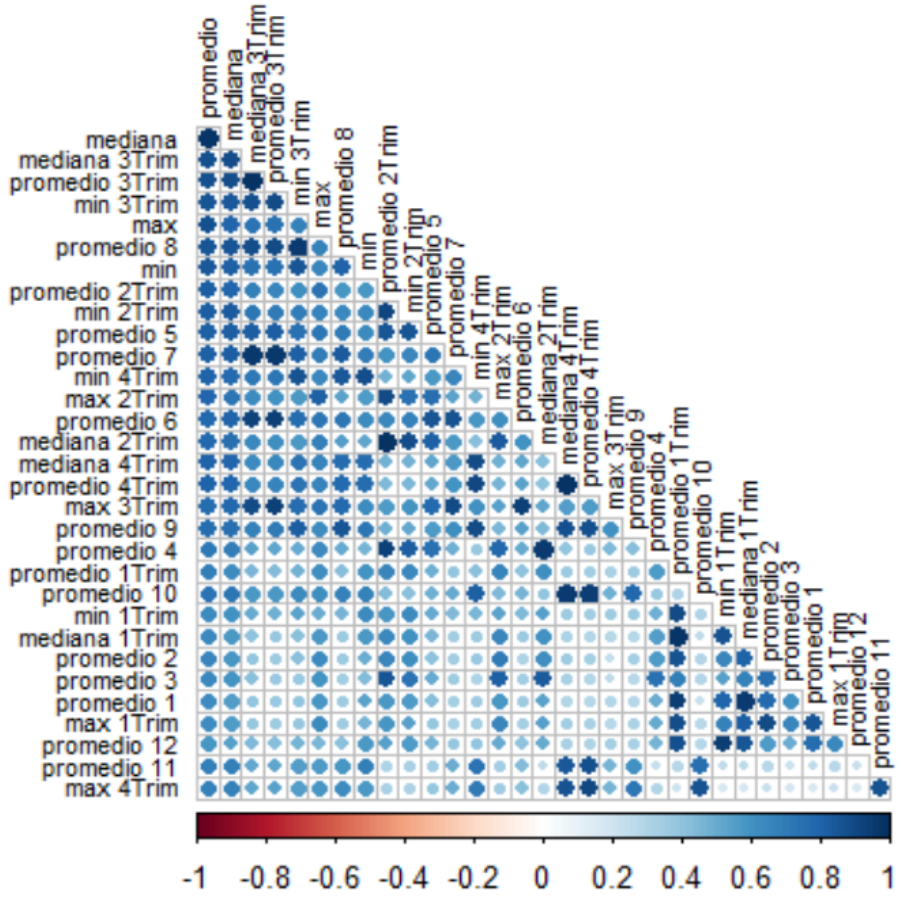

Figura 1.5: Matriz de correlación de la base de datos a utilizar para el análisis que incluye variables creadas para tal fin en base a los valores diarios de caudal del río Paraná.

Por otro lado, se observa una fuerte correlación entre los meses de un mismo trimestre. Ello confirma la correcta separación de temporadas en trimestres.

Finalmente, existen ciertas *features* que tienen un valor alto de correlación con todas las demás variables y que se podría concluir que resumen la totalidad de la crónica por su característica generalista como pueden ser la media y la mediana principalmente. A su vez, la variable que identifica los máximos está muy correlacionada con las *features* del segundo y tercer trimestre que es, como fuera explicado, la época donde se identifican estos valores máximos.

1.3.1.6. Estandarización

Con el objetivo de eliminar las diferencias en las unidades de medida y mitigar el impacto de variables con escalas muy diferentes, se prosiguió a estandarizar los datos.

En este caso, debido al cambio en la tendencia observado a partir de la década de 1970 que fue explicado en la Sección 1.3.1.3, la base de datos ha sido dividida en dos períodos para luego ser estandarizada. El primero de ellos se trata del período comprendido entre 1904 y 1969 y el segundo desde 1970 en adelante.

1.3.2. Río Uruguay

1.3.2.1. Datos

Los datos utilizados en el análisis del río Uruguay corresponden a los registros de la estación número 3802, Paso de los Libres, ubicada en la latitud 29° 43' 09" y longitud 57° 05' 02" que se puede observar en el mapa de la Figura 1.1, en la provincia de Corrientes, obtenidos del portal del Sistema Nacional de Información Hídrica (SNIH). Siguiendo la misma estructura que los datos utilizados para el estudio del río Paraná, en este caso se contó con información desde el día 7 de marzo de 1908 hasta el día 1° de enero de 2021 sumando, de esta manera, más de 40.000 registros. Tal como fue comentado, se han utilizado los datos de los años calendario completos por lo que no se utilizó la información de los años 1908 ni 2021.

En las secciones que se presentan a continuación se replica el análisis exploratorio de los datos tal como fue realizado para el río Paraná.

1.3.2.2. Datos faltantes

Al analizar los datos faltantes del período bajo análisis fueron identificados los registros que no contaban con el dato de caudal y que se muestran en la Tabla 1.2.

Debido a que en la base existen datos faltantes registrados que en ciertos casos corresponden a meses enteros, se procedió a completar los mismos con el dato de promedio mensual obtenido de la misma fuente y la misma estación como valor constante para todos los días de dicho mes. Tal fue el caso de:

- Enero a diciembre de 1920
- Julio de 1931
- Marzo y abril de 1932

Año	Registros con datos faltantes	Fechas con datos faltantes
1911	2	4/2 y 13/5
1913	2	2/6 y 25/6
1914	4	22 al 24/3 y 27/3
1916	1	31/12
1918	1	31/3
1919	1	31/1
1920	366	1/1 al 31/12
1930	5	15/3, 17/5, 31/5, 4/9 y 9/10
1931	31	1/7 al 31/7
1932	61	1/3 al 30/4
1933	2	1/5 y 31/12
1935	2	13 y 14/4
1943	4	1/2 al 4/2
1946	1	14/10
1947	1	23/11
1948	5	1/1, 4/1, 8/1, 9/1 y 12/9
1950	1	31/10
1951	38	1/2, 4/2, 5/2, 6/2, 7/2, 11/2, 17/2, 20/2, 23/2, 1/3, 2/3, 7/3, 11/3, 13/3, 16/3, 17/3, 23/3, 25/3, 27/3, 4/4, 7/4, 11/4, 12/4, 13/4, 14/4, 17/4, 9/10, 14/10, 15/10, 21/10, 6/11, 8/11, 15/11, 2/12, 23/12, 25/12, 30/12 y 31/12
1952	11	1/1, 31/1, 6/2, 9/2, 17/2, 24/2, 25/2, 15/3, 2/4, 27/7 y 31/7
1953	4	17/1, 19/1, 8/12, 20/12
1956	1	10/6
1962	4	7/4, 29/7, 30/7 y 31/7
1963	6	7/9 al 12/9
1966	61	1/11 al 31/12
1967	31	1/10 al 31/10
1969	80	4/5 al 7/5, 13/5 al 29/5, 11/6 al 13/6, 7/7 al 1/9 y del 1/12 al 4/12
1973	33	1/11 al 3/12
1974	12	19/4 al 26/4 y del 12/5 al 15/5
1975	5	7/1, 11/1, 12/1, 12/2 y 13/2
1977	9	22/3 al 26/3, 28/3, 25/12, 26/12 y 27/12
1978	5	4/3 al 7/3 y 7/12
1979	1	16/1
1980	3	23/2, 27/8 y 9/11
1981	1	1/9
1982	2	23/1 y 14/3
1991	28	1/2 al 28/2
2013	4	11/4 y 14 al 16/4

Tabla 1.2: Cantidad de datos faltantes por año y mes de la base de datos utilizada, correspondiente al caudal medio diario del río Uruguay medido en la estación número 3802, Paso de los Libres.

- Noviembre y diciembre de 1966

- Agosto de 1969

- Febrero de 1991

El único mes que satisfacía el requisito de ser un mes completo sin información y que, además, carecía del dato correspondiente al promedio mensual, fue octubre de 1967. En relación a este mes y a los restantes datos faltantes que también cumplían con la restricción de no exceder los 60 días consecutivos sin información, se implementó el método de interpolación basado en *splines*.

1.3.2.3. Medidas estadísticas descriptivas

Seguidamente, se calcularon medidas estadísticas descriptivas de los datos por cada año calendario. Esta información se muestra en los gráficos de la Figura 1.6 donde se pueden ver las medidas por año a lo largo del período. Asimismo, la Figura 1.7 muestra las mismas cuatro medidas por cada uno de los años bajo análisis pero en el mismo gráfico.

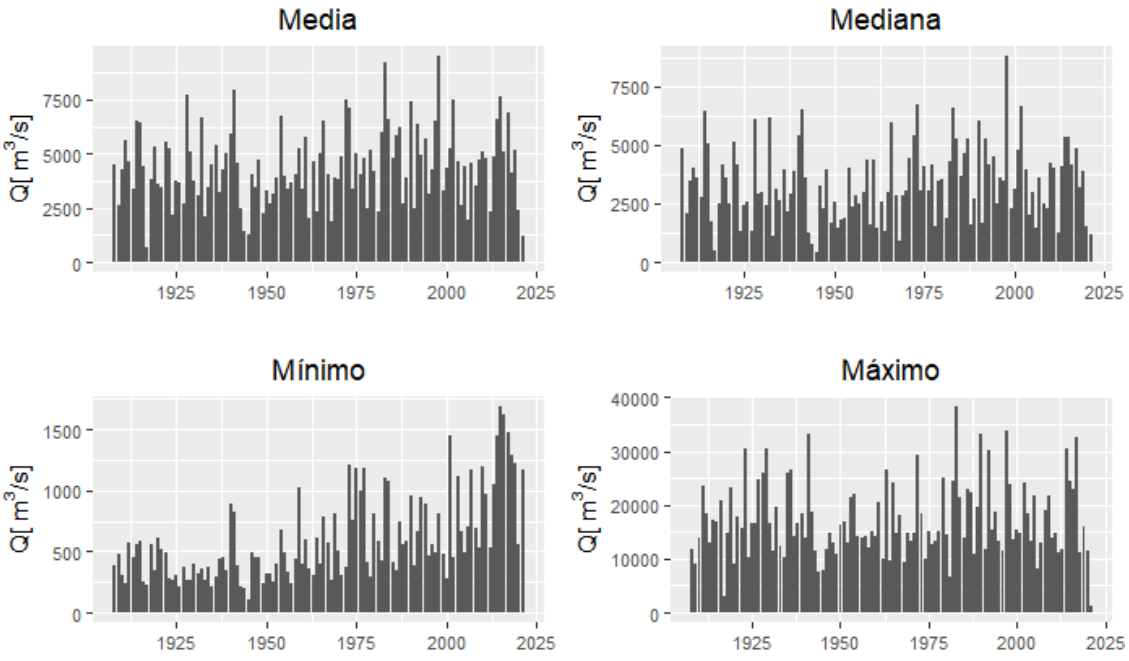

Figura 1.6: Gráficos de barras de medidas estadísticas descriptivas anuales de las crónicas hidrológicas del período 1909-2020 del caudal del Río Paraná, medido en la estación 3802, Paso de los Libres.

En ambas Figuras se puede observar que, al igual que en el caso del río Paraná, los valores máximos son los que se encuentran dentro de un rango más amplio ya que se

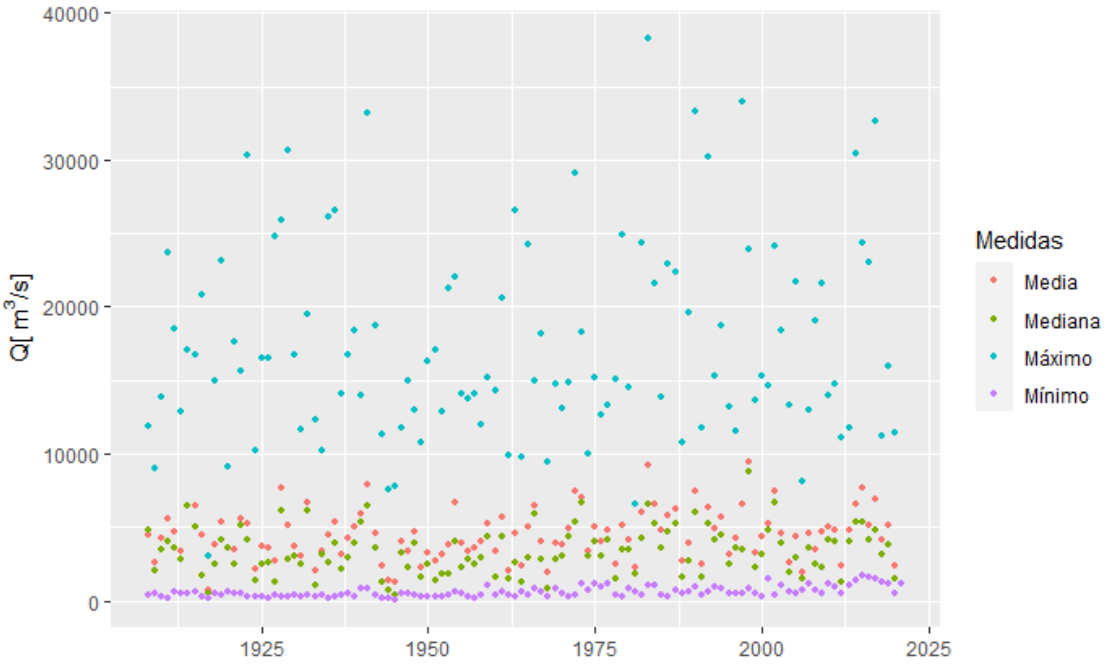

Figura 1.7: Gráfico de medidas estadísticas descriptivas anuales de las crónicas hidrológicas del período 1909-2020 del caudal del río Uruguay, medido en la estación 3802, Paso de los Libres.

encuentran entre 3.039 y 38.256 m^3/s, valores obtenidos de los años 1917 y 1983, respectivamente. Este último dato hace referencia al mismo evento extremo mencionado en la Sección 1.3.1.3.

Por el contrario, los valores mínimos se encuentran dentro de un rango menor al de los máximos y fueron observados en el año 1945 con 104 y el año 2015 con 1.684 m^3/s. El año 1945 pertenece al trienio que comenzó en 1944, también mencionado para el caso del río Paraná, en el cual ocurrió una importante sequía, tal como se menciona en la bibliografía consultada (Genta et al., 1998).

Adicionalmente, atendiendo a los valores mínimos de caudal a lo largo del período, se puede observar que para finales del mismo, más específicamente en la década comprendida entre 2010 y 2019, el 80 % de los valores fueron mayores a 1.000 m^3/s. En contraste, en décadas anteriores, los caudales mínimos tendían a ser inferiores a este umbral, a excepción de casos atípicos.

Por otro lado, el análisis de caracterización del ciclo anual del caudal del río Uruguay

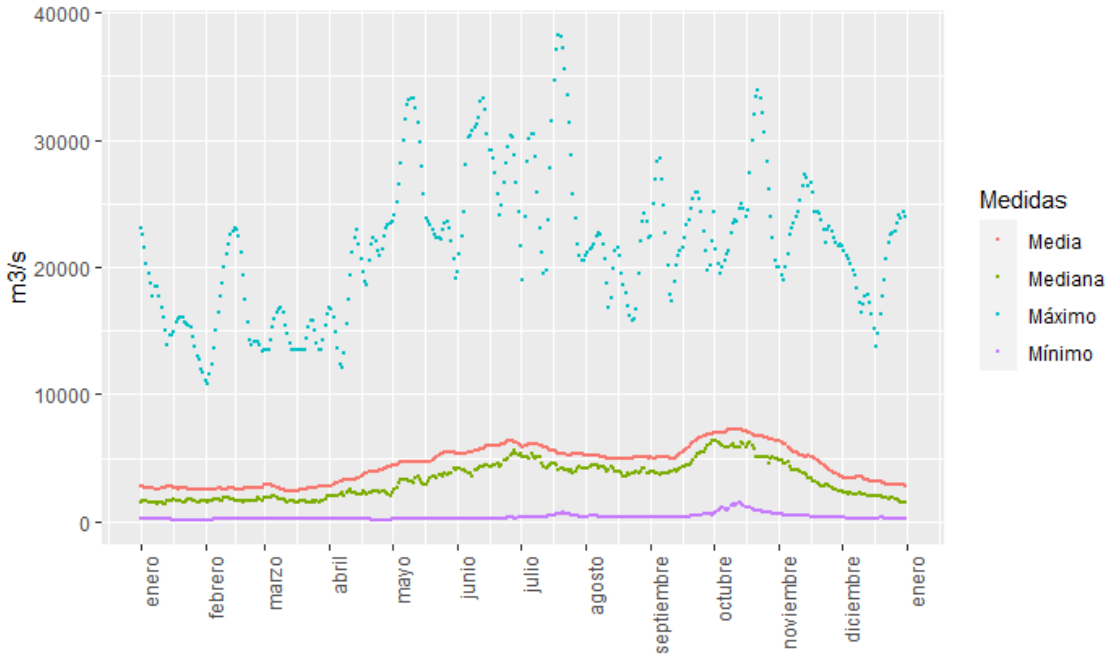

Figura 1.8: Gráfico acumulado de medidas estadísticas descriptivas mensuales de las crónicas hidrológicas del período 1904-2020 del caudal del río Uruguay, medido en la estación 3802, Paso de los Libres.

que se puede observar en la Figura 1.8, muestra que el río Uruguay se caracteriza por contar con una estación húmeda comprendida entre los meses de mayo y noviembre que coincide con el período más seco del río Paraná. Asimismo, la estación más seca se ubica en los meses de enero y febrero, también en oposición al río Paraná.

En relación al comportamiento de los valores mínimos, medios, medianos y máximos, ocurre lo mismo que se podía observar en el caso del río Paraná; es decir, un comportamiento suavizado en los primeros tres pero una mayor variabilidad en el caso de los máximos.

1.3.2.4. *Feature engineering*

El enfoque utilizado para el río Uruguay sigue la misma metodología detallada en la Sección 1.3.1.4 en lo que respecta a la generación de nuevas variables. Además, se adoptó la misma convención de que cada crónica abarca el período comprendido entre diciembre de un año calendario y noviembre del año siguiente, en concordancia con el inicio de la

estación seca.

1.3.2.5. Análisis de correlación

Algunas de las conclusiones que se pueden obtener del análisis de correlación que se observa gráficamente en la Figura 1.9, se describen a continuación.

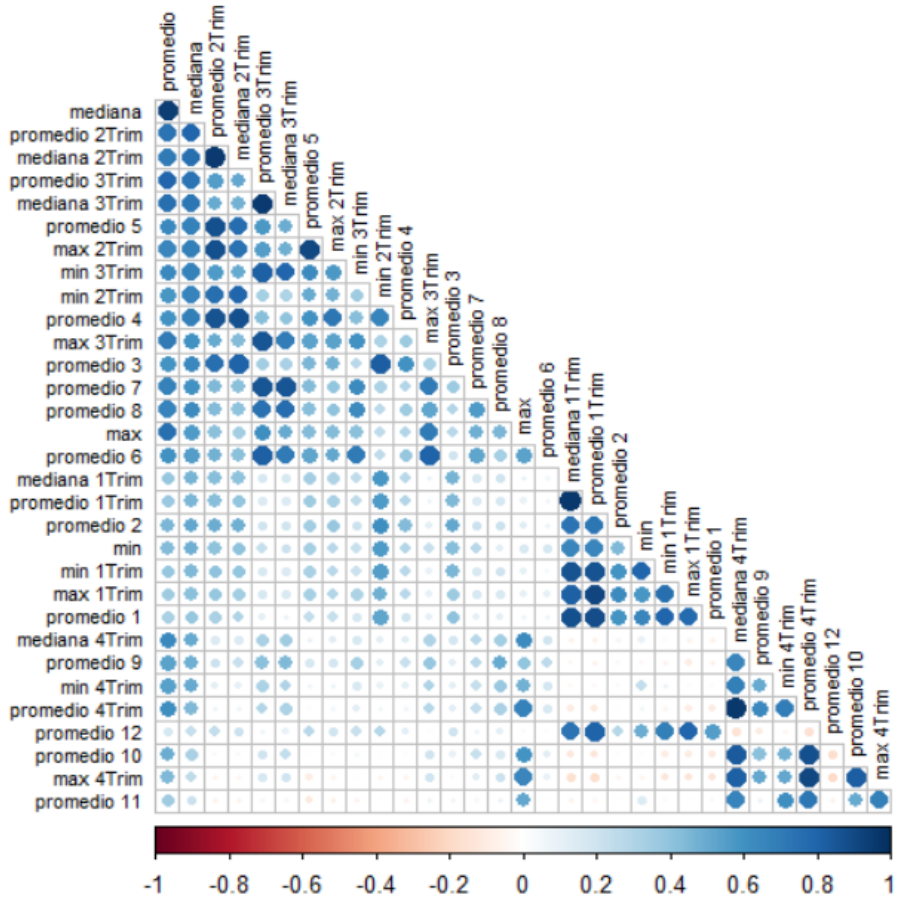

Figura 1.9: Matriz de correlación de la base de datos a utilizar para el análisis que incluye variables creadas para tal fin en base a los valores diarios de caudal del río Uruguay.

Por un lado, existe una alta correlación entre los valores máximos, la media y la mediana anual y los valores correspondientes al tercer y cuarto trimestre del año. Dichos trimestres son los que corresponden al período húmedo del río lo que implica que las medidas estadísticas elegidas para caracterizar cada crónica del río Uruguay se encuentran definidas en los meses húmedos del año y no en los secos.

En contraposición, la segunda conclusión que se obtiene del presente análisis de corre-

lación es que el valor mínimo del año obtiene un alto valor de correlación cuando se lo contrasta con los meses más secos del año.

Por último, se observa una alta correlación entre los meses de un mismo trimestre lo cual confirma la correcta separación en trimestres.

1.3.2.6. Estandarización

Con el objetivo de eliminar las diferencias en las unidades de medida y mitigar el impacto de variables con escalas muy diferentes, se prosiguió a estandarizar los datos. Este proceso siguió el mismo enfoque que se aplicó en el caso del río Paraná, dividiendo el período en dos segmentos y utilizando el año 1970 como punto de corte.

1.4. Cuenca de los ríos Limay, Neuquén y Negro

La cuenca de los ríos Limay, Neuquén y Negro se ubica en el norte de la región Patagónica y se encuentra completamente contenida en el territorio de Argentina. Esta cuenca abarca una extensión que incluye la mayor parte de la provincia de Neuquén y partes de las provincias de Río Negro y Buenos Aires. El río Neuquén drena una superficie de 30.000 km^2, mientras que el río Limay cubre una región de 56.000 km^2. Juntos, estos ríos forman el río Negro, que abarca una cuenca de 116.000 km^2.

En el presente trabajo se hizo énfasis en el estudio del río Limay el cual se encuentra ubicado entre 40° 20' S y 70° 10' O en la región montañosa de los Andes. Particularmente, según lo descripto por (Zabala et al., 2021), la cuenca del río Limay se distingue por ser exorreica y tener una vertiente hacia el Atlántico. La misma tiene una longitud de 430 km y un área de drenaje de aproximadamente 61,600 km^2. En cuanto a su régimen hidrológico natural, exhibe una doble crecida. La primera de ellas durante el invierno, cuando se producen las principales precipitaciones en la cuenca. Así es como las nevadas se acumulan hasta finales de la primavera, cuando el deshielo origina una segunda crecida. Los periodos de estiaje, comunes hacia finales del verano, persisten hasta el inicio de las lluvias otoñales.

En las secciones posteriores se presenta un análisis de los datos relativos a la cuenca

que permitirán respaldar y profundizar la reciente descripción de la cuenca.

1.4.1. Río Limay

1.4.1.1. Datos

Para el caso del río Limay, los datos con los que se trabajó para el análisis provienen de las siguientes fuentes:

- Datos diarios provistos por la Autoridad Interjurisdiccional de las Cuencas de los Ríos Limay, Neuquén y Negro (AIC) para la estación Nahuel Huapi, ubicada en la latitud 41°03'22"S y longitud 71°08'49"O en la provincia de Río Negro desde el día 1° de abril de 1922 hasta el día 30 de noviembre de 2022.

- Datos obtenidos de la base de datos del programa Visual MARGO, programa de optimización de mediano y largo plazo del mercado eléctrico mayorista desarrollado por CAMMESA (Compañía Administradora del Mercado Mayorista Eléctrico S.A.). Estos datos corresponden a la tabla QHYAP que, de acuerdo al Manual de Usuario del programa, se trata de los aportes históricos hidráulicos para cada central hidroeléctrica en m^3/s. En este caso se trata de los aportes del río Limay a la central Alicurá para el período comprendido entre enero de 1943 y diciembre de 2022. Dicha central presenta el primero de cinco diques que se encuentran sobre el mencionado río.

- Datos obtenidos del SNIH a través de su portal web para las siguientes estaciones:

  - Los registros de la estación número 1802, Nahuel Huapi, ubicada en la latitud 41°04'12"S y longitud 71°09'54"O en la provincia de Río Negro, tanto datos diarios como mensuales. Para el caso de los datos diarios se contó con información desde el día 1° de enero de 2006 hasta el día 31 de marzo de 2021 mientras que la información de los datos mensuales cubrió el período comprendido entre el 31 de marzo de 1921 y el 1° de noviembre de 2011.

  - Los registros mensuales de la estación número 1803, Paso Flores, ubicada en la

latitud 40°35'00"S y longitud 70°38'00" O en la provincia de Río Negro desde el día 1° de abril de 1941 hasta el día 1° de noviembre de 2011.

Tanto esta última estación de Paso Flores como la estación Nahuel Huapi que fue referida en los datos obtenidos de la AIC así como también del SNIH, se pueden observar en el mapa de la Figura 1.1.

Una vez obtenidos los datos, se realizó un análisis exploratorio de los mismos con el objetivo de conocer su estructura y su distribución para poder así describirlos. Las principales características de los mismos se presentan a continuación.

1.4.1.2. Datos faltantes

En esta oportunidad se han analizado los datos faltantes para cada una de las bases con las que se contaba de información.

Para el caso de la base diaria de AIC en la estación Nahuel Huapi, los datos faltantes se resumen en la Tabla 1.3 en la que se puede observar que existe un período de 17 años comprendido entre 1966 y 1982 sin información disponible. Sumado a esa fase, se observan lapsos en los años posteriores en los que tampoco se dispone de información.

Siguiendo el estudio de datos faltantes de las bases, la Tabla 1.4 muestra que para el caso de la base de Nahuel Huapi diaria de SNIH los períodos con datos faltantes son sólo dos que se dan en dos años diferentes y que no suman más de 15 días consecutivos cada uno. Debido a que los datos faltantes de esta base coincidían con los de la base diaria de AIC, no resultaron de utilidad para completar los períodos de ausencia de datos.

En cuanto a las bases de datos con información mensual, la Tabla 1.5 muestra los períodos con datos faltantes de la base de datos mensual de SNIH medido en la estación Nahuel Huapi. En algunas oportunidades se trata de períodos que cubren años consecutivos con falta de datos.

Asimismo, la Tabla 1.6 muestra los períodos con datos faltantes para el caso de la base de datos mensual de SNIH medido en la estación Paso Flores. Es relevante aclarar que los datos de esta base utilizados en el análisis fueron aquellos previos al año 1983 ya que ese año fue en el que se inció el llenado del embalse de Alicurá que, como ha sido mencionado,

Año	Período con datos faltantes
1928	1/2 al 29/2
1966	1/4 al 31/12
1967	1/1 al 31/12
1968	1/1 al 31/12
1969	1/1 al 31/12
1970	1/1 al 31/12
1971	1/1 al 31/12
1972	1/1 al 31/12
1973	1/1 al 31/12
1974	1/1 al 31/12
1975	1/1 al 31/12
1976	1/1 al 31/12
1977	1/1 al 31/12
1978	1/1 al 31/12
1979	1/1 al 31/12
1980	1/1 al 31/12
1981	1/1 al 31/12
1982	1/1 al 20/5
1992	1/8 al 31/12
1993	1/1 al 2/4, 30/7 al 15/8 y 31/8
1999	1/4
2000	30/11
2001	1/12 y 2/12
2004	11/6 al 15/6
2005	17/3
2015	9/4
2018	18/8 al 30/8
2021	23/3 al 29/3
2022	17/1 al 24/1, 5/7 al 22/8

Tabla 1.3: Datos faltantes por año de la base de datos utilizada, correspondiente al caudal medio diario del río Limay obtenido de la base de AIC medido en la estación Nahuel Huapi.

Año	Período con datos faltantes
2015	9/4
2018	19/8 al 30/8
2021	24/3 al 29/3

Tabla 1.4: Datos faltantes por año de la base de datos utilizada correspondiente al caudal medio diario del río Limay obtenido de la base de SNIH medido en la estación número 1802, Nahuel Huapi.

es el primero de los cinco diques del río Limay y que se encuentra previo a la estación de Paso Flores.

Por último, la base de datos de CAMMESA no presentó períodos con datos faltantes.

Debido a que los períodos con ausencia de información eran mayores a 60 días corridos, se procedió a tratarlos de acuerdo a la metodología desarrollada en la Sección 1.2.2 para dichos casos. Tras ese análisis, se obtuvieron los resultados que se muestran en la Tabla 1.7 que se presenta a continuación:

Año	Período con datos faltantes
1930	Diciembre
1931	Octubre
1932	Noviembre
1933	Noviembre
1934	Noviembre
1935	Noviembre
1936	Noviembre
1937	Noviembre
1938	Noviembre
1939	Noviembre
1940	Julio
1941	Octubre
1943	Octubre
1946	Octubre
1963	Diciembre
1964	Octubre
1965	Octubre
1966	Marzo a diciembre
1966	Marzo a diciembre
1967 a 1972	Año completo
1973	Enero a marzo
1976	Abril a junio
1977	Febrero y marzo
1979	Abril a diciembre
1980	Enero a marzo
1982 a 2009	Año completo
2010	Diciembre

Tabla 1.5: Datos faltantes por año de la base de datos utilizada correspondiente al caudal medio mensual del río Limay obtenido de la base de SNIH medido en la estación número 1802, Nahuel Huapi.

Año	Período con datos faltantes
1941	Octubre
1943	Octubre
1946	Octubre
1963	Diciembre
1964	Octubre
1965	Octubre
1966	Octubre
1967	Octubre
1968	Octubre
1969	Octubre
1970	Enero
1981	Noviembre
1982	Febrero a abril
1983	Noviembre y diciembre
1984 a 2009	Año completo
2010	Diciembre

Tabla 1.6: Datos faltantes por año de la base de datos utilizada correspondiente al caudal medio mensual del río Limay obtenido de la base de SNIH medido en la estación número 1803, Paso Flores.

Base	Periodicidad	Correlación	R^2 ajustado	RMSE
SNIH Paso Flores	Mensual	0,96	0,93	30,76
CAMMESA	Semanal	0,98	0,95	22,49

Tabla 1.7: Análisis de correlación y resultado del modelo de regresión lineal realizados entre la base de Nahuel Huapi de AIC y las bases de Paso de los Libres y de CAMMESA.

Con respecto al tratamiento de la base mensual de Nahuel de SNIH, debido a que ésta presentaba correlación perfecta al ser estudiada con la base de AIC mensualizada, no fue realizado el modelo de regresión lineal explicado sino que los datos fueron completados directamente con la información disponible.

Finalmente, como resultado de la aplicación de los modelos mencionados, se procedió a la imputación de datos faltantes en la base de datos de AIC. La secuencia de completado se determinó considerando varios criterios, tales como la periodicidad de las bases de datos, la correlación existente entre las variables y el error cuadrático medio de cada modelo, en ese orden de prioridad.

En primer lugar, se realizó la imputación de datos utilizando la base de datos de CAMMESA como fuente principal. Esta decisión se basó en el hecho de que la base de CAMMESA posee la menor periodicidad en comparación con las bases de Paso Flores y

Nahuel, además de presentar un RMSE menor en relación al modelo de Paso Flores.

A continuación, se procedió a completar la base de datos de AIC utilizando la información de la base de Nahuel mensual. Esta elección se debió a la correlación perfecta que existía entre la base de Nahuel y la base de AIC, lo que la convirtió en una fuente confiable para la imputación de datos faltantes.

Por último, se utilizó la base mensual de Paso Flores para llevar a cabo la imputación de datos en la base de AIC. Esta decisión se tomó considerando los criterios previamente mencionados y en función de su disponibilidad.

1.4.1.3. Medidas estadísticas descriptivas

A continuación, se realizaron cálculos para determinar medidas estadísticas descriptivas de los datos para cada año calendario. Estos resultados se representan en los gráficos de la Figura 1.10, donde se presenta un desglose de estas medidas a lo largo del período, año tras año. Del mismo modo, en la Figura 1.11 se muestran estas mismas cuatro medidas para cada uno de los años analizados, pero esta vez en un solo gráfico.

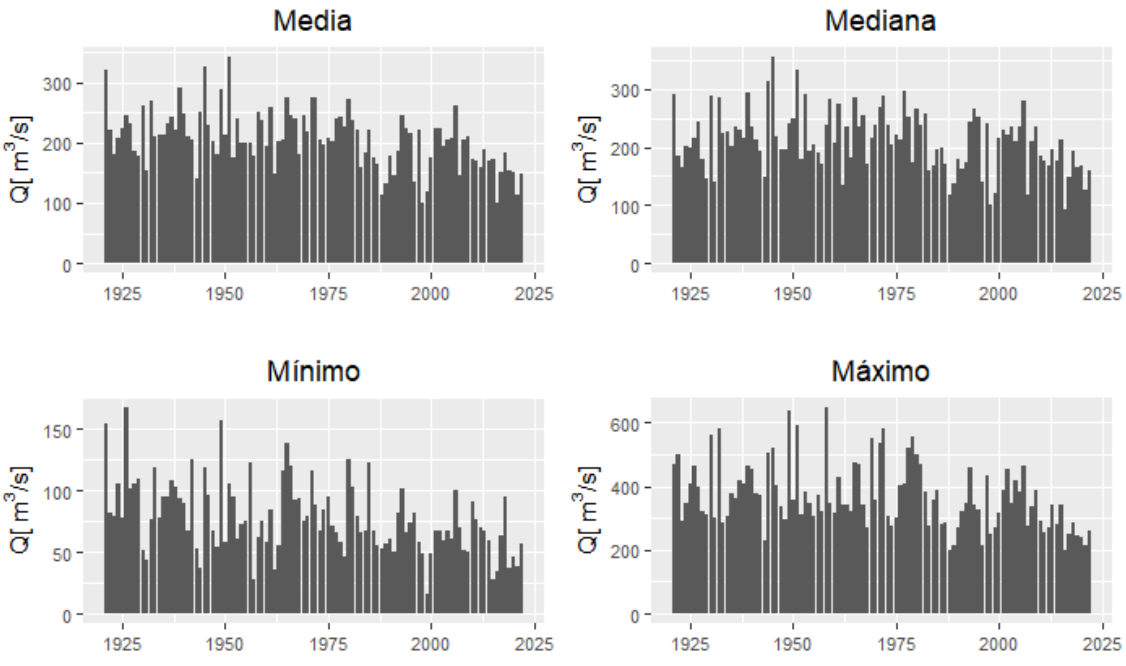

Figura 1.10: Gráficos de barras de medidas estadísticas descriptivas anuales de las crónicas hidrológicas del período 1922-2021 del caudal del río Limay.

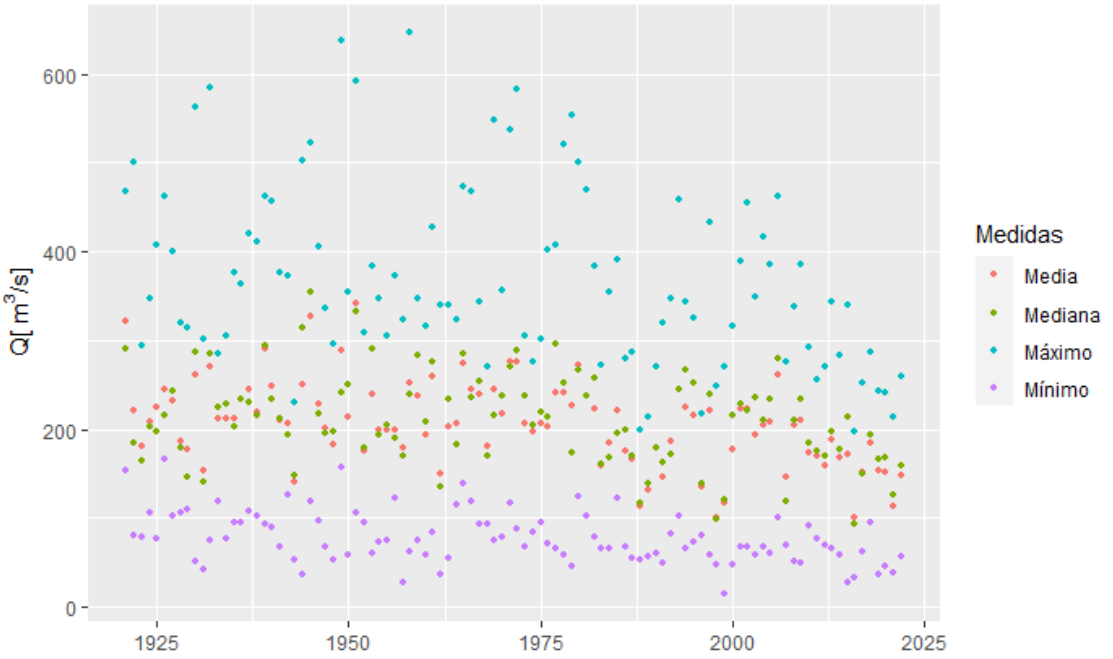

Figura 1.11: Gráfico de medidas estadísticas descriptivas anuales de las crónicas hidrológicas del período 1922-2021 del caudal del río Limay.

Al igual que en los casos anteriores, en la Figura 1.11 se observa que los valores máximos anuales del río Limay se encuentran dentro de un rango de amplitud más grande cuyo mínimo es 197,9 y su máximo es 646,8 m^3/s, valores obtenidos de los años 2016 y 1958, respectivamente. Asimismo, es importante resaltar que dicha amplitud disminuye en la medida que avanza el período.

Por el contrario, los valores mínimos se encuentran dentro de un rango menor al de los máximos y fueron observados en el año 1999 con 15,3 m^3/s y el año 1926 con 167 m^3/s.

Por otro lado, el análisis de caracterización del ciclo anual del caudal del río Limay que se puede observar en la Figura 1.12, muestra que el río Limay se caracteriza por contar con dos estaciones húmedas, una en el período invernal y otra cercana al mes de noviembre. Asimismo, la estación más seca se ubica en los meses de febrero y mayo.

De acuerdo a la bibliografía consultada (Seoane y López, 2007), estas observaciones se respaldan en el hecho de que el proceso de precipitación está fuertemente influenciado por las barreras montañosas aguas arriba y por la topografía montañosa local. Es por ello que se observan dos máximos de caudal en el año generados por dos procesos diferentes: la

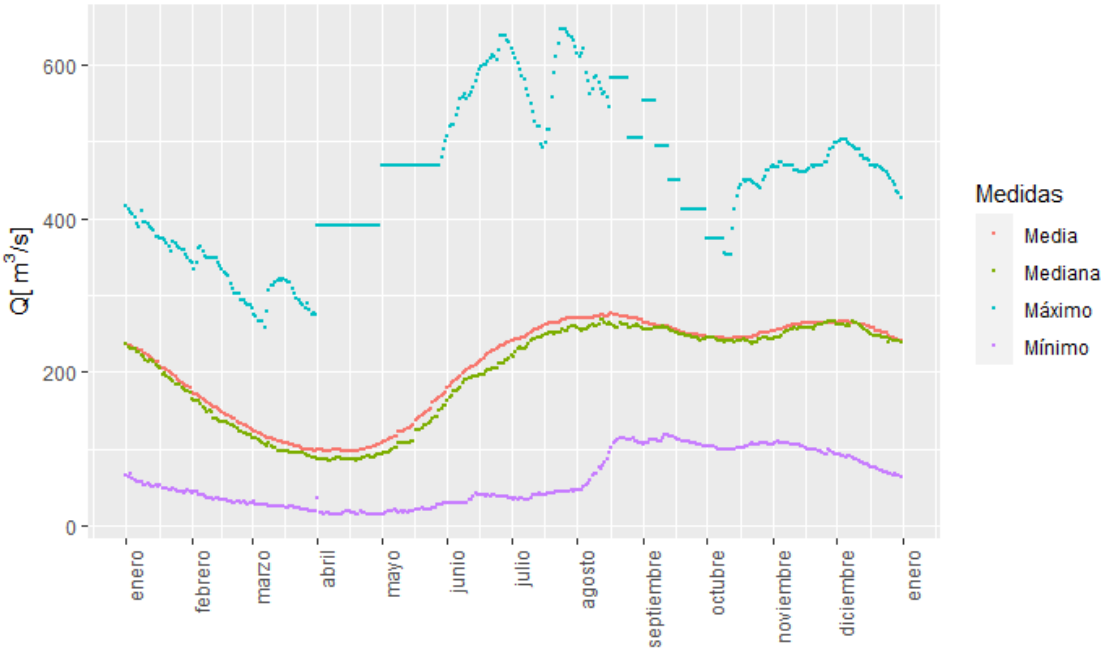

Figura 1.12: Gráfico acumulado de medidas estadísticas descriptivas mensuales de las crónicas hidrológicas del período 1922-2021 del caudal del río Limay.

precipitación en el período comprendido entre junio y agosto y el deshielo de nieve entre octubre y diciembre.

Por otro lado, se puede señalar que una vez más se repite el comportamiento de los valores máximos en los que se observa más variabilidad en el año, observando sus máximos en la época invernal.

Finalmente, es relevante destacar que en la Figura 1.12 se evidencian eventos en los cuales un valor se repite a lo largo de todo un mes o una semana. En estas situaciones, el valor máximo coincide con el promedio mensual o semanal de los meses en los que no se disponía de datos diarios, lo que resultó en que el análisis se viera obligado a utilizar los valores mensuales o semanales, según fuera el caso.

1.4.1.4. *Feature engineering*

El enfoque utilizado para el río Limay sigue la misma metodología detallada en la Sección 1.3.1.4 en lo que respecta a la generación de nuevas variables. Además, se adoptó la misma convención de que cada crónica abarca el período comprendido entre diciembre

de un año calendario y noviembre del año siguiente, en concordancia con el inicio de la estación seca.

1.4.1.5. Análisis de correlación

Algunas de las conclusiones que se pueden obtener del análisis de correlación que se observa gráficamente en la Figura 1.13, se describen a continuación.

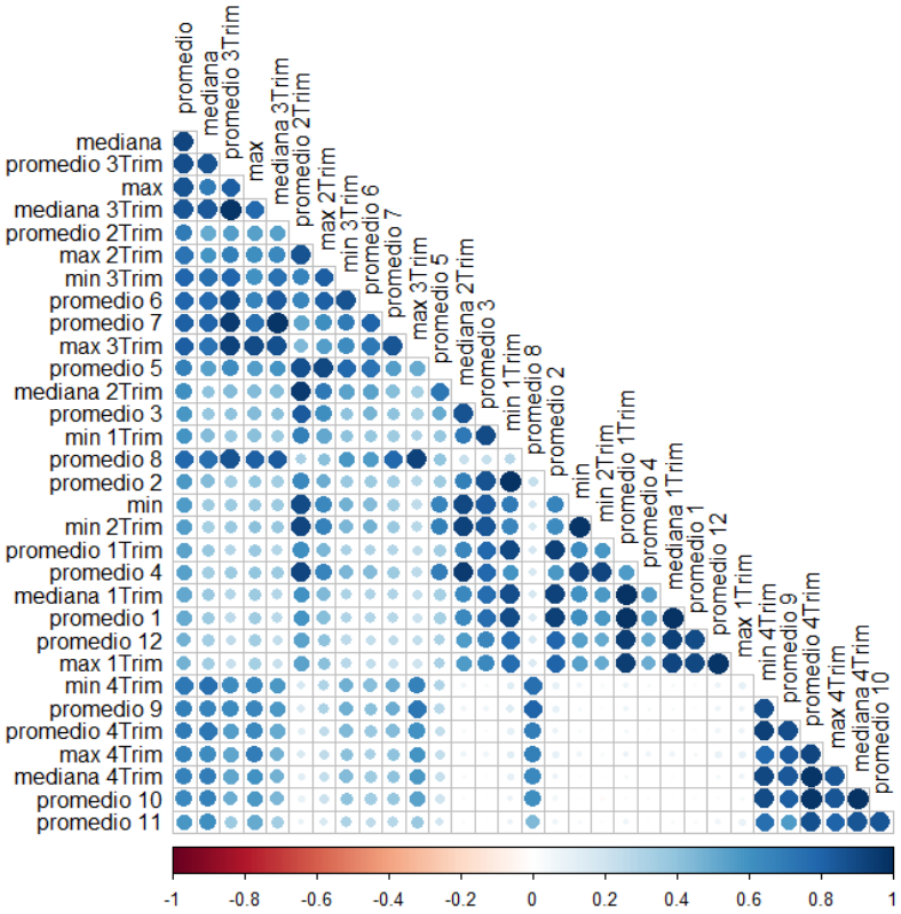

Figura 1.13: Matriz de correlación de la base de datos a utilizar para el análisis que incluye variables creadas para tal fin en base a los valores de caudal del río Limay.

Por un lado, existe una alta correlación entre la media y la mediana anual y los valores del segundo y tercer trimestre que coinciden con la época húmeda del año. Ello implica que las medidas de posición que caracterizan a cada crónica del río Limay se encuentran definidas en los meses húmedos del año y no en los secos.

Por otro lado, la segunda conclusión que se obtiene del presente análisis es que la

correlación entre las *features* del primer trimestre en relación al tercer y cuarto trimestre es muy baja.

1.4.1.6. Estandarización

Con el objetivo de eliminar las diferencias en las unidades de medida y mitigar el impacto de variables con escalas muy diferentes, se prosiguió a estandarizar los datos. Este proceso siguió el mismo enfoque que se aplicó en el caso de los ríos Paraná y Uruguay, dividiendo el período en dos segmentos y utilizando el año 1970 como punto de corte.

1.5. Conclusiones

El estudio exploratorio desarrollado en el presente capítulo arrojó diversas conclusiones por cada uno de los ríos bajo análisis así como también resultados generales que serán presentados a continuación.

En primer lugar, en relación a las características de la serie de tiempo histórica del río Paraná, a partir de la década de 1970 se observa un cambio en la tendencia de los datos, especialmente en los valores mínimos y medianos, confirmado por el test de Chow. Asimismo, los valores máximos muestran una mayor variabilidad en la serie histórica en comparación con los mínimos.

Con respecto al ciclo anual de dicho río, el mismo revela una estación húmeda de diciembre a mayo, una estación seca en septiembre, y extremos de caudal máximo de junio a agosto, coincidiendo con eventos de El Niño.

En términos de correlación, hay baja correlación entre el 1° y 3° trimestre, alta correlación entre el 2° y 3° trimestre, y correlación entre la media y la mediana, las cuales resumen bien la crónica. Además, el 2° y 3° trimestre correlacionan con los máximos.

En segundo lugar, en relación a la caracterización del río Uruguay, los valores máximos también se encuentran en un rango más amplio que los mínimos. Particularmente, para el caso de los valores mínimos, entre 2010 y 2019 el 80 % de los valores mínimos fueron mayores a 1.000 m^3/s cuando previamente a dicha década estos valores eran menores a ese umbral para todos los años.

Asimismo, lo que define al ciclo anual para el río Uruguay es que la estación húmeda es entre mayo y noviembre que coincide con el período seco del Paraná mientras que la estación seca es entre enero y febrero, también en contraposición con lo que sucede en la misma época en el río Paraná.

En cuanto a la correlación de sus variables, la misma es alta entre los valores máximo, media y mediana y los valores del 3° y 4° trimestre. También hay mucha correlación entre el mínimo y los meses secos.

Por último, en el caso del río Límay los valores máximos también se encuentran en un rango más amplio pero dicha amplitud ha disminuido en el último período.

Adicionalmente, en lo que respecta a las características de su ciclo anual, el período seco es entre diciembre y abril mientras que el húmedo es entre junio y noviembre.

En cuanto al análisis de correlación del río Limay, existe una alta valoración de esta medida entre la mediana y la media y el 2° y 3° trimestre. Por el contrario, existe una baja correlación entre el 1° trimestre y el 4° trimestre.

Por último, existen algunas características comunes a las series temporales de los tres ríos bajo análisis entre las que se encuentran que todos los ríos muestran una amplia variabilidad en los valores máximos, a pesar de que su tendencia pueda variar en el tiempo. Adicionalmente, los ciclos anuales muestran patrones distintivos en cada río, con estaciones húmedas y secas que coinciden en el caso del río Uruguay y Limay pero que ambas son opuestas al comportamiento del río Paraná.

En cuanto a la correlación entre las variables, la misma varía entre los ríos pero, en general, se observa que para todos los ríos existe una alta correlación entre la media y la mediana y los trimestres más húmedos del año.

2. CLUSTERIZACIÓN

2.1. Resumen

Mediante las características de cada río analizadas en el capítulo anterior, en la presente sección se aplicaron técnicas de agrupamiento temporal con el objetivo de encontrar patrones comunes en la serie de crónicas hidrológicas de los ríos.

2.2. Metodología

Las técnicas utilizadas para lograr el objetivo de agrupar crónicas fueron, en primera instancia, el análisis de componentes principales, cuyo objetivo fue el de simplificar y resumir la información contenida en el conjunto de datos preservando las relaciones esenciales entre las variables para así facilitar la interpretación de los datos. En segundo lugar, se aplicó el método de K-means para agrupar crónicas de acuerdo a características comunes.

A continuación se explican las técnicas aplicadas.

2.2.1. Análisis de componentes principales

El primer método utilizado fue el de componentes principales que tuvo como objetivo el de reducir la dimensión de los datos, eliminar la multicolinealidad y permitir una mejor visualización de los datos.

Según lo explicado por (Chan y Rey, 2019) basándose en el estudio de (Hotelling, 1933), el análisis de componentes principales se trata de una técnica mediante la cual se realizan combinaciones lineales de variables que tienen una alta correlación dentro de un conjunto de variables que no la tienen resultando así en una menor cantidad de *features* pero que conservan la variabilidad con la que se cuenta en el conjunto original de los datos; es decir, maximizando la varianza. A lo que se denomina componentes principales es al nuevo conjunto de variables.

Es debido a esta finalidad por maximizar la varianza que el método sólo tiene sentido

si existen variables originales que se encuentren correlacionadas entre sí así como también variables latentes. Es decir, variables que no están explícitamente pero que, una vez realizada la técnica de componentes principales, surgen como resultado de la combinación lineal con otras. Es en esta instancia que se vuelve de suma relevancia el análisis realizado en el capítulo anterior de análisis exploratorio y, puntualmente, en el análisis de correlación. Es de suponer que, una vez realizada la técnica, las componentes principales sean la combinación lineal de variables que ya se observó que tienen una gran correlación entre sí.

Si nombráramos a las variables originales como $X_1, X_2, \ldots, X_p$ siendo la variable Y_1 una combinación lineal de ellas, entonces

$$Y_i = \sum_{j=1}^{p} a_{ij} X_j = a_{i1} X_1 + a_{i2} X_2 + ... + a_{ip} X_p \tag{2.1}$$

Donde los coeficientes a_{ij} se denominan cargas (del inglés *loadings*). De tal manera, se busca aquella variable Y_1 tal que tenga norma unitaria y que, a su vez, tenga la varianza máxima entre todas las posibles combinaciones lineales de $X_1, X_2, \ldots, X_p$.

Para la realización de esta técnica se utilizaron las librerías *base* y *stats* de R R Core Team (2021b). En primer lugar se obtuvieron los autovalores y autovectores de la matriz de covarianza muestral para analizar cuántas componentes sería relevante utilizar para el análisis utilizando el criterio de Kaiser y qué porción de variabilidad explican las mismas utilizando el criterio de porcentaje de variabilidad explicada.

El criterio de Kaiser, de acuerdo a lo desarrollado por (Chan y Rey, 2019) consiste en utilizar las m primeras componentes principales a partir de la matriz de correlaciones respetando el criterio de que sus autovalores resulten iguales o mayores que 1:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 1 \text{ y } \lambda_{m+1} < 1 \tag{2.2}$$

Sin embargo, se recomienda utilizar el siguiente criterio, en base a estudios de simulación de Montecarlo:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0{,}7 \tag{2.3}$$

Extendiéndose tal criterio a la matriz de covarianzas, se eligen las primeras m componentes tales que:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq \frac{tr(\sum)}{p} \text{ y } \lambda_{m+1} < \frac{tr(\sum)}{p} \tag{2.4}$$

Nuevamente, puede considerarse la sugerencia de utilizar la siguiente cota inferior:

$$0{,}7\frac{tr(\sum)}{p} \tag{2.5}$$

En segundo lugar se estudiaron las cargas o *loadings* de cada una de las componentes que son los coeficientes de las combinaciones lineales y se estudió su significado y su signo para analizar lo que permite explicar cada componente.

Finalmente, se graficaron las dos primeras componentes principales en un gráfico de *biplot* usando la librería Vu (2011) de R con el objetivo de visualizar en un mismo gráfico las componentes, sus cargas y, finalmente, las observaciones originales dentro del mismo. Ello permitió un posterior estudio gráfico de los posibles clústeres.

2.2.2. Clusterización

El análisis de clústeres se basa en la división de los datos en grupos según sus características. A la hora de entender la información con la que se cuenta, el análisis de clústeres permite asignar grupos a partir de ciertas características que contiene una observación.

En su libro, (Tan y Kumar, 2014) explican que el objetivo de la clusterización es que los objetos dentro de un grupo estén relacionados o sean similares entre ellos pero diferentes de los objetos de los demás grupos. Por lo tanto, cuanto más homogéneos son los objetos de un grupo y mayor es la diferencia entre los grupos, mejor es la clusterización.

En cuanto al método de clasificación, la clusterización es considerada como clasificación no supervisada debido a que se realiza la categorización de los datos sin etiquetas o clases previamente definidas.

El análisis de clusterización consistió en la aplicación del algoritmo de K-means utilizando la librería *stats* de R R Core Team (2021a) y, a continuación, se analizaron las

características de cada uno de los clústeres calculando las mismas métricas de estadística descriptiva que se habían utilizado para el análisis exploratorio de los datos.

Con respecto a la técnica de clusterización de K-means, la misma se trata de una clusterización por particiones. Ello quiere decir que es una división de los datos en subconjuntos que no son superpuestos de manera tal que cada objeto se encuentra en un grupo diferente. Asimismo, es una técnica de clusterización basada en prototipos que, en este caso, refieren a un centroide calculado a partir de la media de un grupo de puntos.

El primer aspecto a definir en el algoritmo de *K-means* es el número de clústeres a realizar. Para ello, se aplicó una técnica también descripta por (Tan y Kumar, 2014) mediante la cual se utilizan dos medidas para encontrar el número óptimo de clústeres a realizar. La primera de ellas es el coeficiente de *Silhouette* que es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de clusterización midiendo cuán similares son las observaciones a su propio clúster en comparación con los demás clústeres. Asimismo, se puede calcular la medida de *Silhouette* de cada clúster calculado el promedio del *Silhouette* de las observaciones del mismo así como también la medida de *Silhouette* de todo el dataset al calcular el promedio de todos los puntos.

La fórmula del coeficiente de Silhouette se puede expresar de la siguiente manera:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}} \tag{2.6}$$

Donde:

- $s(i)$ es el coeficiente de Silhouette para el punto i.
- $a(i)$ es la distancia promedio del punto i a todos los demás puntos en el mismo clúster.
- $b(i)$ es la distancia promedio más pequeña del punto i a todos los puntos en cualquier otro clúster, donde i no pertenece.

La segunda medida a evaluar se trata de la suma del cuadrado de los errores (SSE por sus siglas en inglés) que, en el caso del análisis de clústeres, permite distinguir la distancia

promedio entre las observaciones de un clúster y su centroide. Por definición, este valor tiene una tendencia a descender cuando K aumenta.

En el caso de el SSE, la fórmula se expresa de la siguiente manera:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \overline{x}_j)^2 \tag{2.7}$$

Donde:

- n es el número total de observaciones.
- k es el número total de clústeres.
- x_{ij} es el valor de la observación i en la variable j.
- $\overline{x}_j$ es la media de la variable j en todos los datos.

En conclusión, lo que se busca con este análisis es encontrar aquel punto en el que se forme el llamado codo en el cual el SSE cae significativamente cuando aumenta el K en una unidad. Finalmente, se busca el K que, paralelamente, cumpla con los dos efectos mencionados previamente: un valor de *Silhouette* en aumento y una caída del valor de SSE.

Una vez elegido el número de clústeres, se selecciona ese mismo número de centroides iniciales desde donde comenzar el cálculo. Cada punto es entonces asignado al centroide más cercano y cada conjunto de puntos asignado al mismo centroide es un clúster. Seguidamente, el centroide de cada clúster es actualizado basado en los puntos asignados a cada clúster. Esta operación de asignación y actualización se repite hasta que ninguna observación cambia de clúster o, de manera equivalente, hasta que los centroides permanecen iguales.

En resumen, el algoritmo utilizado se resumiría de la siguiente manera:

1. Siendo K el número seleccionado de vecinos más cercanos y D el número de observaciones.
2. Para cada observación, $z = (x', y')$ realizar:

a) Computar $d(x', x)$ como la distancia entre z y cada una de las observaciones $(x, y) \in D$.

b) Seleccionar $D_z \subseteq D$ como el set de k observaciones más cercanas a z.

c) Siendo $y' = \underset{v}{\operatorname{argmax}} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$

3. Fin del bucle.

2.3. Datos

Para el análisis realizado en el presente capítulo se unificaron las bases de datos desarrolladas en el capítulo anterior. El objetivo de la consolidación de los datos fue estudiar patrones comunes de las crónicas históricas de los tres ríos en su conjunto.

Tras su consolidación y debido a la alta correlación entre las variables de un mismo trimestre hallada en el capítulo anterior, se prosiguió a eliminar de la base a aquellas variables con información mensual para preservar únicamente aquellas con información trimestral y anual. Adicionalmente, fueron creadas dos nuevas variables anuales correspondientes al valor del primer y tercer cuartil del caudal del año.

De esta manera, se obtuvo una única base con las 17 variables creadas y con 297 observaciones: 99 por cada río analizado. Las 99 observaciones corresponden a los años comprendidos en el período 1922-2020 ya que eran los años con los que se contó información completa de los tres ríos. Las mismas se detallan a continuación:

- La mediana, el máximo y el mínimo de cada trimestre del año (12 variables):
  - *mediana 1Trim*: mediana de caudal del primer trimestre
  - *mediana 2Trim*: mediana de caudal del segundo trimestre
  - *mediana 3Trim*: mediana de caudal del tercer trimestre
  - *mediana 4Trim*: mediana de caudal del cuarto trimestre
  - *max 1Trim*: máximo de caudal del primer trimestre
  - *max 2Trim*: máximo de caudal del segundo trimestre
  - *max 3Trim*: máximo de caudal del tercer trimestre

- *max 4Trim*: máximo de caudal del cuarto trimestre
- *min 1Trim*: mínimo de caudal del primer trimestre
- *min 2Trim*: mínimo de caudal del segundo trimestre
- *min 3Trim*: mínimo de caudal del tercer trimestre
- *min 4Trim*: mínimo de caudal del cuarto trimestre

- La mediana, máximo, mínimo, primer y tercer cuartil de cada crónica (5 variables):
  - *mediana*: mediana de caudal de la crónica
  - *max*: máximo caudal de la crónica
  - *min*: mínimo caudal de la crónica
  - *Cuartil 1*: primer cuartil del caudal de la crónica
  - *Cuartil 3*: tercer cuartil del caudal de la crónica

2.4. Resultados

A continuación se presentarán los resultados obtenidos a partir de la aplicación de cada uno de los métodos desarrollados en la sección anterior.

2.4.1. Análisis de componentes principales

Como fuera comentado, el primer paso para el análisis de componentes principales fue estudiar el valor de los autovalores obtenidos a partir de la matriz de correlación utilizando el criterio de Kaiser. Para ello, se extrajeron los autovalores de las primeras componentes siendo el resultado el que se muestra en la Tabla 2.1.

Componente	1	2	3	4	5	6
Autovalor	9,27	2,65	1,38	1,17	0,62	0,45

Tabla 2.1: Autovalores de la matriz de correlación de la base de datos.

De acuerdo a este criterio, los autovalores deben resultar iguales o mayores que 1 ya que, si la varianza fuera inferior a ese valor, se estaría explicando menor variabilidad que sus variables originales. Por lo tanto, se podría afirmar que las primeras 4 componentes son relevantes para el presente análisis.

Asimismo, se utilizó un segundo criterio de selección de componentes a través del análisis de porcentaje de variabilidad explicada. En este caso, los resultados fueron los que se muestran en la siguiente Tabla 2.2:

Componente	1	2	3	4	5	6
Variabilidad explicada	55 %	16 %	8 %	7 %	4 %	3 %

Tabla 2.2: Porcentaje de variabilidad explicada en base a los autovalores de la matriz de correlación de la base de datos.

Mediante este criterio se puede observar que las dos primeras componentes ya tienen una explicabilidad del 71 %, siendo la primera componente la que mayor explicabilidad tiene con un 55 %. Esto permite que en un gráfico de dos dimensiones ya se pueda observar el comportamiento de los datos de acuerdo a las variables que más los describen.

Seguido a ello, se realizó un gráfico de las cargas o *loadings* de las dos primeras componentes y se analizaron sus resultados.

En el caso de la primera de ellas, graficadas sus cargas en la Figura 2.1, se podría decir que se trata de una componente de tamaño ya que todas sus cargas tienen el mismo signo. Ello quiere decir que en esta componente se podrá ver que las observaciones tendrán un valor más alto si el valor de todas sus variables también lo es.

Esta componente presenta las cargas con valores más altos en términos absolutos cuando se trata de variables relacionadas con valores máximos. Particularmente, se trata de los valores máximos anuales y de los trimestres 3 y 2. Trasladando estas conclusiones al análisis exploratorio aplicado en el capítulo previo, los valores máximos de los tres ríos ocurren durante el tercer trimestre.

Por el contrario, la Figura 2.2 muestra las cargas de la segunda componente, que se trata de una componente de forma ya que sus cargas tienen tanto signos positivos como negativos.

El de la segunda componente es un caso en el que existe una gran diferencia en la magnitud de dichas cargas. Por ejemplo, la magnitud del máximo del cuarto trimestre es significativamente mayor que la del mínimo del tercer trimestre.

Asimismo, se puede decir que los *loadings* más relevantes de esta componente se dividen en dos grupos: aquellos con carga positiva y, por otro lado, los de carga negativa. El primero

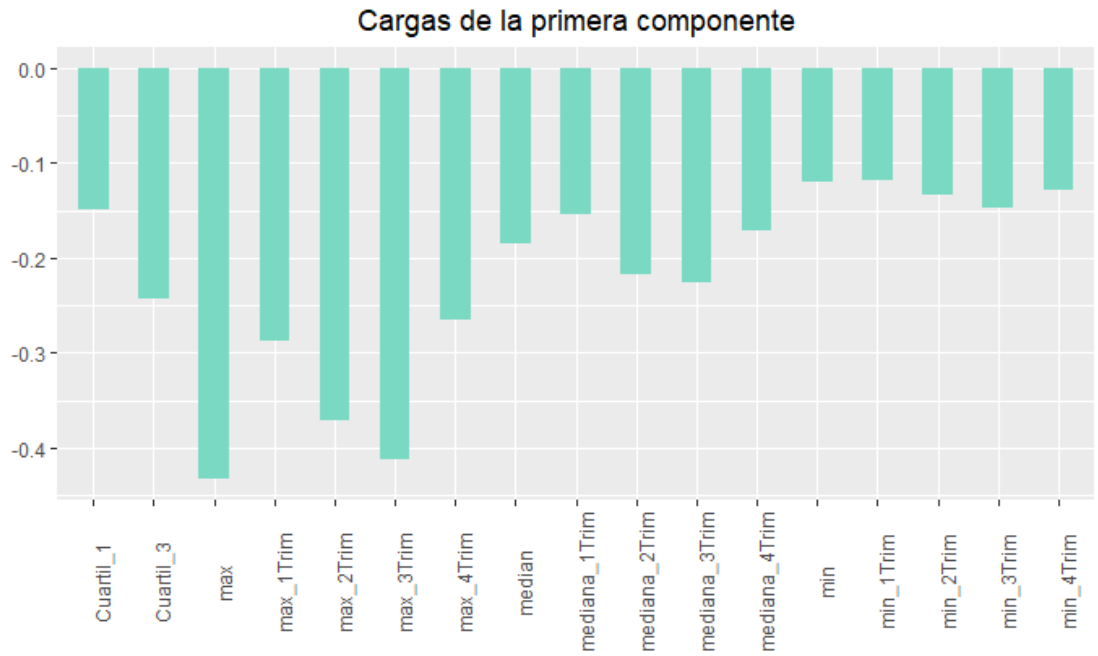

Figura 2.1: Cargas de la primera componente construida a partir de las variables de la base de datos de caudal de los tres ríos analizados.

de ellos está conformado con variables vinculadas con el cuarto trimestre o, en algunos casos, del tercer trimestre mientras que las cargas negativas de mayor valor absoluto están vinculadas con medidas del primer trimestre.

Para finalizar con el análisis de componentes principales se procedió a realizar un biplot con las dos primeras componentes así como también con las cargas analizadas. Por último, fueron graficadas las observaciones de la base de datos original pero en función de las nuevas componentes. Ello se puede observar en la Figura 2.3 donde el eje de abscisas representa la primera componente, el eje de ordenadas la segunda, las flechas marrones identifican a los *loadings* y los puntos negros son las observaciones.

A continuación se pasarán a detallar ciertas conclusiones que fueron obtenidas a partir de la interpretación del *biplot* graficado:

- La primera componente, como ya fue explicado, tiene todas sus cargas con el mismo signo. Ésta explica el 55 % y se podría decir que permite interpretar cuán húmedas o secas son las crónicas. Al ver la dirección de las cargas de las variables mínimo, máximo, mediana y media, se puede afirmar que, cuanto más a la izquierda se en-

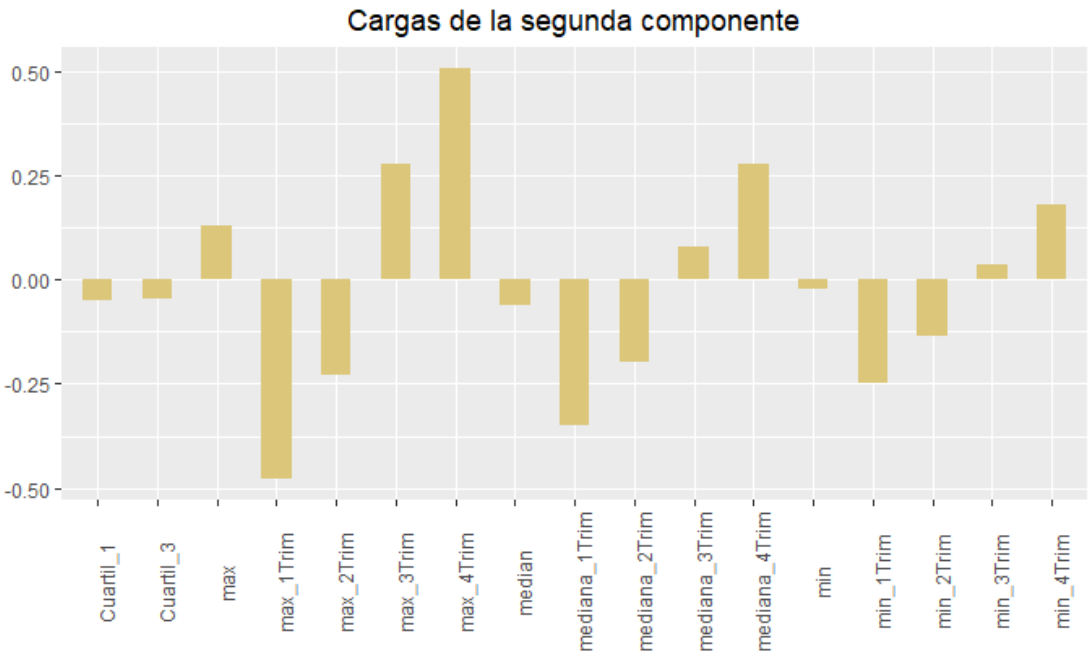

Figura 2.2: Cargas de la segunda componente construida a partir de las variables de la base de datos de caudal de los tres ríos analizados.

cuentra ubicada la crónica, más húmeda es. Eso quiere decir que sus valores máximos, mínimos, medios y medianos son más altos. Por lo tanto, a dicha componente se la podría denominar humedad de la crónica. En este caso, los *loadings* correspondientes a las variables mencionadas y a aquellas del tercer trimestre forman ángulos muy pequeños por lo que se puede afirmar que se trata de variables que sí se encuentran correlacionadas.

- La segunda componente explica la variación que existe entre las temporadas de los ríos contraponiendo cargas del primer y del cuarto trimestre, siendo éstas ortogonales. Cuanto más húmedo fue el cuarto trimestre de las crónicas, más al norte del biplot se encuentra situado aquel registro del biplot. En el caso del río Paraná, un cuarto trimestre con valores relativamente altos significaría una época seca inusualmente húmeda mientras que para el caso de los ríos Uruguay y Limay ello significaría que se trató de un año con una época húmeda extremadamente húmeda. En contraposición, cuanto más al sur se encuentre situado el elemento, significa que se trata de una crónica con un primer trimestre más húmedo que, en el caso del río Paraná

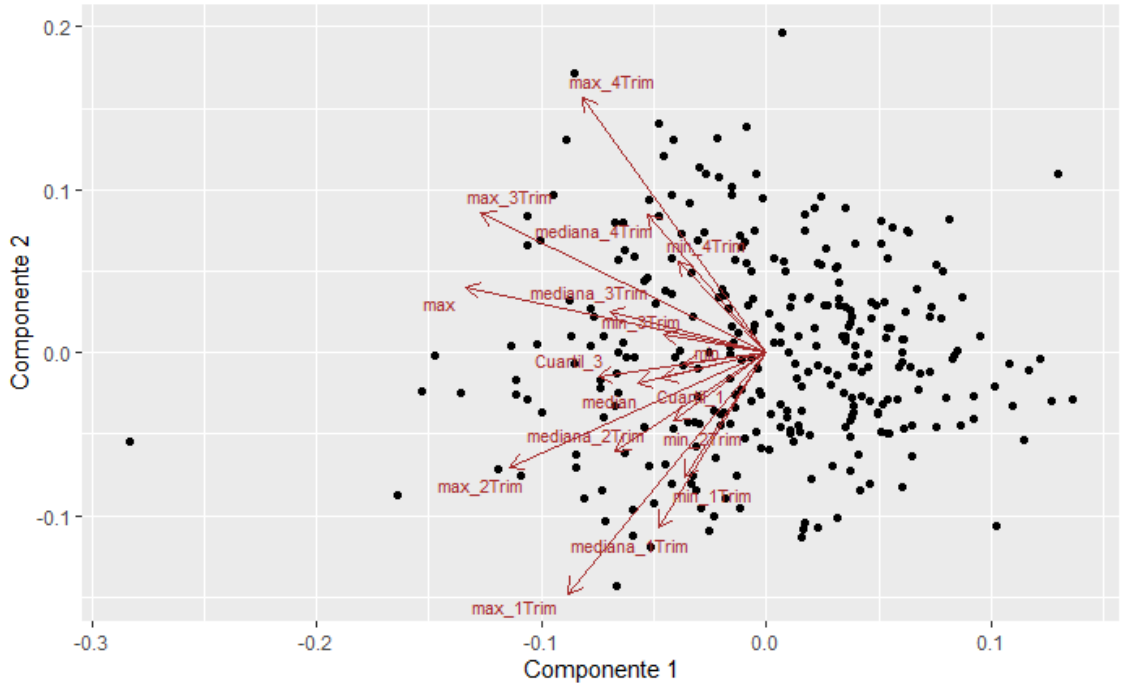

Figura 2.3: Biplot creado con las primeras dos componentes principales donde se grafican las observaciones correspondientes a cada crónica hidrológica. Las flechas marrones se refieren a las cargas de las componentes.

significaría una época húmeda extremadamente húmeda y en el caso del río Uruguay significaría una época seca inusualmente húmeda. En el caso del río Limay el primer trimestre se trata del comienzo de la época seca por lo que sería un caso similar al del río Uruguay.

- Sabiendo que las cargas que se encuentran más juntas son las que tienen más correlación entre ellas y que las que se encuentran ortogonales entre sí son las que no están correlacionadas, entonces se puede reforzar lo observado en el trabajo de análisis exploratorio y en la matriz de correlación que se observa en la Figura 2.4. Por ejemplo, se puede ver que existe una fuerte correlación entre las variables anuales tales como el valor máximo anual, el mínimo y la mediana con las cargas del segundo y tercer trimestre. Asimismo, se puede notar la falta de correlación en las cargas ortogonales entre sí que son las del primer y cuarto trimestre.

- En el gráfico de biplot se pueden analizar los años hidrológicos individualmente para confirmar las observaciones realizadas en los ítems anteriores. Para citar un ejemplo,

se estudió de manera independiente la crónica que se encuentra alejada en valores muy altos del eje de abscisas y se confirmó que se trata del año 1983 de Paraná, un año particularmente húmedo.

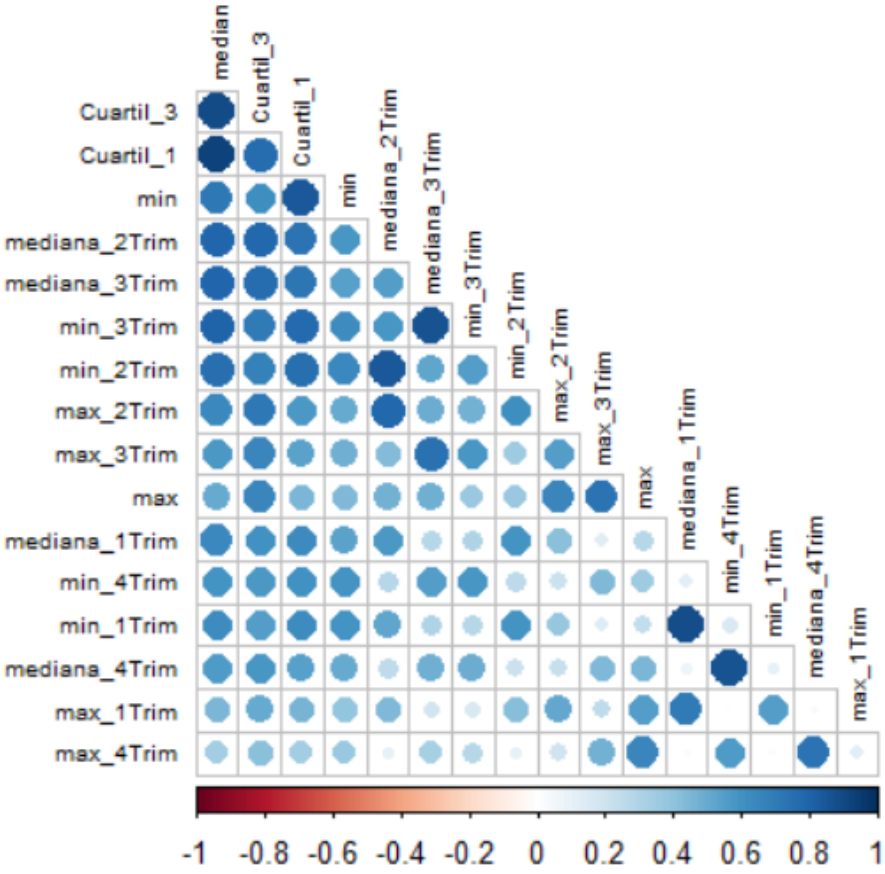

Figura 2.4: Matriz de correlación de la base de datos consolidada en base a los valores diarios de caudal de los ríos Paraná, Uruguay y Limay.

2.4.2. Clusterización utilizando K-means

Tal como fue explicado, el primer paso previo a la aplicación del algoritmo de clusterización es la selección del número de grupos. Para ello, se calcularon las métricas de *Silhouette* y de SSE para el rango de clústeres de 2 a 20 y se observaron sus resultados. Ello se puede analizar a partir de la Imagen 2.5 donde se puede ver en el eje de abscisas el número de clústeres y en el eje de ordenadas cada una de esas medidas. El gráfico superior muestra la medida *Silhouette* y el inferior, la suma de los cuadrados del error.

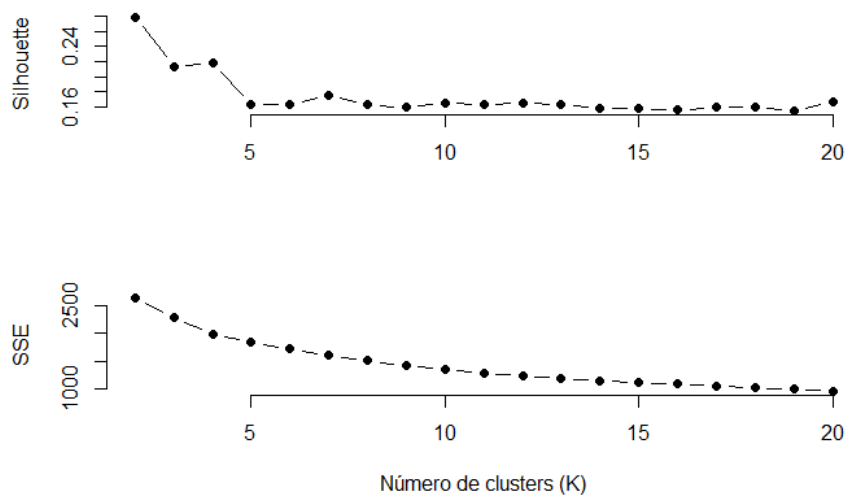

Figura 2.5: El Gráfico superior expone la medida de *Silhouette* y el Gráfico inferior la suma de los cuadrados del error (SSE) en base al número de clústeres. El objetivo del análisis de estas medidas es el de definir el número de clústeres a realizar.

La Figura 2.5 muestra la evolución del coeficiente de *Silhouette* en la medida que aumenta K. El objetivo es encontrar aquel K que muestre un lugar en el que el *Silhouette* se encuentre en aumento. Asimismo, la Figura debajo muestra la evolución del SSE en la medida que aumenta el número de clústeres. Por definición, este valor tiene una tendencia a descender cuando K aumenta. Por lo tanto, lo que se busca es aquel punto en el que SSE se encuentre en caída mientras que el coeficiente de *Silhouette* se encuentre en un punto alto. En este caso, ese punto se encuentra en los 4 clústeres debido a la estabilidad e incluso pequeño aumento del *Silhouette* en un valor alto previo a su caída coincidente con un valor de SSE en disminución.

Una vez definido el número de clústeres a realizar, se procedió a aplicar el algoritmo de K-means en los datos y se obtuvieron los 4 grupos deseados.

Con el objetivo de analizar a qué corresponde cada uno de ellos y qué características propias tienen, se los graficó en el biplot creado en la sección previa que se puede ver en la Figura 2.6.

En esta Imagen también fueron identificadas algunas de las crónicas con la referencia de su año para así hacer un análisis contrastándolo con los datos que ya se tenían de experiencia en el campo. Así es como se pueden estudiar las crónicas que se encuentran,

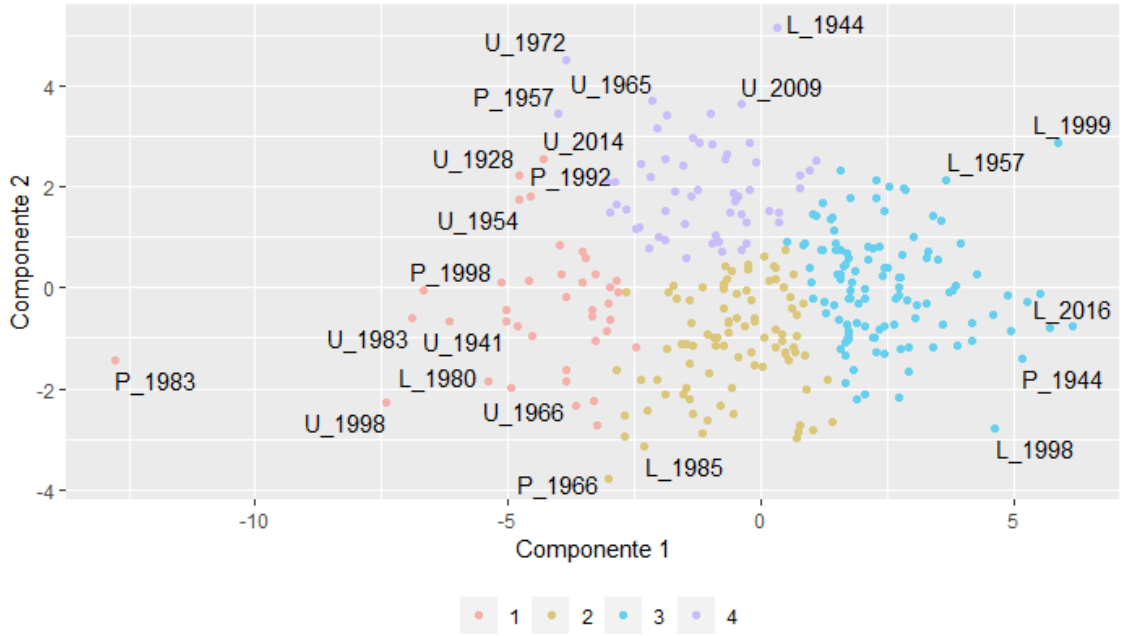

Figura 2.6: Biplot de las dos primeras componentes principales y representación de las crónicas hidrológicas en el mismo. Cada color de las observaciones corresponde a un clúster diferente y el río al cual se refiere cada crónica fue identificado de acuerdo a su letra inicial: U para el caso del Río Uruguay, P para Paraná y L para Limay.

gráficamente, en la periferia del cúmulo de observaciones. Entre ellas se puede resaltar nuevamente el caso de la crónica de 1983, tanto del río Paraná como del río Uruguay que se encuentran agrupadas en un mismo clúster. Como fue explicado en el capítulo anterior, ambas corresponden a crónicas extremadamente húmedas lo cual confirma la descripción realizada acerca de la primera componente: cuanto más a la izquierda del gráfico de biplot, más húmeda es la crónica. Por el contrario, a la derecha del gráfico nos encontramos con crónicas como la de 1944 del río Uruguay o la de 1999 del río Limay que fueron descriptas como crónicas de valores mínimos extremos.

Asimismo, se tomaron como ejemplo a analizar los años 1966 del río Paraná y 1985 del río Limay que se encuentran situados con bajos valores de la segunda componente. En ambos casos se trata de crónicas con valores inusualmente bajos en el último trimestre del año pero extremadamente altos en el primero, tal como se describió a dicha componente.

Por el contrario, al analizar las crónicas de los años 1972 y 1944 de los ríos Uruguay y Paraná respectivamente, se encontraron casos de valores secos en el primer trimestre pero

extremadamente húmedos o con picos máximos de caudal en el cuarto trimestre.

Seguidamente, con el objetivo de observar la serie temporal de crónicas hidrológicas distinguiéndolas según su reciente clusterización y sus medidas estadísticas, fueron confeccionadas las Figuras 2.7, 2.8 y 2.9 que permiten observar la evolución de las series de valores de medianas, mínimos y máximos respectivamente por crónica según el clúster al cual pertenecen.

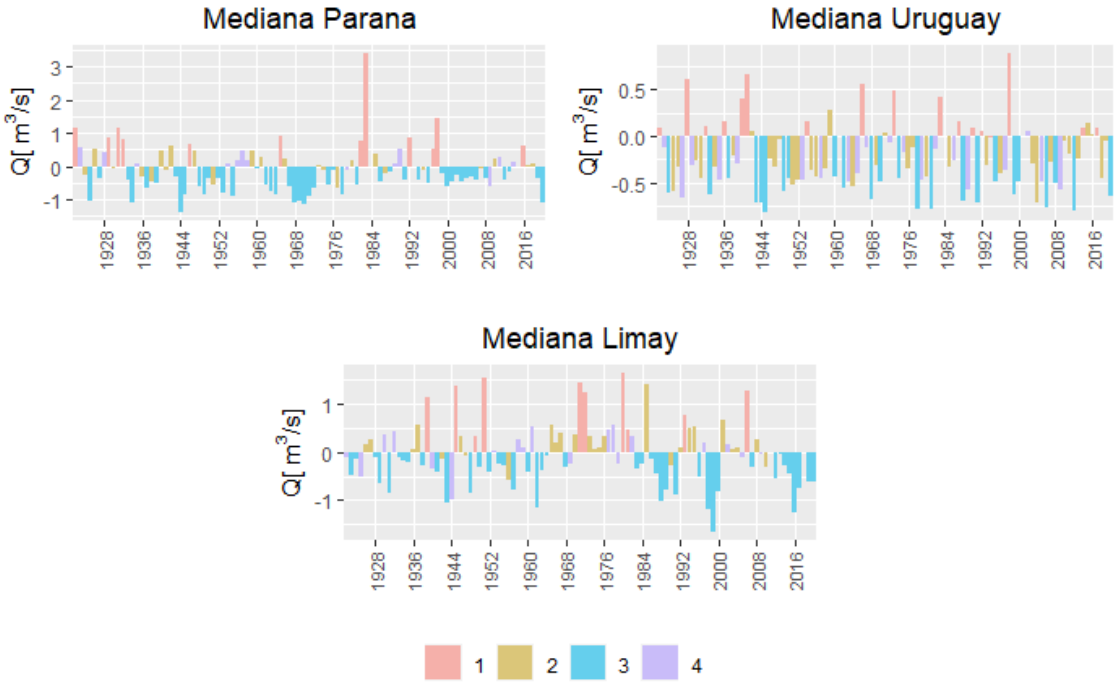

Figura 2.7: Gráfico de barras del valor de mediana de caudal de cada crónica hidrológica del período bajo análisis diferenciadas por color según el clúster al cual pertenece cada una.

Para complementar el análisis y entender la clusterización de acuerdo al comportamiento de las crónicas a lo largo del año, se procedió a graficar cada una de ellas desde diciembre hasta noviembre y superponerlas separándolas por clúster y por río. Las mismas se observan en las Figuras 2.10, 2.11, 2.12 y 2.13 donde se las puede encontrar separadas por río correspondiendo las mencionadas Figuras a los clústeres 1 a 4, respectivamente.

Utilizando como soporte las Figuras mencionadas así como también el biplot de la Figura 2.6, a continuación se procederá a describir cada uno de los clústeres y explicar sus características:

- **Clúster 1:** Se trata del agrupamiento con crónicas más húmedas. Ello se puede

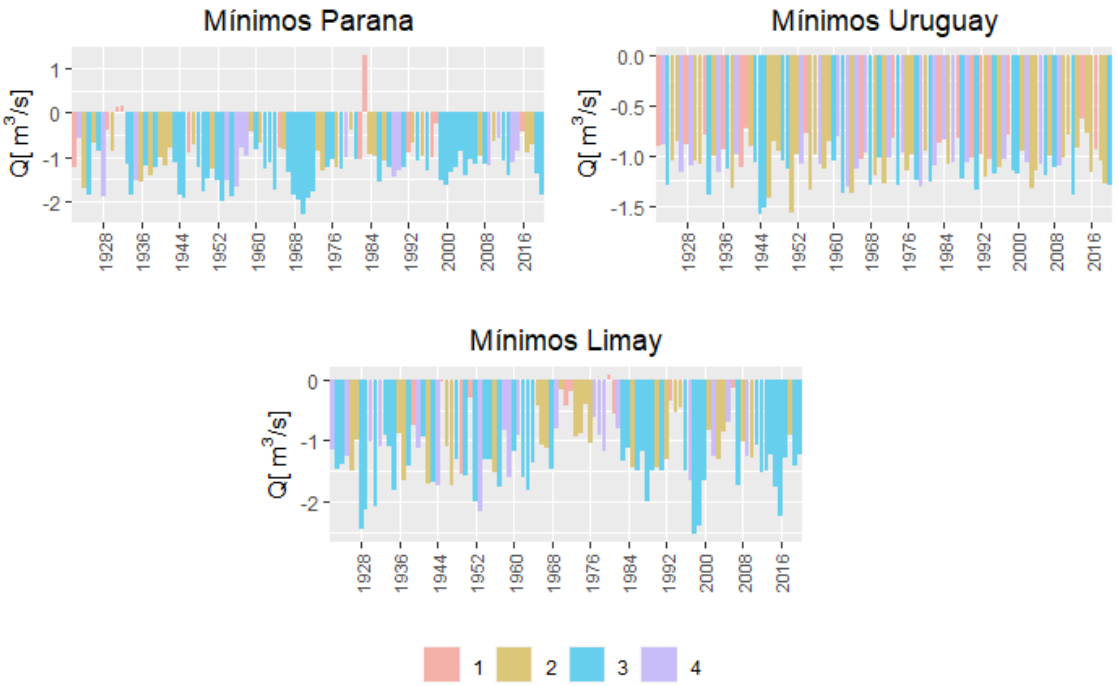

Figura 2.8: Gráfico de barras del valor mínimo de caudal de cada crónica hidrológica del período bajo análisis diferenciadas por color según el clúster al cual pertenece cada una.

observar en la ubicación de los registros dentro del biplot (Figura 2.6) así como también en la Figura 2.7 donde se observa que los valores de la mediana de las crónicas siempre tiene valores positivos. Adicionalmente, los valores mínimos (Figura 2.8) tienden a ser más altos en la mayoría de los casos en comparación con los demás clústeres. A pesar de estas características observadas, no son las únicas crónicas con valores llamativamente altos de valores máximos sino que, como se puede observar en la Figura 2.9, el podio lo comparten con las crónicas del clúster 4 que presenta valores máximos significativamente altos en el cuarto trimestre. Finalmente, al analizar la Figura 2.10 se confirma que se trata de un clúster de crónicas muy húmedas debido a que su caudal con frecuencia es positivo durante todo el año.

- **Clúster 2:** En el caso del segundo agrupamiento, de acuerdo a su ubicación en el biplot de la Figura 2.6 se trata de crónicas de humedad media con un primer trimestre húmedo. En relación a los valores de la mediana anual representados en la Figura 2.7, este clúster no presenta valores extremos de esa métrica. Lo que es más, de acuerdo a cada río se pueden observar características distintas de la mediana

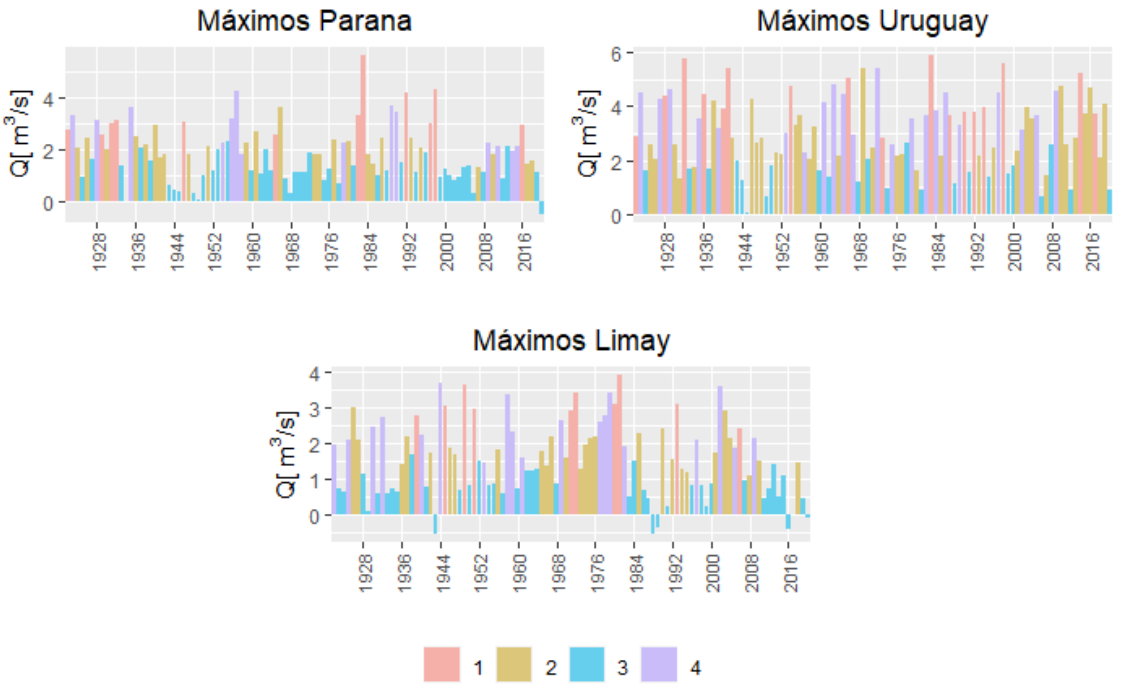

Figura 2.9: Gráfico de barras del valor máximo de caudal de cada crónica hidrológica del período bajo análisis diferenciadas por color según el clúster al cual pertenece cada una.

anual de caudal. Por ejemplo, en el caso del río Uruguay existen valores de mediana que son relativamente bajos. Por el contrario, en el caso de los mínimos de caudal representados en la Figura 2.8, se encuentran valores de mínimos extremos a pesar de no tratarse de un clúster originalmente seco. Ello se puede observar principalmente en el río Uruguay y está vinculado con las características propias del cuarto trimestre que es significativamente seco en este clúster. Con respecto a los máximos que se observan en la Figura 2.9 no existen observaciones relacionadas con éstos ya que no presenta casos de extremos de caudal. Por último, en el gráfico de líneas de la Figura 2.11 se observa que el río Uruguay es el que más crónicas aporta a este clúster. Adicionalmente, en los gráficos se puede ver con claridad que son crónicas que tienen el comportamiento estacional previamente desarrollado: un comienzo húmedo con un primer trimestre con valores altos de caudal pero que finalizan el período siendo comparativamente más secos que su inicio.

- **Clúster 3:** De acuerdo a lo observado en el biplot, el tercer agrupamiento reúne crónicas particularmente secas. Al observar el gráfico de valores de caudal mediano

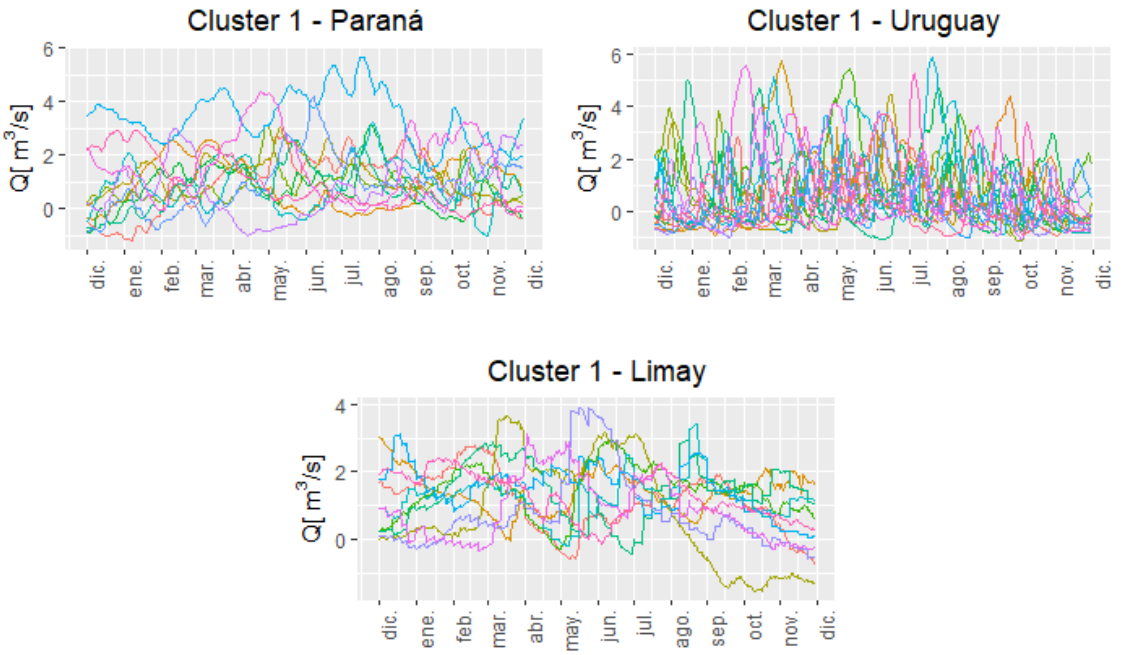

Figura 2.10: Gráfico de líneas por río del caudal medio diario de todas las crónicas hidrológicas pertenecientes al clúster número 1.

anual (Figura 2.7), se puede notar que sus valores son siempre negativos. Algo que se puede resaltar de este clúster es que temporalmente existen períodos de varios años en los que los mismos se clasificaron dentro del mismo clúster 3. Tal es el caso del río Limay en los últimos años bajo análisis o del río Paraná en los años cercanos a 1968 y la década del 2000. Con respecto a las mediciones de valores mínimos (Figura 2.8), éstos son valores extremadamente bajos. Asimismo, los valores máximos (Figura 2.9) son significativamente secos en términos relativos. Por último, la Figura 2.12 muestra que se trata generalmente de crónicas con valores estables a lo largo del año pero que comienzan muy secas y que no se recuperan conforme pasan los meses.

- **Clúster 4:** En el caso del último agrupamiento de crónicas descripto se observan años de humedad media en lo referido al caudal de los ríos pero con mucha variación entre el primer y el cuarto trimestre, teniendo este último valores significativamente más altos de caudal de los ríos. En cuanto a los valores de las medianas anuales (Figura 2.7) se trata de valores que no son extremos sino todo lo contrario. Se podría afirmar que se trata de crónicas con humedad mediana. En relación a sus valores mínimos, se

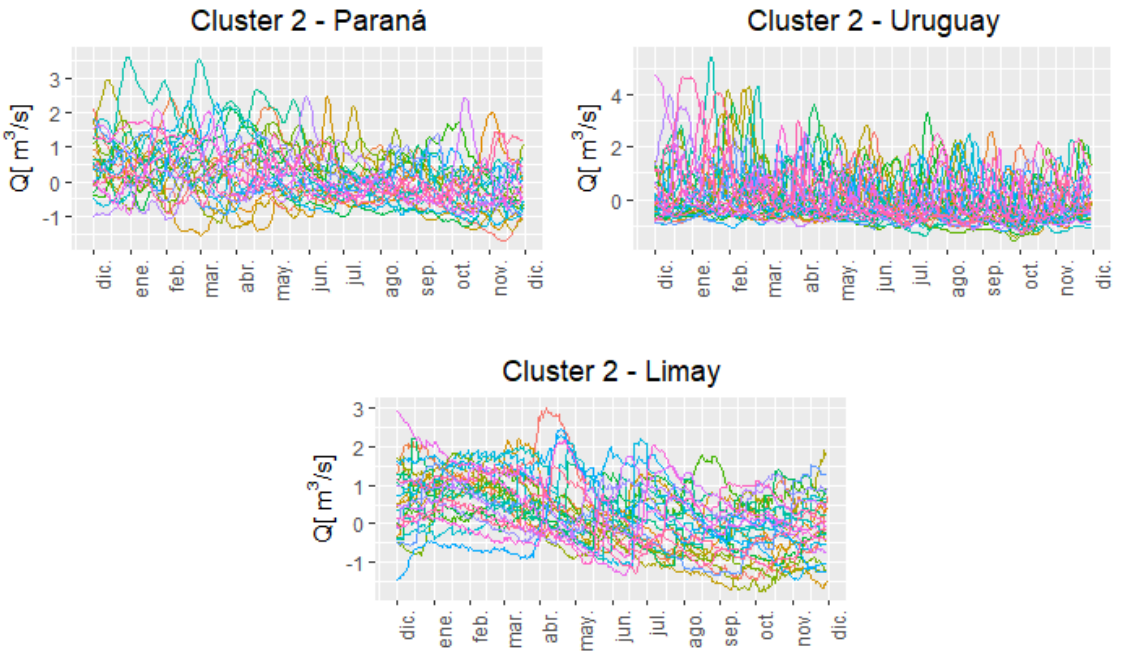

Figura 2.11: Gráfico de líneas por río del caudal medio diario de todas las crónicas hidrológicas pertenecientes al clúster número 2.

observa en la Figura 2.8 que existen algunos casos en los que existen valores mínimos llamativamente bajos lo cual podría significar valores hallados en aquellos primeros meses del año en los que se transita un período seco. Por el contrario en la Figura 2.9 se experimentan valores máximos cercanos a los observados en el clúster 1 que se puede suponer que provienen del cuarto trimestre que es el trimestre de recuperación de este clúster de crónicas. Finalmente, la Figura 2.13 muestra con claridad que este agrupamiento presenta crónicas que comienzan el período con valores muy bajos en el primer trimestre pero que en la medida que avanza el año se recuperan y terminan con un cuarto trimestre con valores muy altos de caudal.

Una vez finalizada la descripción de cada uno de los clústeres, se procedió a crear tres matrices de confusión con el objetivo de mostrar los casos en los cuales para un mismo año hidrológico existía coincidencia entre los ríos en cuanto a la asignación de clústeres. Adicionalmente, fue confeccionada una cuarta matriz con el objetivo de analizar las crónicas de los tres ríos en su conjunto.

La Tabla 2.3 muestra que la mayor coincidencia se encuentra en los clústeres 2 y

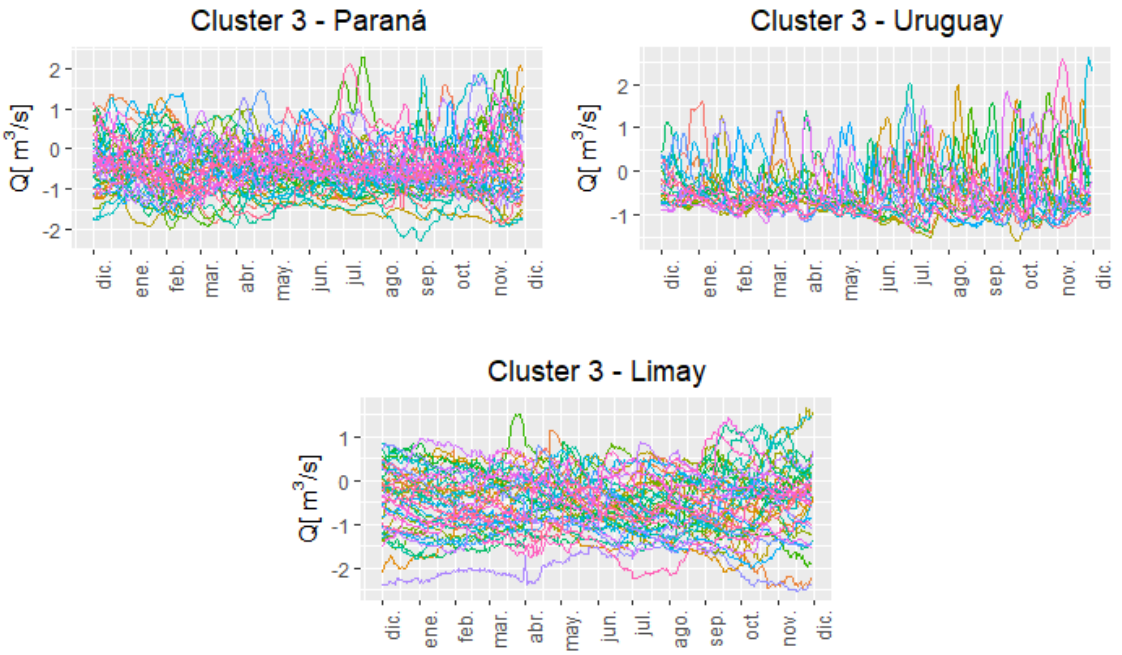

Figura 2.12: Gráfico de líneas por río del caudal medio diario de todas las crónicas hidrológicas pertenecientes al clúster número 3.

3 que son aquellos que tienen más crónicas y, adicionalmente, que tienen un régimen de humedad media a seca. Asimismo, se trata de años en los que se observa un primer trimestre húmedo a medio pero un cuarto trimestre que no es extremadamente húmedo. Recordando el régimen de estos dos ríos, el río Uruguay presenta su época seca en el primer trimestre mientras que el río Paraná en ese trimestre está atravesando su temporada húmeda por lo que podría tratarse de crónicas en las que el Paraná tiene un régimen normal mientras que el Uruguay se encuentra en un año húmedo.

En relación a los restantes cuadrantes de dicha matriz, no se observan patrones de coincidencia entre las crónicas.

		Paraná			
		Clúster 1	**Clúster 2**	**Clúster 3**	**Clúster 4**
Uruguay	**Clúster 1**	4	7	1	4
	Clúster 2	3	14	15	4
	Clúster 3	0	2	21	0
	Clúster 4	4	2	10	5

Tabla 2.3: Matriz de confusión de los clústeres de Uruguay y Paraná por crónica hidrológica.

En cuanto a la matriz de confusión de la Tabla 2.4 que cruza los resultados de la

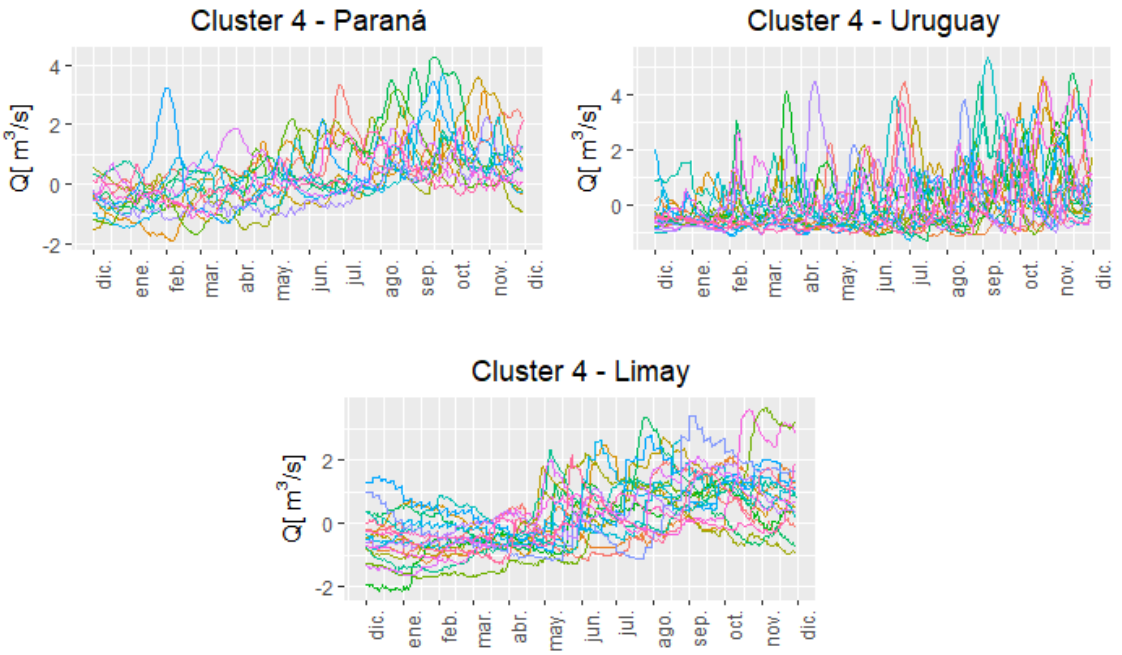

Figura 2.13: Gráfico de líneas por río del caudal medio diario de todas las crónicas hidrológicas pertenecientes al clúster número 4.

clusterización para los ríos Uruguay y Limay, se puede ver una situación similar a aquella descripta para el caso Uruguay-Paraná. En este caso existe una diferencia que está relacionada con el régimen de estos ríos. En ambos casos la temporada más seca se halla en los primeros dos trimestres del año mientras que la húmeda está en los últimos dos. Esto puede resultar en una mayor cantidad de aciertos de coincidencia pero que se deben a la similitud del régimen de ambos ríos. Por otro lado, cabe resaltar que no existió ningún año en el cual se esté atravesando por una crónica del clúster 1 que es el más húmedo de los cuatro.

		Limay			
		Clúster 1	**Clúster 2**	**Clúster 3**	**Clúster 4**
Uruguay	**Clúster 1**	0	6	8	2
	Clúster 2	4	12	14	6
	Clúster 3	4	5	12	2
	Clúster 4	2	4	7	8

Tabla 2.4: Matriz de confusión de los clústeres de Uruguay y Limay por crónica hidrológica.

Adicionalmente, la Tabla 2.5 muestra la matriz de confusión de los ríos Paraná y Limay a partir de la cual se puede obtener una conclusión común con respecto a la tabla anterior

relacionada con la falta de entrecruzamiento de las crónicas del primer clúster. Por otro lado, con respecto a los casos en los que se comparte o se confunde entre el clúster 2 y el 3, se podría decir que se trata de crónicas en las que el Limay pudo haber tenido una época seca de características más húmedas.

		Limay			
		Clúster 1	**Clúster 2**	**Clúster 3**	**Clúster 4**
Paraná	**Clúster 1**	0	3	5	3
	Clúster 2	3	11	6	5
	Clúster 3	7	11	22	7
	Clúster 4	0	2	8	3

Tabla 2.5: Matriz de confusión de los clústeres de Paraná y Limay por crónica hidrológica.

Finalmente, la Tabla 2.6 permite observar la matriz de confusión de los ríos Paraná, Uruguay y Limay en su conjunto con el objetivo de estudiar si existen patrones en el comportamiento de sus crónicas a través del análisis de los clústeres creados. Es así que se puede arribar a las siguientes conclusiones:

- Existe un patrón común relacionado con el tercer cluster que permite identificar una tendencia a los casos en los que los tres ríos presentan crónicas secas durante el mismo año.

- Asimismo, vinculado con este fenómeno, otro patrón que se repite es el de años con crónicas secas en los ríos Limay y Paraná pero coincidente con una crónica del río Uruguay de humedad media y sólo con el primer trimestre húmedo.

- Por otro lado, se observa también el patrón de los tres ríos con el primer trimestre húmedo pero con el resto del año con humedad media.

- También se resalta que no se han observado años en los que los tres ríos hayan tenido crónicas húmedas.

- Finalmente, no es significativo el caso en el que una cuenca compense a la otra en términos de humedad de sus crónicas.

		Limay			
Paraná	**Uruguay**	**Clúster 1**	**Clúster 2**	**Clúster 3**	**Clúster 4**
Clúster 1	**Clúster 1**	0	1	2	1
	Clúster 2	0	1	2	0
	Clúster 3	0	0	0	0
	Clúster 4	0	1	1	2
Clúster 2	**Clúster 1**	0	3	3	1
	Clúster 2	3	6	2	3
	Clúster 3	0	2	0	0
	Clúster 4	0	0	1	1
Clúster 3	**Clúster 1**	0	1	0	0
	Clúster 2	1	4	8	2
	Clúster 3	4	3	12	2
	Clúster 4	2	3	2	3
Clúster 4	**Clúster 1**	0	1	3	0
	Clúster 2	0	1	2	1
	Clúster 3	0	0	0	0
	Clúster 4	0	0	3	2

Tabla 2.6: Matriz de confusión de los clústeres de Paraná, Uruguay y Limay por crónica hidrológica.

2.5. Conclusiones

Los análisis llevados a cabo en el presente capítulo permitieron obtener conclusiones vinculadas tanto con la creación de las nuevas componentes principales mediante la combinación de variables como con la clusterización de las crónicas hidrológicas de acuerdo a características comunes del comportamiento del caudal.

Tras el análisis de las componentes principales derivadas de los datos de los tres ríos, se ha observado que la primera componente está fuertemente asociada con los valores máximos y con los del segundo y tercer trimestre, coincidiendo con los picos de caudal máximo durante el tercer trimestre en los tres ríos. Esto indica la naturaleza de las crónicas en términos de humedad y sequedad.

Por otro lado, la segunda componente captura la variabilidad entre las estaciones, destacando las diferencias entre las cargas del primer y cuarto trimestre. Asimismo, se encontró una correlación significativa entre las variables anuales, como el máximo, el mínimo y la mediana, y las cargas del segundo y tercer trimestre.

En cuanto a la clusterización de las crónicas utilizando el algoritmo kmeans, se identificaron cuatro clústeres distintivos:

- Cluster 1: crónicas más húmedas.
- Cluster 2: crónicas con humedad media, especialmente en el primer trimestre.

- Cluster 3: crónicas secas.
- Cluster 4: humedad media, con valores secos al inicio del año y un cuarto trimestre muy húmedo.

Finalmente, en cuanto al estudio de los patrones en el comportamiento hidrológico de los tres ríos para un mismo año hidrológico, se arribó a las conclusiones que se presentan a continuación.

En primer lugar, se ha observado un patrón común relacionado con el tercer cluster, que indica una tendencia hacia años en los que los tres ríos presentan crónicas secas simultáneamente. Además, se han identificado patrones en los que los ríos Limay y Paraná muestran crónicas secas, mientras que el río Uruguay presenta una crónica de humedad media, específicamente en el primer trimestre.

También se ha observado un patrón en el que los tres ríos tienen un primer trimestre húmedo pero el resto del año presenta humedad media. Sin embargo, no se han encontrado años en los que los tres ríos tengan crónicas húmedas simultáneamente.

En conclusión, no se ha observado un patrón significativo de compensación entre las cuencas en términos de humedad de las crónicas.

3. PRONÓSTICO

3.1. Resumen

Partiendo del agrupamiento de las crónicas hidrológicas desarrollado en el capítulo anterior, en la presente sección fueron desarrollados diversos modelos probabilísticos con el objetivo de predecir las características futuras del año hidrológico en curso, basándose en el comportamiento de los ríos involucrados.

De tal modo, se utilizaron las variables creadas en las secciones anteriores con el fin de construir modelos que predigan la pertenencia de cada crónica a los clústeres.

Adicionalmente, fue estudiada la influencia de oscilaciones climáticas como ENSO (El Niño-Southern Oscillation) y PDO (Pacific Decadal Oscillation). Dichos efectos también han sido estudiados por autores como (González et al., 2017a) hallando una alta correlación entre el fenómeno en cuestión y las precipitaciones, especialmente en el Noreste y Sur de los Andes en primavera y, en menor medida, en otoño. (Garbarini et al., 2016) también han estudiado sus efectos sobre las precipitaciones, encontrando que las estaciones de transición presentan señales más fuertes que el invierno y el verano.

3.2. Datos

La base de datos utilizada para la construcción de los modelos predictivos fue aquella consolidada en el capítulo anterior, que incluye información sobre los caudales de los tres ríos. Sin embargo, para esta investigación, se seleccionaron específicamente variables asociadas al primer, segundo y tercer trimestre. De este modo, para cada método de predicción seleccionado, se desarrollaron tres modelos distintos: uno utilizando únicamente las variables del primer trimestre, otro con las del primer y segundo trimestre, y finalmente, un tercer modelo que incorpora variables desde el primer hasta el tercer trimestre.

Además de los datos de caudal, se incorporaron índices relacionados con anomalías en la temperatura de la superficie del mar (SST por sus siglas en inglés o TSM por sus

siglas en español) en regiones específicas del Océano Pacífico tropical. Las características de aquellos índices utilizados en el presente trabajo se detallan a continuación:

- **NIÑO 1+2:** Las mediciones del índice se realizan en la región 0-10S, 90W-80W. Es la región más pequeña y oriental de las de SST Niño y corresponde a la zona de la costa de Sudamérica donde El Niño fue reconocido por primera vez por las poblaciones locales. Este índice tiende a tener la mayor variabilidad de todos los índices de SST Niño.

- **NIÑO 3:** Las mediciones del índice se realizan en la región 5N-5S, 150W-90W.

- **NIÑO 3.4:** Las mediciones del índice se realizan en la región 5N-5S, 170W-120W. Las anomalías de Niño 3.4 pueden considerarse como la representación de las SST ecuatoriales promedio a lo largo del Pacífico desde aproximadamente la línea de cambio de fecha hasta la costa de Sudamérica. Los eventos de El Niño o La Niña se definen cuando las SST de Niño 3.4 superan los +/- 0.4°C durante seis meses o más.

- **NIÑO SOI:** Sus siglas en inglés corresponden al nombre de Índice de Oscilación del Sur. Es la diferencia entre los valores estandarizados de presión en superficie en Darwin y los valores estandarizados en Tahití. El signo es opuesto al de los índices Niño y es más ruidoso que esos índices.

- **NIÑO PDO:** Sus siglas en inglés corresponden al nombre de Oscilación Decadal del Pacífico. Es la componente principal líder de las anomalías mensuales de la temperatura superficial del mar en el océano Pacífico Norte.

Los datos de los mencionados índices provienen de la página web del Physical Sciences Laboratory (PSL), detallados para cada uno de los índices. Los mismos fueron obtenidos a través de su portal web en el siguiente url: `https://psl.noaa.gov/enso/dashboard.html`.

3.3. Metodología

La técnica utilizada para lograr el objetivo de pronosticar crónicas futuras fue el análisis de diferentes métodos de predicción entre los que se encontraron el de K vecinos más cercanos, el método de árboles de decisión y la regresión logística.

A continuación se explican los tres métodos elegidos así como también la forma de dividir la base en conjuntos y las métricas utilizadas para evaluar los modelos.

3.3.1. Métodos de predicción

3.3.1.1. K vecinos más cercanos

Según (Hastie et al., 2009), el método de *K vecinos más cercanos* (KNN por sus siglas en inglés) es un algoritmo de aprendizaje supervisado que es comúnmente utilizado para problemas de clasificación. En este enfoque, los nuevos puntos de datos se clasifican o predicen en función de cómo se relacionan con los puntos de datos vecinos más cercanos en el espacio de características.

Dado que la proximidad entre los puntos de datos se determina según la distancia entre ellos, inicialmente debemos seleccionar la distancia a aplicar, comúnmente representada por la distancia euclidiana. Esta se define mediante la siguiente fórmula:

$$d(i) = \|x_i - x_0\|, \quad \text{para todo } i \in \mathbb{N} \tag{3.1}$$

Donde $d(i)$ es la distancia entre el punto de datos x_i y el nuevo punto x_0.

Adicionalmente, el hiperparámetro a definir es el número de vecinos (K) que aplicaría el algoritmo. El mismo representa el número de puntos de datos más cercanos que se considerarán para tomar una decisión.

Por último, se debe tomar una decisión relacionada con el método a utilizar para determinar la clase del nuevo punto.

En este caso, se empleó el software estadístico R (versión 4.1.1) y, en particular, se utilizó el paquete *class* (Venables y Ripley, 2002a) bajo la función *knn* (Venables y Ripley, 2002b) para aplicar el mencionado algoritmo, el cual se define con la distancia euclidiana

y un sistema de votación.

Las ventajas de este algoritmo son que es fácil de entender e implementar, no asume distribuciones particulares en los datos y es útil cuando la frontera de datos es irregular.

Por otro lado, las desventajas de este algoritmo son que es sensible a datos atípicos y que puede volverse computacionalmente costoso para grandes conjuntos de datos.

Por último, resulta relevante analizar el compromiso (o *trade off*) entre elegir un K alto que implique un alto sesgo, baja varianza y subajuste (o *underfitting*) en contraposición a un K bajo que signifique un bajo sesgo pero alta varianza y sobreajuste (u *overfitting*).

3.3.1.2. Árboles de decisión

El segundo método de predicción utilizado en el presente capítulo es el de árboles de decisión cuya descripción es detallada por (Alpaydin, 2020). Tal como se describe en el libro, se trata de una técnica de aprendizaje supervisado mediante la cual se parte de un nodo inicial que incluye todo el conjunto de datos. A continuación se elige la característica o *feature* que mejor divide los datos en subconjuntos homogéneos. La métrica comúnmente utilizada es la ganancia de información o la impureza de Gini. En este caso se utilizó el índice de Gini que se define de la siguiente manera:

$$\text{Gini} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}} \tag{3.2}$$

Donde:

- n es el número total de individuos o elementos en la muestra.
- x_i es el valor de la variable de interés para el i-ésimo individuo.
- $\bar{x}$ es el promedio de los valores de la variable de interés en la muestra.

Este mismo proceso se repite en cada subconjunto resultante de manera recursiva hasta alcanzar un criterio de parada, como una profundidad máxima o un número mínimo de observaciones en un nodo.

Asimismo, cada hoja del árbol representa una clase en el caso de clasificación o un valor en el caso de regresión. La decisión en cada hoja se basa en la mayoría de las observaciones

en esa hoja.

En este caso, en el cual el método es utilizado con el fin de predecir, cada nueva observación se mueve a través del árbol, siguiendo las divisiones basadas en sus características hasta llegar a una hoja.

La principal ventaja de este método es su interpretabilidad. Sin embargo, su principal deventaja es que los árboles pueden ser propensos al sobreajuste.

En este caso se empleó el software estadístico R (versión 4.1.1) y, en particular, se utilizó el paquete *rpart* (Therneau et al., 2021), bajo la función *rpart*, para aplicar el mencionado algoritmo.

Por último, los hiperparámetros del algoritmo empleado a través de la mencionada función, de acuerdo a la bibliografía de la librería utilizada son:

- *maxdepth:* Establece la profundidad máxima de cualquier nodo del árbol final, contando el nodo raíz como profundidad 0.
- *minsplit:* Establece el número mínimo de observaciones que deben existir en un nodo para intentar una división.
- *minbucket:* Se tata del número mínimo de observaciones en cualquier nodo terminal (hoja).
- *cp:* Corresponde a un parámetro de complejidad mediante el cual cualquier división que no mejore el resultado probablemente será podada y, por lo tanto, el programa no necesita seguirla.

El criterio utilizado en el presente trabajo fue el de establecer todos los hiperparámetros en valores correspondientes a casos en los que el modelo sobreajustaría a excepción de *minsplit*. De esta manera, el algoritmo utiliza a este último como único hiperparámetro para regular el sobreajuste.

3.3.1.3. Regresión logística

El tercer método utilizado en el presente análisis se trata de la regresión logística que, tal como es descripta por (James et al., 2013), se trata de un método estadístico

cuya función es predecir la probabilidad de pertenencia de un evento a una determinada categoría.

Tomando el caso más simplificado posible que correspondería al de una clasificación binaria con una única variable predictora, y tomando los datos del presente trabajo como ejemplo, se podría modelar la probabilidad de que un registro pertenezca, por ejemplo, al clúster 1 dado el caudal máximo del primer trimestre.

En ese caso, la probabilidad de pertenencia al clúster 1 dado el caudal máximo del primer trimestre puede escribirse como:

$$Pr(\text{clúster} = 1|\text{Caudal máximo 1° trimestre}) \tag{3.3}$$

Los valores resultantes varíarían entre 0 y 1. Por lo tanto, para cualquier valor dado de caudal máximo del primer trimestre, se puede realizar una predicción acerca de si pertenecerá o no al clúster número 1.

Bajo esta técnica, la función utilizada para modelar la probabilidad es la función logística debido a su característica de arrojar resultados entre 0 y 1 para todos los valores ingresados. De esta manera, la función utilizada es la siguiente:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{3.4}$$

Donde:

- β_0, β_1 son los coeficientes para la clase k.
- X es la variable independiente.
- e es la base del logaritmo natural.

Para ajustar el modelo se utiliza el método de máxima verosimilitud y se observa que se pueden predecir probabilidades de ocurrencia cercanas a cero, pero nunca por debajo de cero y, por el contrario, probabilidades de ocurrencia cercanas a uno, pero nunca por encima de uno. La función logística siempre producirá una curva en forma de S por lo que, independientemente del valor de X, obtendremos una predicción sensata a diferencia

del modelo lineal que arrojaría valores por encima de uno o por debajo de cero. Por este motivo es que el modelo logístico es mejor para capturar el rango de probabilidades que el modelo de regresión lineal.

Una vez explicado el modelo en su escenario más simplificado, se considerará el problema de predecir una respuesta binaria utilizando múltiples predictores. Por analogía con la extensión de la regresión lineal simple a múltiple, se puede generalizar la función de la siguiente manera:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{3.5}$$

Donde:

- p es la cantidad de predictores y es un número natural mayor que cero.
- $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes que determinan la influencia de cada variable independiente en la predicción de la probabilidad.
- $X_1, X_2, \ldots, X_p$ son las variables independientes.
- e es la base del logaritmo natural, que transforma la salida de la función lineal en una probabilidad válida en el rango $(0, 1)$.

Asimismo, la regresión logística para el caso de más de dos clases se conoce comúnmente como regresión logística multinomial. A diferencia de la regresión logística binomial, que se utiliza para problemas de clasificación binaria, la regresión logística multinomial se emplea cuando hay tres o más categorías en la variable dependiente.

La formulación general de la regresión logística multinomial se puede expresar de la siguiente manera. Suponiendo que se cuenta con K clases, la probabilidad de que una observación i pertenezca a la clase k se expresa mediante la función:

$$P(Y_i = k) = \frac{e^{\beta_{k0}+\beta_{k1}X_{i1}+\beta_{k2}X_{i2}+...+\beta_{kp}X_{ip}}}{\sum_{j=1}^{K} e^{\beta_{j0}+\beta_{j1}X_{i1}+\beta_{j2}X_{i2}+...+\beta_{jp}X_{ip}}} \tag{3.6}$$

Donde:

- p es la cantidad de predictores y es un número natural mayor que cero.
- Y_i es la variable dependiente para la observación i siendo i un número natural mayor que cero.
- K es el número de clases y es un numero natural mayor que cero.
- $\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp}$ son los coeficientes para la clase k.
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ son las variables independientes para la observación i.
- e es la base del logaritmo natural.

En resumen, la regresión logística multinomial extiende el concepto de la regresión logística binomial a la predicción de múltiples categorías.

Asimismo, la función de costo de la regresión logística se denomina función de pérdida logarítmica y se puede definir de la siguiente manera:

$$\text{Costo} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$$

Donde:

- n es el número de observaciones en el conjunto de datos.
- y_i es el valor verdadero de la variable dependiente para la observación i siendo i un número natural mayor que cero.
- $\hat{p}_i$ es la probabilidad predicha de que la observación i pertenezca a la clase positiva.

Por último, al modelo se le pueden aplicar técnicas de regularización. La regularización es utilizada en el aprendizaje automático y la estadística para evitar el sobreajuste en modelos, especialmente cuando éstos tienen un gran número de parámetros.

La idea central de la regularización es agregar un término de penalización a la función de pérdida (o costo) del modelo, que castiga los modelos que tienen coeficientes de parámetros demasiado grandes. Esta penalización ayuda a prevenir que los coeficientes tomen valores extremos, reduciendo así la complejidad del modelo y mejorando su capacidad de generalización.

Las dos penalizaciones para aplicar la regularización en el presente trabajo son las de Lasso y Ridge.

La penalización por Lasso introduce un término de regularización en la función de pérdida del modelo. Este término está proporcional a la suma de los valores absolutos de los coeficientes del modelo. Esta técnica tiene la propiedad de hacer que algunos coeficientes sean exactamente cero, lo que efectivamente realiza la selección de variables, eliminando algunas características del modelo.

Por consiguiente, la función de costo con penalización Lasso quedaría expresada de la siguiente manera:

$$\text{Costo con penalización Lasso} = \text{Costo} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.7}$$

Donde:

- λ es el parámetro de penalización.
- p es el número de predictores siendo p un número natural mayor que cero.
- β_j es el coeficiente de regresión asociado al predictor j.

En cambio, la penalización por Ridge agrega el término de penalización proporcional a la suma de los cuadrados de los coeficientes. Evita que los coeficientes tomen valores extremos, pero generalmente no los reduce a cero. Ridge es útil cuando hay multicolinealidad entre las variables independientes.

De esta manera, la función de costo quedaría expresada de la siguiente manera:

$$\text{Costo con penalización Ridge} = \text{Costo} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.8}$$

Donde:

- λ es el parámetro de penalización.
- p es el número de predictores siendo p un número natural mayor que cero.
- β_j son los coeficientes de regresión.

Asimismo, profundizando la definición de *lambda* (λ), un valor más alto de λ implica una mayor penalización y, por lo tanto, coeficientes más pequeños. Ajustar su valor permite equilibrar la precisión del modelo con la magnitud de los coeficientes.

En este caso se empleó el software estadístico R (versión 4.1.1) y, en particular, se utilizó el paquete *glmnet* (Jerome Friedman y Narasimhan, 2021), bajo la función *glmnet*, para aplicar el mencionado algoritmo.

Para la aplicación de esta técnica, en el presente trabajo fueron realizados dos modelos: uno para la penalización Ridge ($\alpha = 0$) y otro para la penalización Lasso ($\alpha = 1$). Adicionalmente, en cada uno de los modelos creados, el hiperparámetro a optimizar se trató del mencionado parámetro de penalización *lambda* (λ).

3.3.2. División de la base en conjuntos

La base de datos utilizada en este estudio se dividió en varios conjuntos con el propósito de entrenar, validar y evaluar los modelos predictivos de manera robusta. La separación se llevó a cabo de la siguiente manera:

3.3.2.1. Holdout

El primer conjunto seleccionado consistió en el período temporal más reciente de cada río, abarcando desde el año 2001 hasta el 2020 y totalizando un 20 % de los datos. Los mismos fueron separados para constituir el conjunto de *holdout* que se reservó intacto durante todo el proceso de modelado y fue utilizado exclusivamente al final del análisis para evaluar el rendimiento real del modelo que finalmente fue elegido.

3.3.2.2. Conjunto de entrenamiento y testeo

Este conjunto consiste en el 80 % restante de los datos y se divide, a su vez, en 2 subconjuntos que se detallan a continuación:

- **Testeo (20 %):** Se eligieron los 20 años anteriores al período de *holdout* (1981-2000) para formar el conjunto de prueba (*test set*). El objetivo de la utilización de este conjunto es el de elegir el modelo de mejor rendimiento.

- **Entrenamiento (60 %):** Los años restantes, una vez excluido el conjunto de prueba, son los utilizados para entrenar y validar los modelos. Para ello, fue implementada la técnica de validación cruzada (CV por sus siglas en inglés) con 3 grupos o *folds*. El procedimiento se divide en tres instancias en las cuales se utilizan dos de los grupos como grupos de entrenamiento y el tercero como grupo de validación en el cual se evalúa el rendimiento del modelo. En cada una de las instancias se alternan los grupos de entrenamiento y validación. Asimismo, la evaluación del rendimiento se realiza utilizando métricas que serán descriptas en la próxima sección. De esta manera, se proporcionan tres evaluaciones independientes del rendimiento del modelo, cada una basada en un conjunto de datos de validación diferente. La métrica de rendimiento se promedia sobre los tres folds para obtener una evaluación general. El objetivo en esta instancia de entrenamiento es la de elegir los hiperparámetros óptimos de cada modelo.

Para cada conjunto, se estudió la presencia de observaciones que correspondieran a todos los clústeres, con el objetivo de asegurar la representatividad al momento de construir los modelos.

Estas divisiones estratégicas garantizan una adecuada representación de los distintos períodos en cada fase del análisis. Además, se seleccionaron años consecutivos en la serie temporal para mantener la continuidad temporal y capturar posibles patrones o tendencias a lo largo del tiempo. Esta elección mejora la capacidad de los modelos para generalizar y prever eventos futuros basándose en la información histórica disponible.

3.3.3. Métricas

Las métricas de evaluación de los modelos utilizadas en el presente trabajo se presentan a continuación:

- **Exactitud (Accuracy):** La exactitud mide la proporción de predicciones correctas entre el total de predicciones.

Su fórmula es la siguiente:

$$Exactitud = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

Donde:

- Verdaderos Positivos (VP): El número de instancias positivas que fueron clasificadas correctamente como positivas.
- Verdaderos Negativos (VN): El número de instancias negativas que fueron clasificadas correctamente como negativas.
- Falsos Positivos (FP): El número de instancias negativas que fueron clasificadas incorrectamente como positivas.
- Falsos Negativos (FN): El número de instancias positivas que fueron clasificadas incorrectamente como negativas.

En otras palabras, la exactitud mide la fracción de predicciones correctas en comparación con el total de predicciones realizadas. Su debilidad radica en aquellos casos en los que existe desbalance en las clases.

- **Precisión:** Esta métrica se centra en la proporción de instancias positivas identificadas correctamente entre todas las instancias que el modelo ha predicho como positivas. En otras palabras, mide la precisión del modelo cuando clasifica una instancia como positiva. La métrica de *precisión* proporciona información sobre la calidad de las predicciones positivas del modelo, indicando qué tan precisa es la clasificación cuando el modelo predice una instancia como positiva. Es especialmente útil cuando se busca minimizar los falsos positivos, es decir, cuando la clasificación incorrecta de instancias negativas como positivas es costosa o no deseada.

  Su fórmula es la siguiente:

$$Precisión = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

- **Recall:** Esta métrica se centra en la proporción de instancias positivas que fueron identificadas correctamente entre todas las instancias que son realmente positivas. En otras palabras, mide la capacidad del modelo para capturar todas las instancias positivas. La métrica de *recall* es particularmente útil cuando es importante capturar la mayoría de las instancias positivas y minimizar los falsos negativos. Es adecuada en situaciones donde clasificar incorrectamente una instancia positiva como negativa puede tener consecuencias significativas.

  Su fórmula es la siguiente:

$$Recall = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

- **Índice Kappa:** Es una métrica de concordancia que evalúa la consistencia entre las clasificaciones observadas y las clasificaciones esperadas por azar, ajustando la probabilidad de concordancia aleatoria. Los valores del Índice Kappa se encuentran en el rango [-1, 1] donde 1 significa una concordancia perfecta entre las clasificaciones observadas y esperadas, 0 es una concordancia igual a la que se esperaría por azar y, por último, -1 es una discordancia total. El Índice Kappa puede ser útil en situaciones donde la proporción de casos positivos y negativos es desigual y en general es una métrica útil para evaluar la confiabilidad de las clasificaciones, especialmente cuando hay desbalance en las clases.

  Su fórmula es la siguiente:

$$\textit{Índice Kappa} = \frac{\text{Clasificación observada - Clasificación esperada}}{\text{1 - Clasificación esperada}}$$

- **F-score:** El *F-score* es una métrica que combina las métricas de precisión (*precision*) y *recall*. Es particularmente útil cuando hay un desequilibrio en las clases, ya que equilibra la importancia de falsos positivos y falsos negativos.

  Su fórmula es la siguiente:

$$F - score = \frac{\text{Precisión} * \text{Recall}}{\text{Precisión} + \text{Recall}}$$

El *F-score* alcanza su mejor valor en 1 (Precisión y Recall perfectas) y su peor valor en 0.

En aquellos casos en los que no existen verdaderos positivos para alguna clase, el resultado del cálculo queda indefinido.

Debido a que en el presente caso existe un desbalance de clases ya que existen agrupamientos de crónicas con un número significativamente menor que otros, se decidió hacer un análisis de todas las métricas mencionadas pero prestando especial atención a la métrica de *F-score*.

Sin embargo, en aquellos casos en los que la cantidad de valores indefinidos de la métrica no permitieran observar la evolución de los resultados en la medida que varían los hiperparámetros, se definió enfocar la atención en la métrica de *Accuracy* para evaluar la proporción de predicciones correctas en relación al total de predicciones.

3.4. Resultados

A continuación se presentarán los resultados obtenidos a partir de la aplicación de cada uno de los métodos desarrollados en la sección anterior.

En una primera sección se mostrarán los resultados de la división de la base de datos en conjuntos y en las secciones siguientes se presentarán los modelos confeccionados separados de acuerdo a las variables utilizadas en cada uno de ellos.

3.4.1. Representatividad de los clústeres en los conjuntos de registros

Tal como fue explicado, una vez dividida la base de datos en conjuntos, se estudió la presencia de observaciones que correspondieran a todos los clústeres en cada uno de ellos, con el objetivo de asegurar la representatividad al momento de construir los modelos. El resultado de dicho análisis se encuentra detallado en la Tabla 3.1 donde se observa que para todos los grupos de registros existe un número de casos representativo de cada clúster.

Clúster	Casos *holdout*	Casos testeo	Casos *fold* 1	Casos *fold* 2	Casos *fold* 3
1	4	13	10	6	6
2	21	13	16	19	20
3	25	25	18	25	21
4	10	9	13	10	13

Tabla 3.1: Cantidad de registros correspondiente a cada clúster por grupo de registros utilizados en la construcción de los modelos.

3.4.2. Modelos con variables del primer trimestre

En una primera instancia, se elaboraron 4 modelos: uno por cada método desarrollado utilizando aquellas variables que correspondieran al primer trimestre. Las mismas se detallan a continuación:

- Mediana del primer trimestre
- Máximo del primer trimestre
- Mínimo del primer trimestre

3.4.2.1. Modelo de K vecinos más cercanos

El primer paso del análisis consistió en trabajar con los registros correspondientes al conjunto de entrenamiento aplicando la técnica de validación cruzada. Gracias a ello, se obtuvieron 3 conjuntos de resultados por cada métrica a calcular.

Los resultados de dichos cálculos se pueden observar en la Figura 3.1 en la que se muestra la evolución de los resultados del cálculo de las mencionadas métricas de acuerdo a la variación del hiperparámetro a analizar que, en este caso, se trató del número de vecinos más cercanos a considerar. Como fue mencionado previamente, en aquellos casos en los que no existen verdaderos positivos para alguna categoría, la métrica de *F-score* permanece indefinida.

Asimismo, en cada una de las Figuras se adicionó el cálculo del promedio del resultado de los tres *folds* por cada valor del hiperparámetro. De esta manera se puede observar que existe un patrón común en todas las métricas a analizar que comienza con un valor bajo de las mismas cuando la cantidad de vecinos elegida es baja y que comienza a aumentar cuando se acerca al valor de cinco o seis vecinos. Una vez alcanzado aquel punto, el valor

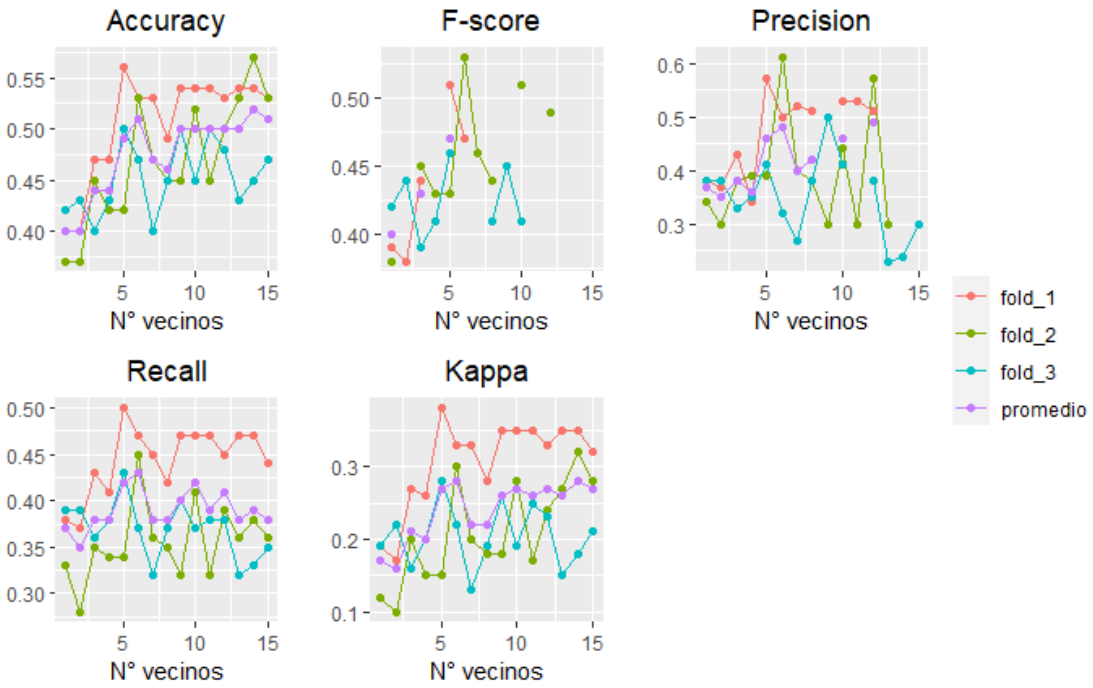

Figura 3.1: Gráficos de métricas por cada *fold* y promedio de los resultados aplicando la técnica de K vecinos más cercanos.

de las métricas desciende para volver a ascender cercano a los ocho o diez vecinos que es cuando retoma o supera el resultado anterior de cinco vecinos.

A modo de resumen, la Figura 3.2 muestra únicamente los valores promedio de los *folds* para cada una de las métricas. En este gráfico se puede observar cómo la línea que grafica el promedio representa con claridad la situación planteada en el párrafo anterior: el resultado de las métricas comienza en valores bajos y asciende en la medida que aumenta el número de vecinos hasta llegar a seis vecinos. Una vez logrado ese valor, desciende abruptamente y vuelve a retomar valores similares entre los diez y los catorce vecinos, según la métrica.

Finalmente, con el objetivo de hallar el valor más adecuado del hiperparámetro utilizado, se calcularon los resultados del modelo pero siendo aplicados sobre los mismos *folds* sobre los que había sido entrenado. Fue así que se graficaron sus resultados en la Figura 3.3 junto con aquellos obtenidos al aplicar el modelo en los folds de testeo. Por lo tanto, en dicha Figura fueron comparados los resultados del modelo en dos instancias:

- Los resultados obtenidos tras aplicar el modelo a los *folds* con los que había sido

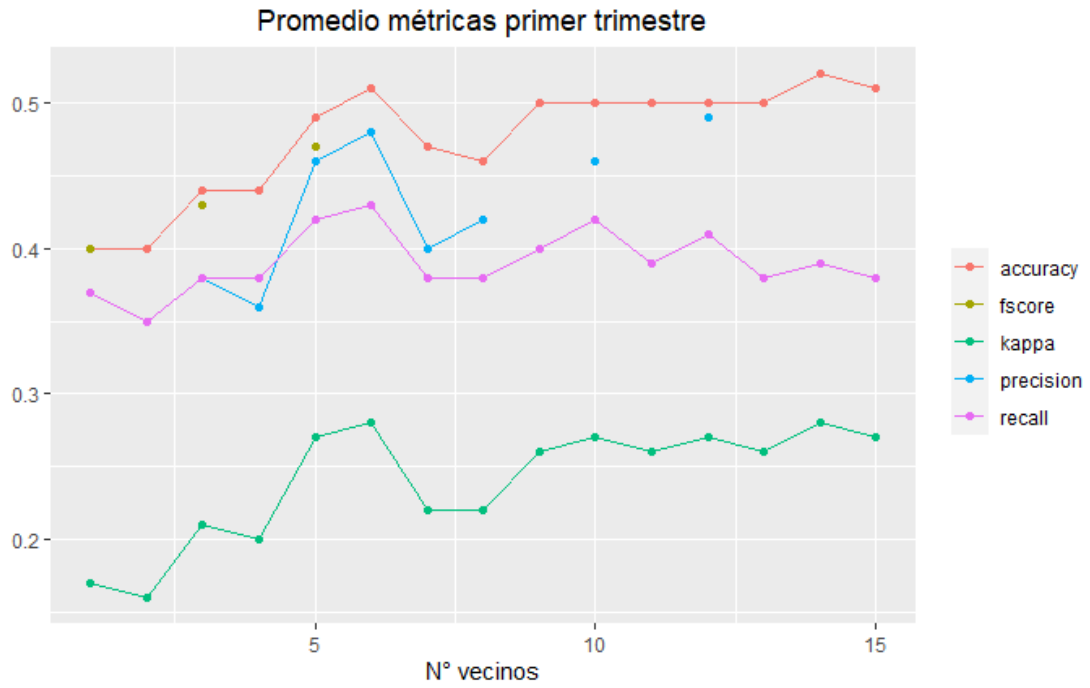

Figura 3.2: Gráfico de promedio de métricas luego de aplicada la validación cruzada con 3 *folds* para la técnica de K vecinos más cercanos.

entrenado (*train*).

- Los resultados obtenidos tras aplicar el modelo a los *folds* con los que no había sido entrenado (*test*).

En la Figura en cuestión se puede observar la evolución del valor de la métrica de *accuracy* en la medida que aumenta la cantidad de vecinos utilizados. En la misma se observa una caída en el resultado de dicha métrica para el caso del set de entrenamiento mientras que ocurre lo contrario en el set de testeo.

Con respecto a la elección de la métrica de *accuracy* para el análisis, la misma radica en la imposibilidad de utilizar *f-score* debido a la gran cantidad de resultados no definidos por la misma. En ese caso y tal como fue explicado, se consideró que *accuracy* sería la idónea.

A través del análisis de la Figura 3.3 se busca aquel valor de hiperparámetro en el que se localice un aumento del valor de *accuracy* en el conjunto de testeo y una caída local en el conjunto de entrenamiento. Ese punto se observa en la cantidad de doce vecinos más

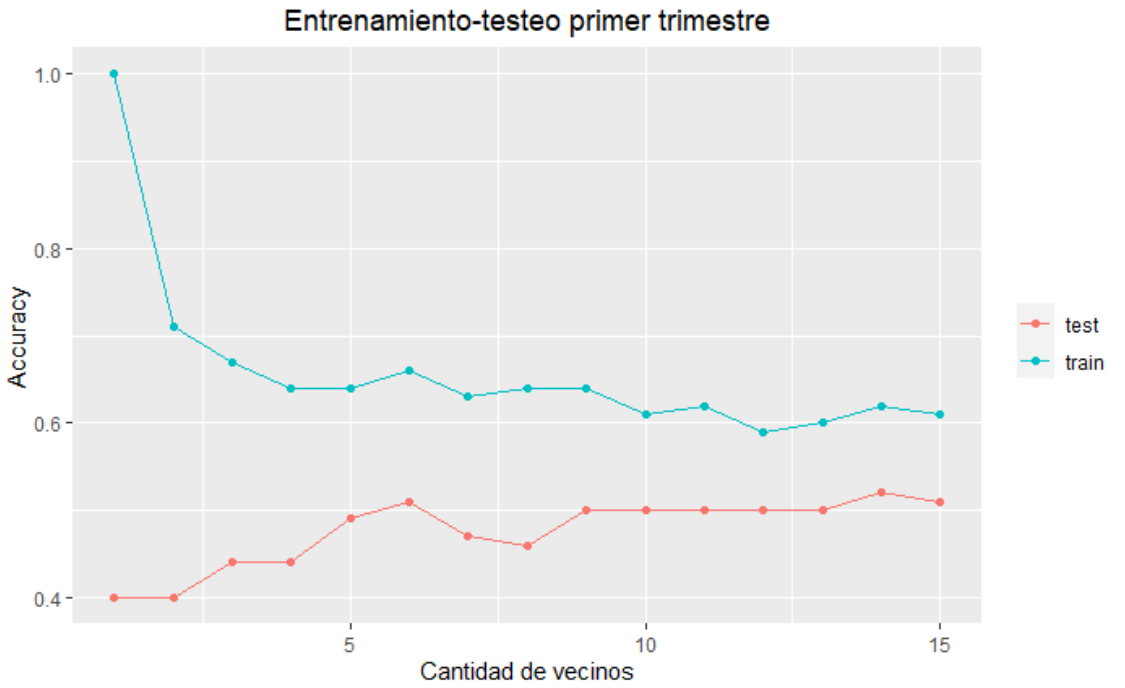

Figura 3.3: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de K vecinos más cercanos y la métrica de *accuracy* utilizando las variables correspondientes al primer trimestre.

cercanos.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.2 donde se observa que no se encuentra definido el valor de *F-Score* pero el de *accuracy* asciende a 0,55. Estos resultados serán comparados con aquellos que provengan de los siguientes modelos construidos.

Métrica	**Valor**
Accuracy	0,55
F-score	Indefinido
Precision	0,51
Recall	0,43
Kappa	0,33

Tabla 3.2: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de K vecinos más cercanos realizado con variables del primer trimestre.

3.4.2.2. Modelo de árboles de decisión

En la presente sección se describen los resultados obtenidos tras la construcción del modelo de árboles de decisión con las variables correspondientes al primer trimestre del año hidrológico.

Tal como fue realizado en la sección anterior, el primer paso del análisis consistió en trabajar con los registros correspondientes al conjunto de entrenamiento aplicando la técnica de validación cruzada. Gracias a ello, se obtuvieron 3 conjuntos de resultados por cada métrica a calcular.

Los resultados de dichos cálculos se pueden observar en la Figura 3.4 en la que se muestra la evolución de los resultados del cálculo de las mencionadas métricas de acuerdo a la variación del hiperparámetro a analizar que, en este caso, se trató del número mínimo de observaciones que deben existir en un nodo para intentar una división (*minsplit*).

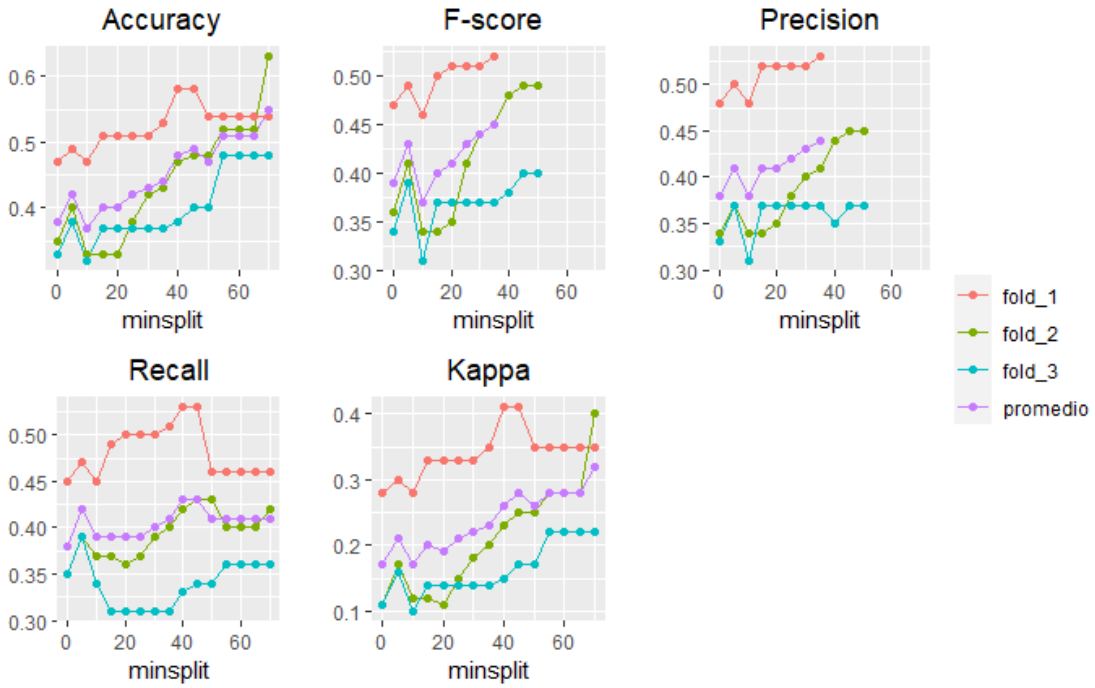

Figura 3.4: Gráficos de métricas por cada *fold* y promedio de los resultados aplicando la técnica de árboles de decisión.

Al analizar el promedio de cada una de las métricas graficadas, se puede observar que existe un patrón común que comienza con un valor bajo de las mismas cuando el *minsplit* es cercano a cero pero que asciende en la medida que éste aumenta.

A modo de resumen, la Figura 3.5 muestra únicamente los valores promedio de los *folds* para cada una de las métricas. En este gráfico se puede observar cómo la línea que grafica el promedio representa con claridad la situación planteada en el párrafo anterior: el resultado de las métricas comienza en valores bajos con un máximo relativo en 5 *minsplit* y asciende en la medida que aumenta el valor de dicho hiperparámetro.

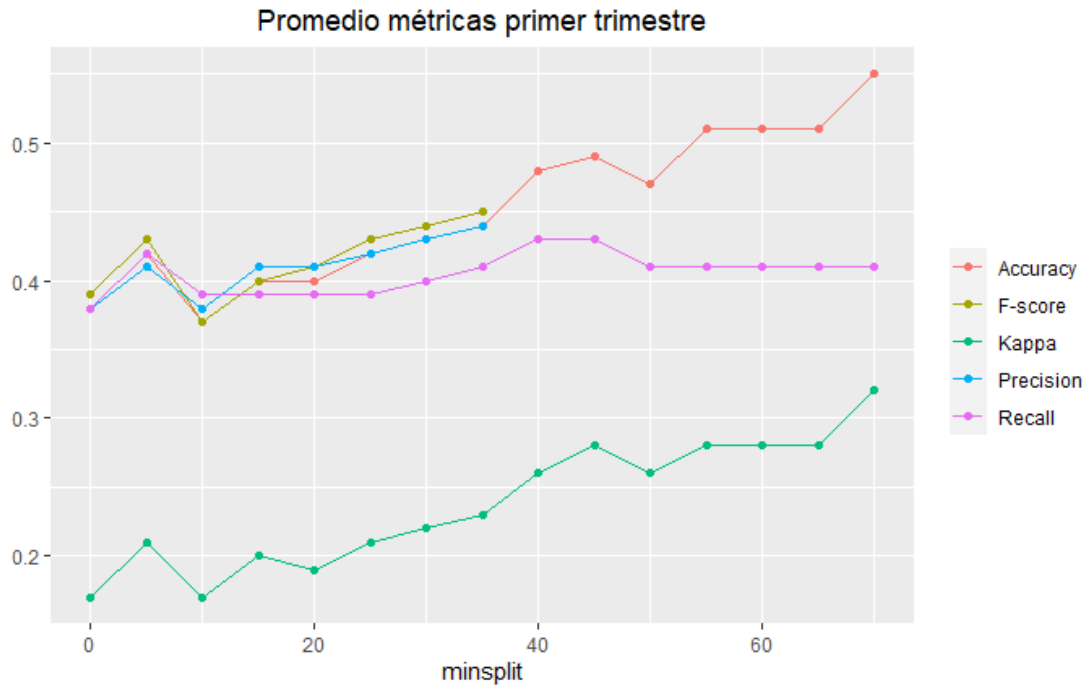

Figura 3.5: Gráfico de promedio de métricas luego de aplicada la validación cruzada con 3 *folds* para la técnica de árboles de decisión.

Finalmente, se aplicó el modelo sobre los datos de los *folds* con los que había sido entrenado y se graficaron sus resultados en la Figura 3.6 junto con los resultados del promedio de la validación cruzada en los folds de testeo. En la misma se puede observar la evolución del valor de la métrica de *accuracy* en la medida que aumenta el valor de *minsplit*. La elección de dicha métrica para el análisis radica en la imposibilidad de utilizar la métrica de *f-score* debido a la gran cantidad de resultados no definidos por la misma. En ese caso y tal como fue explicado, se consideró que *accuracy* sería la idónea.

A través del análisis de la Figura 3.6 se busca aquel valor del hiperparámetro en el que se localice un aumento relativamente significativo del valor de *accuracy* en el conjunto de testeo y una caída local también significativa en el conjunto de entrenamiento. Ese punto

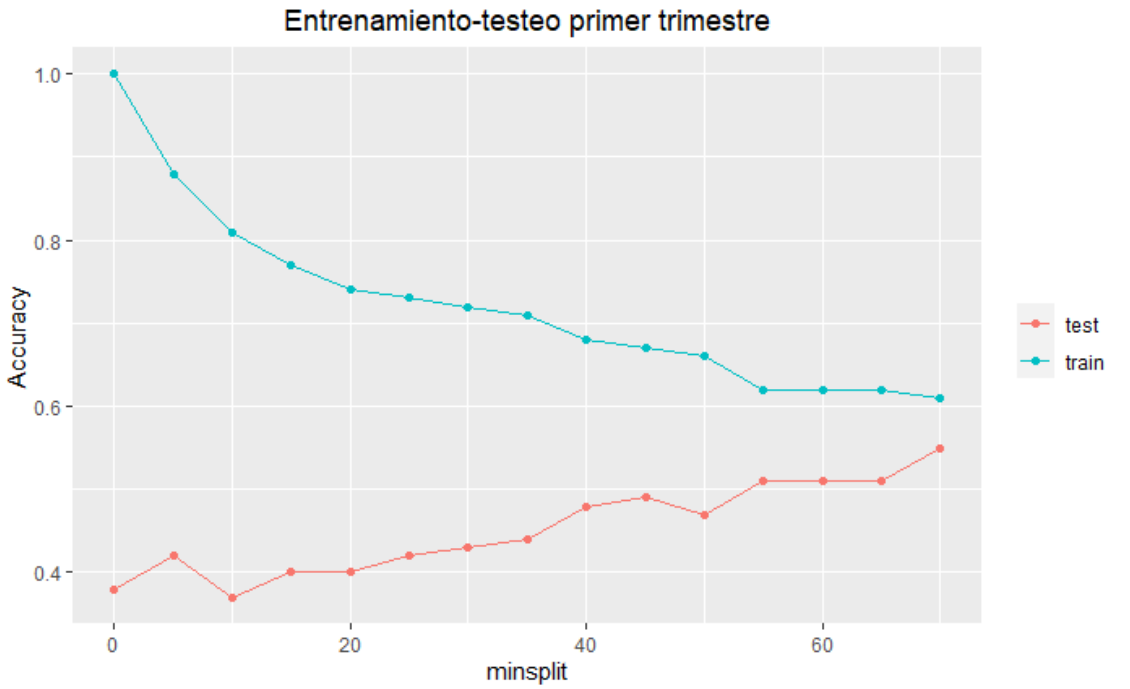

Figura 3.6: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de árboles de decisión y la métrica de *accuracy* utilizando las variables correspondientes al primer trimestre.

se observa en el valor de *minsplit* igual a cincuenta y cinco.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.3 donde se observa que no se encuentra definido el valor de *F-Score* ni de Precisión pero el de *accuracy* asciende a 0,60. Estos resultados serán comparados con los de los demás modelos construidos.

Métrica	**Valor**
Accuracy	0,60
F-score	Indefinido
Precision	Indefinido
Recall	0,47
Kappa	0,39

Tabla 3.3: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de árboles de decisión realizado con variables del primer trimestre.

3.4.2.3. Modelo de regresión logística con $alpha = 0$

En la presente sección se describen los resultados obtenidos tras la construcción del modelo de regresión logística con las variables correspondientes al primer trimestre del año hidrológico y con el valor de *alpha* igualando a cero; es decir, con penalización del tipo Ridge.

Tal como fue realizado en la sección anterior, el primer paso del análisis consistió en trabajar con los registros correspondientes al conjunto de entrenamiento aplicando la técnica de validación cruzada. Gracias a ello, se obtuvieron 3 conjuntos de resultados por cada métrica a calcular.

Para hallar el intervalo de valores de *lambda* (λ) con los cuales graficar y analizar el posible valor del hiperparámetro, se realizó un trabajo iterativo en el cual se partió de un determinado intervalo de valores que luego se fue acotando hasta llegar a aquellos que se muestran en las Figuras a continuación.

Los resultados de dichos cálculos se pueden observar en la Figura 3.7 en la que se muestra la evolución de los resultados del cálculo de las mencionadas métricas de acuerdo a la variación del hiperparámetro *lambda* (λ). Recordando la sección anterior, dicho hiperparámetro controla la fuerza de la regularización. Un valor más alto de λ implica una mayor penalización y, por lo tanto, coeficientes más pequeños. Ajustar su valor permite equilibrar la precisión del modelo con la magnitud de los coeficientes.

Debido a que el patrón de comportamiento de las métricas en la medida que *lambda* aumenta no se ve tan claro en la Figura 3.7, es de utilidad apoyar el análisis también en la Figura 3.8 que muestra únicamente los valores promedio. Allí no se ven los resultados de las métricas de *F-score* y de Precisión pero para el resto de las métricas se observa que el valor permanece constante o con un leve aumento, dejando de lado ciertos rangos en particular y que en valores de *lambda* cercanos a 0,007 así como también 0,0098 presenta valores relativamente más altos.

Finalmente, se aplicó el modelo sobre los datos de los *folds* con los que había sido entrenado y se graficaron sus resultados en la Figura 3.9 junto con los resultados del promedio de la validación cruzada en los *folds* de testeo. En la misma se puede observar

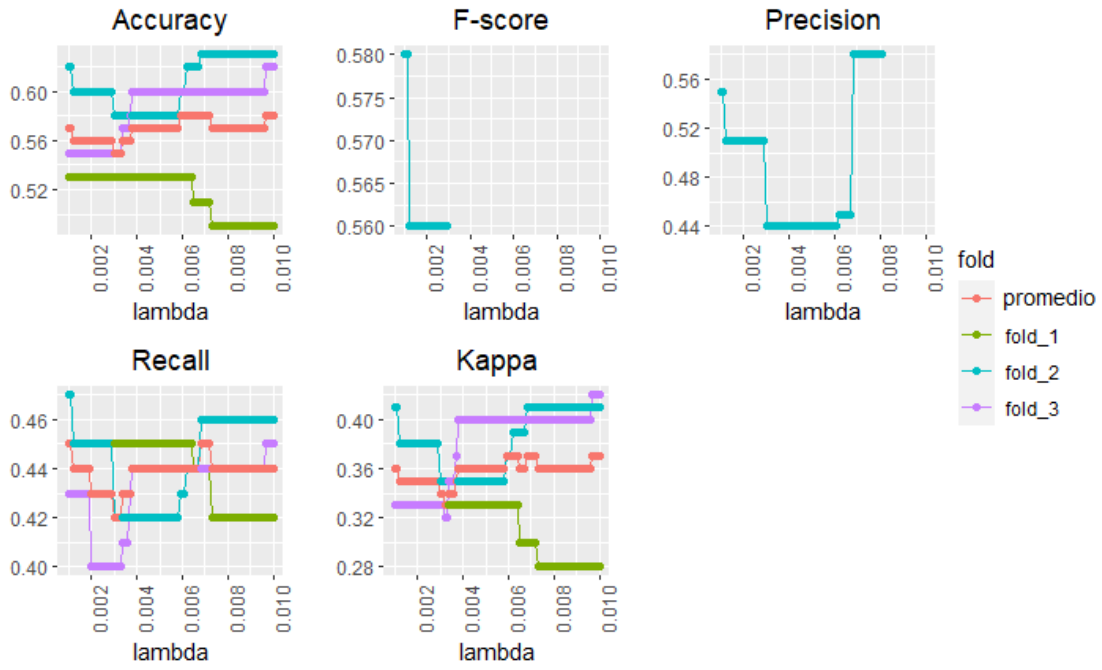

Figura 3.7: Gráficos de métricas por cada *fold* y promedio de los resultados aplicando la técnica de regresión logística con *alpha* igual a cero.

la evolución del valor de la métrica de *accuracy* en la medida que aumenta el valor de *lambda*.

A través del análisis de la Figura 3.9 se busca aquel valor del hiperparámetro en el que se localice un aumento relativamente significativo del valor de *accuracy* en el conjunto de testeo y una caída local también significativa en el conjunto de entrenamiento. Ese punto se observa en el valor de *lambda* igual a 0,0098.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.4 donde se observa que no se encuentra definido el valor de *F-Score* ni de Precisión pero el de *accuracy* asciende a 0,57. Estos resultados serán comparados con los de los demás modelos construidos.

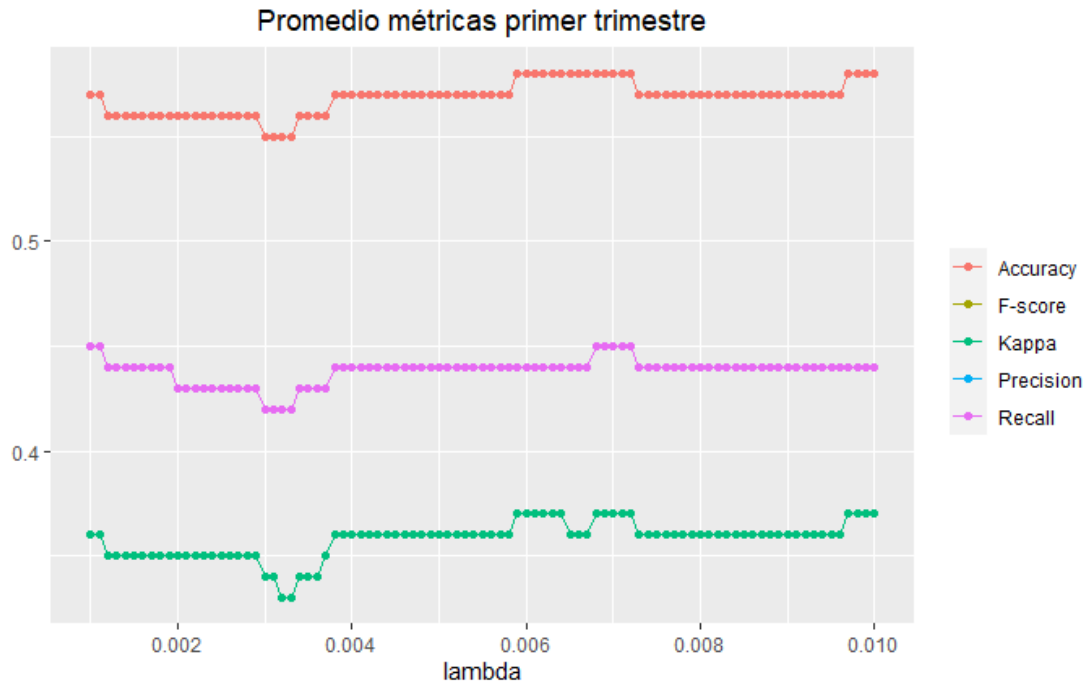

Figura 3.8: Gráfico de promedio de métricas luego de aplicada la validación cruzada con 3 *folds* para la técnica de regresión logística con *alpha* igual a cero.

Métrica	**Valor**
Accuracy	0,57
F-score	Indefinido
Precision	Indefinido
Recall	0,43
Kappa	0,34

Tabla 3.4: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística realizado con variables del primer trimestre y con el valor de *alpha* igual a cero.

3.4.2.4. Modelo de regresión logística con $alpha = 1$

En la presente sección se describen los resultados obtenidos tras la construcción del modelo de regresión logística con las variables correspondientes al primer trimestre del año hidrológico y con el valor de *alpha* igualando a uno; es decir, con penalización del tipo Lasso.

Tal como fue realizado en las secciones previas, el primer paso del análisis consistió en trabajar con los registros correspondientes al conjunto de entrenamiento aplicando la técnica de validación cruzada. Gracias a ello, se obtuvieron 3 conjuntos de resultados por cada métrica a calcular.

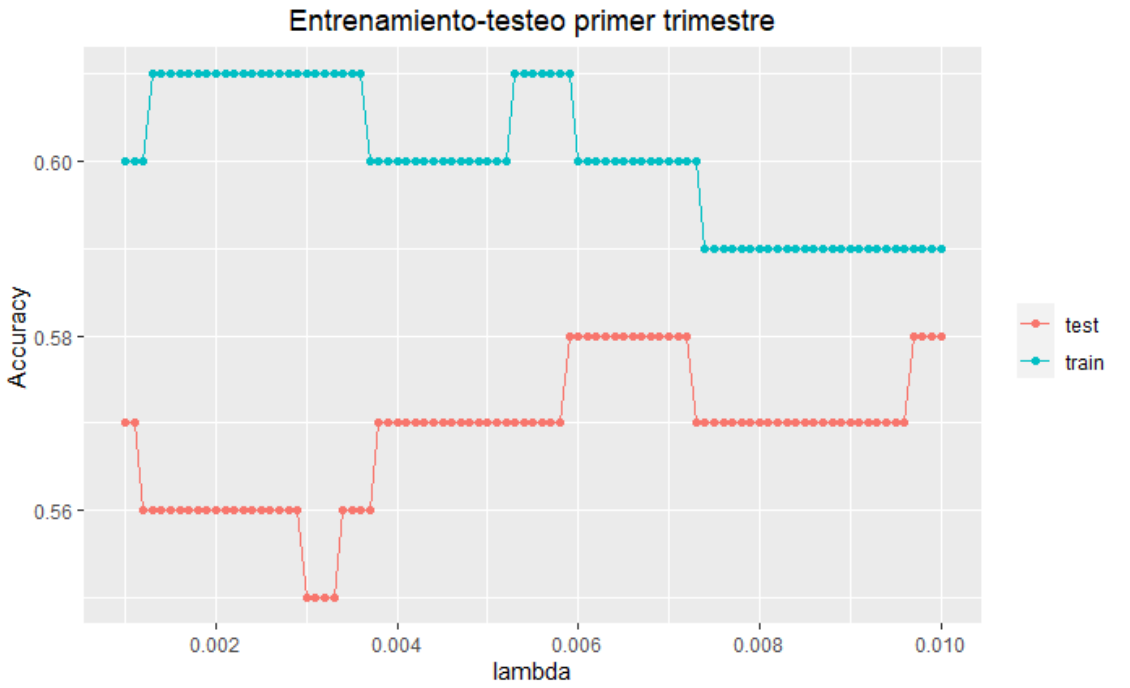

Figura 3.9: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a cero y la métrica de *accuracy* utilizando las variables correspondientes al primer trimestre.

Los resultados de dichos cálculos se pueden observar en la Figura 3.10 en la que se muestra la evolución de los resultados del cálculo de las mencionadas métricas de acuerdo a la variación del hiperparámetro *lambda* (λ).

Al analizar el promedio de cada una de las métricas graficadas, se puede observar que existe un patrón común que asciende levemente en la medida que aumenta el valor de *lambda* llegando a su valor máximo entre 0,1 y 0,2 pero que luego desciende abruptamente a partir del valor de *lambda* cercano a 0,3.

A modo de resumen, la Figura 3.11 muestra únicamente los valores promedio de los *folds* para cada una de las métricas. En este gráfico se puede observar cómo la línea que grafica el promedio representa con claridad la situación planteada en el párrafo anterior: el resultado de las métricas comienza en un leve ascenso con un máximo entre 0,1 y 0,2 para luego descender de manera repentina.

Finalmente, se aplicó el modelo sobre los datos de los *folds* con los que había sido entrenado y se graficaron sus resultados en la Figura 3.12 junto con los resultados del

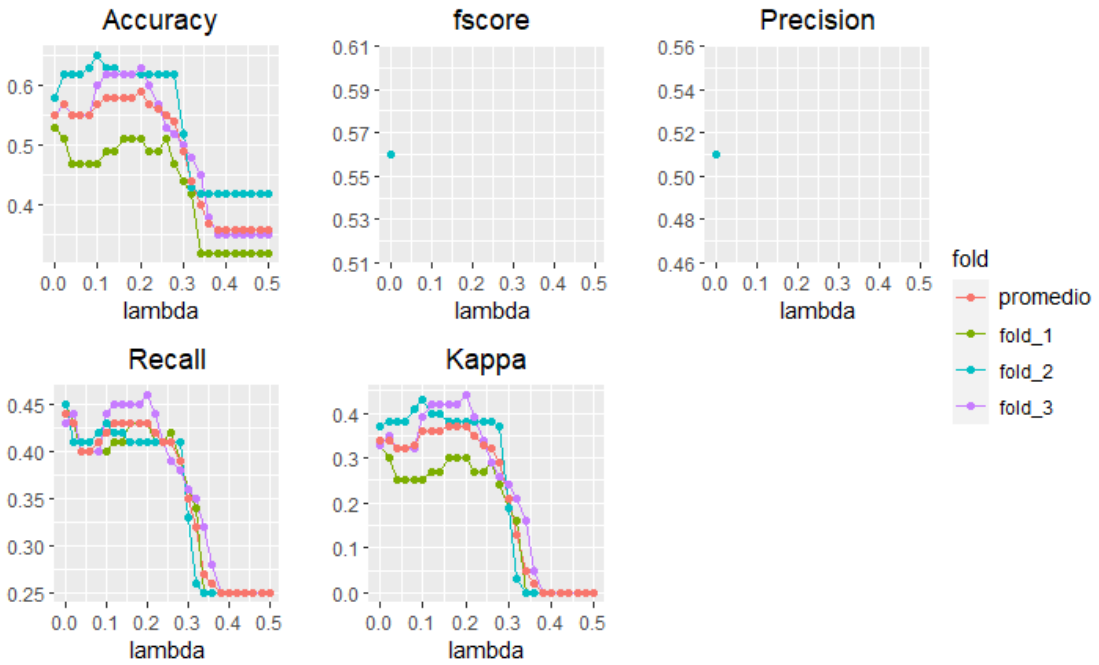

Figura 3.10: Gráficos de métricas por cada *fold* y promedio de los resultados aplicando la técnica de regresión logística con *alpha* igual a uno.

promedio de la validación cruzada en los folds de testeo. En la misma se puede observar la evolución del valor de la métrica de *accuracy* en la medida que aumenta el valor de *lambda*.

A través del análisis de la Figura 3.12 se busca aquel valor del hiperparámetro en el que se localice un aumento significativo del valor de *accuracy* en el conjunto de testeo y una caída local en el conjunto de entrenamiento. Ese punto se observa en el valor de *lambda* igual a 0,14.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.5 donde se observa que no se encuentra definido el valor de *F-Score* ni de Precisión pero el de *accuracy* asciende a 0,53.

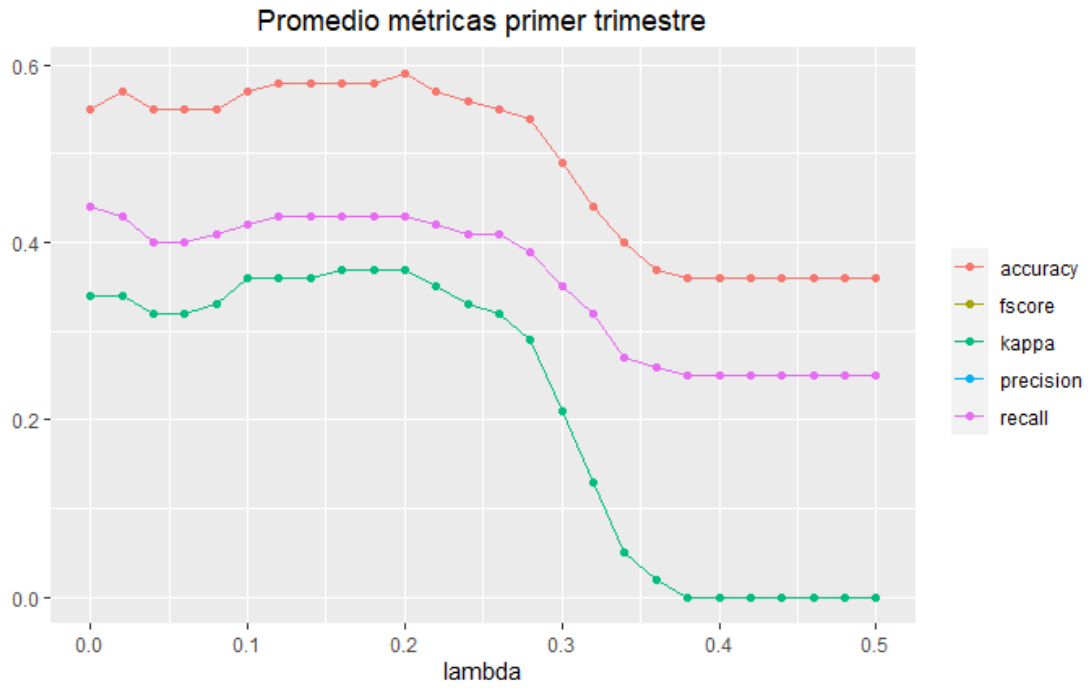

Figura 3.11: Gráfico de promedio de métricas luego de aplicada la validación cruzada con 3 *folds* para la técnica de regresión logística con *alpha* igual a uno.

Métrica	**Valor**
Accuracy	0,53
F-score	Indefinido
Precision	Indefinido
Recall	0,39
Kappa	0,28

Tabla 3.5: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística realizado con variables del primer trimestre y con el valor de *alpha* igual a uno.

3.4.2.5. Resumen de resultados de los modelos con variables del primer trimestre

Para finalizar el análisis de los modelos que utilizan las variables correspondientes al primer trimestre, en la Tabla 3.6 se procedió a mostrar los resultados de los cuatro modelos para facilitar su comparación.

En la misma se observa que, para este conjunto de datos, el modelo de árboles de decisión es el que presenta mejores resultados. Sin embargo, cabe destacar que el modelo de K vecinos más cercanos es el que tiene menor cantidad de métricas no definidas.

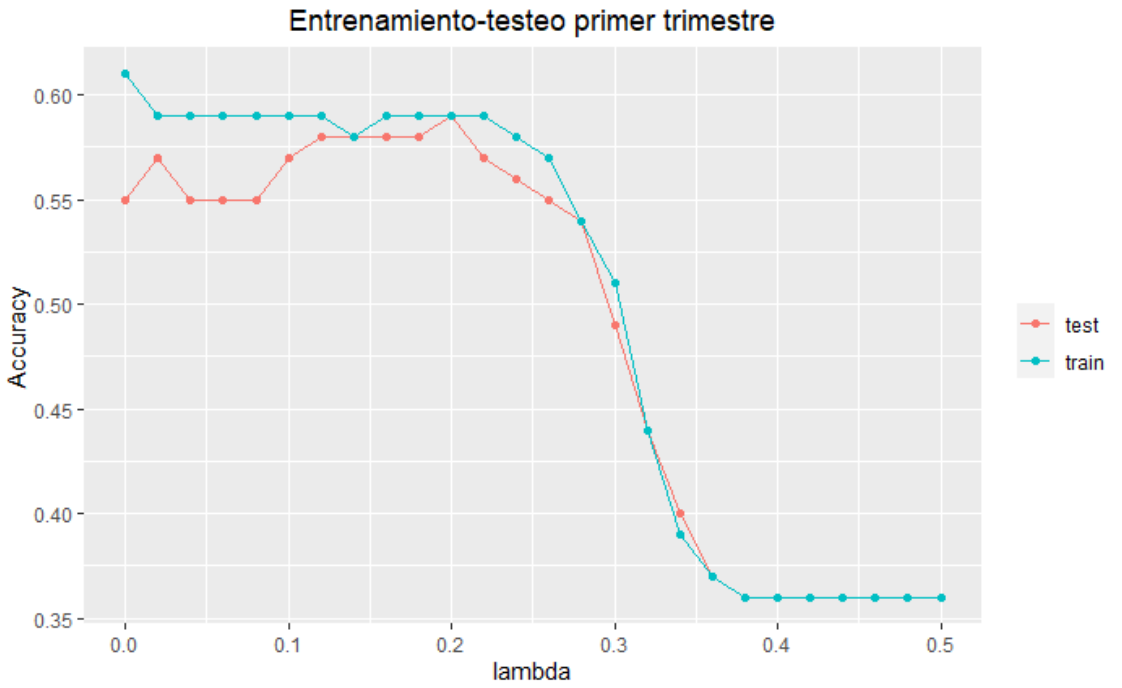

Figura 3.12: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a uno y la métrica de *accuracy* utilizando las variables correspondientes al primer trimestre.

Métrica	**KNN**	**Árboles**	**Regresión 0**	**Regresión 1**
Accuracy	0,55	0,60	0,57	0,53
F-score	Indefinido	Indefinido	Indefinido	Indefinido
Precision	0,51	Indefinido	Indefinido	Indefinido
Recall	0,43	0,47	0,43	0,39
Kappa	0,33	0,39	0,34	0,28

Tabla 3.6: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para los cuatro modelos realizados sobre el conjunto de datos correspondientes a las variables del primer trimestre del año hidrológico.

3.4.3. Modelos con variables del segundo trimestre

Seguidamente, aplicando la misma metodología llevada adelante en la sección previa, se elaboraron 4 nuevos modelos que, en este caso, adicionaron las variables del segundo trimestre en su confección. Es decir que se utilizaron las siguientes variables:

- Mediana del primer trimestre
- Mediana del segundo trimestre
- Máximo del primer trimestre

- Máximo del segundo trimestre
- Mínimo del primer trimestre
- Mínimo del segundo trimestre

3.4.3.1. Modelo de K vecinos más cercanos

Después de aplicar la técnica de validación cruzada junto con el algoritmo de K vecinos más cercanos, se han obtenido resultados que han sido analizados utilizando la métrica de *F-score*. La elección de esta métrica se justifica en secciones anteriores de este estudio. Es relevante destacar que aún quedan valores del hiperparámetro sin resultados para la métrica seleccionada. No obstante, los resultados obtenidos hasta el momento proporcionan información valiosa para el estudio en curso, permitiendo extraer conclusiones significativas.

En la Figura 3.13 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

Primeramente, se puede observar la evolución del valor de la métrica de *F-score* en la medida que aumenta la cantidad de vecinos utilizados. Para el caso del conjunto de entrenamiento, se observa una caída del valor de la métrica estabilizándose en un valor cercano a un *F-score* de 0,65 a partir de los 8 vecinos más cercanos en adelante. Por otro lado, para el caso de los resultados obtenidos en el conjunto de testeo, se observa inicialmente un ascenso en el valor de la métrica para el caso de 2 vecinos más cercanos pero que luego asciende y se estabiliza en un valor de métrica de *F-score* cercano a 0,6.

Seguidamente, a través del análisis del gráfico, se busca aquel valor de hiperparámetro en el que se localice un aumento del valor de *F-score* en el conjunto de testeo y una caída local en el conjunto de entrenamiento. Uno de estos puntos se observa en la cantidad de ocho vecinos más cercanos.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

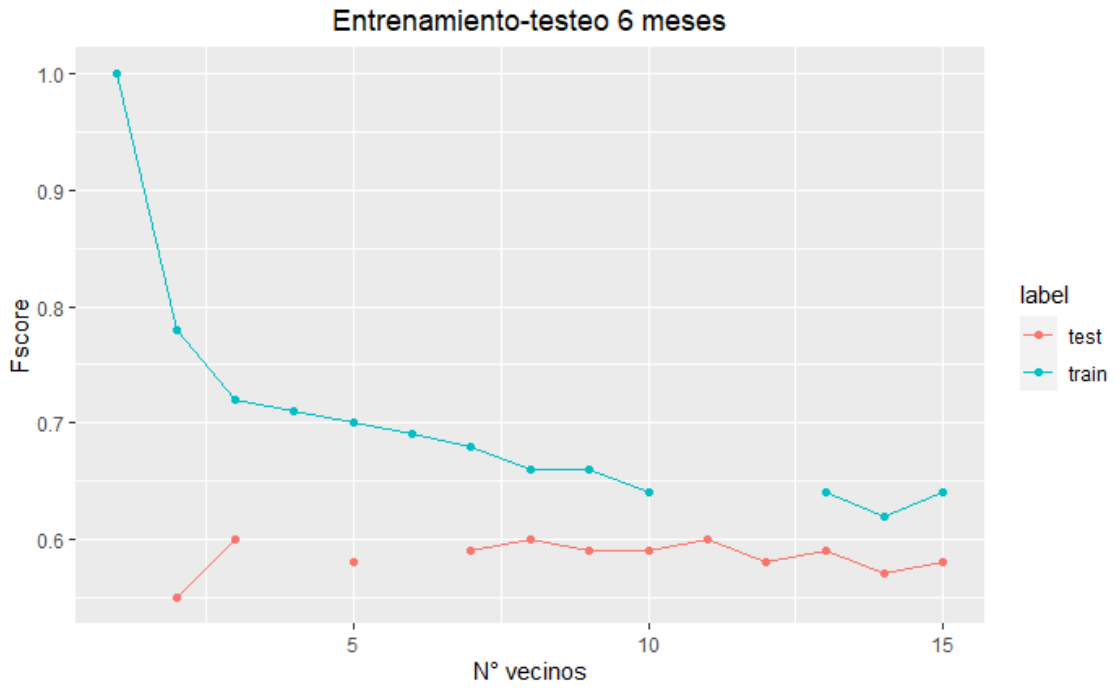

Figura 3.13: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de K vecinos más cercanos y la métrica de *F-score* utilizando las variables correspondientes al primer y segundo trimestre.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.7 donde se observa que el valor de *F-Score* asciende a 0,64 y que el de *accuracy* corresponde a 0,65; es decir, 0,10 puntos por encima que el valor obtenido en la misma métrica pero en el modelo de K vecinos modelado con 3 meses de información.

Métrica	**Valor**
Accuracy	0,65
F-score	0,64
Precision	0,63
Recall	0,57
Kappa	0,50

Tabla 3.7: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de K vecinos más cercanos realizado con variables del primer y segundo trimestre.

3.4.3.2. Modelo de árboles de decisión

Después de aplicar la técnica de validación cruzada junto con el algoritmo de árboles de decisión, se han analizado sus resultados correspondientes a la métrica de *F-score*. En este caso, se han obtenido datos de la misma hasta el valor de 45 *minsplit*.

En la Figura 3.14 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

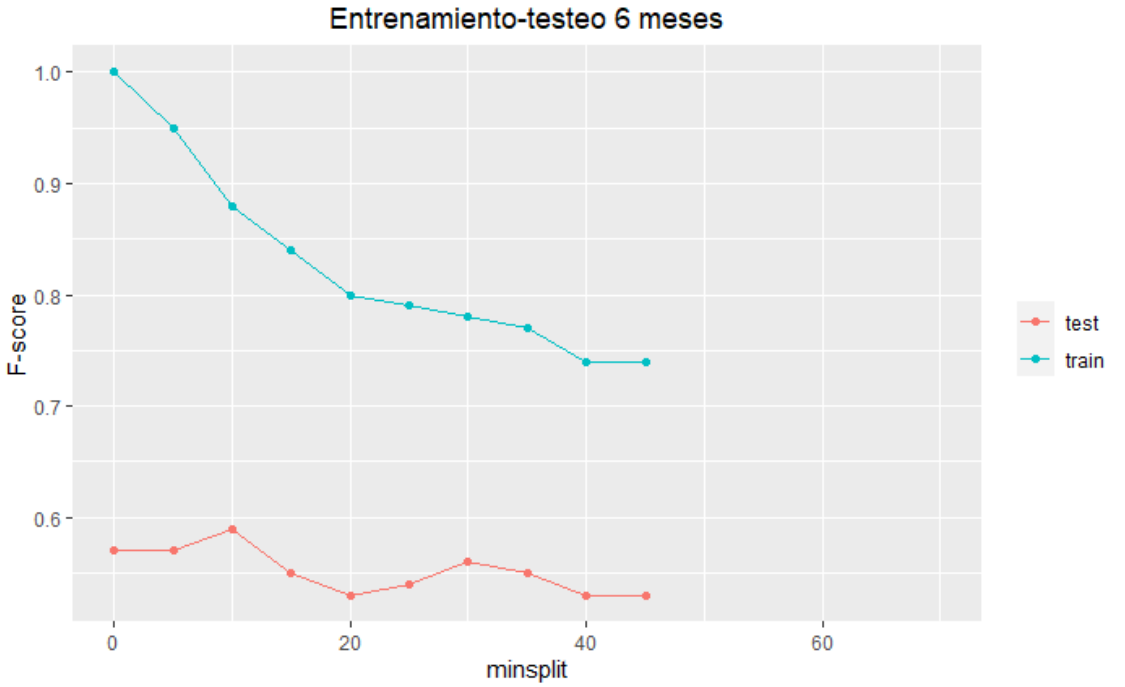

Figura 3.14: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de árboles de decisión y la métrica de *F-score* utilizando las variables correspondientes al primer y segundo trimestre.

Primeramente, se puede observar la evolución del valor de la métrica de *F-score* en la medida que aumenta el valor de *minsplit*. Para el caso del conjunto de entrenamiento, se observa una caída pronunciada hasta 20 *minsplit*. A partir de ese valor, continúa el descenso aunque menos abrupto. Asimismo, en el caso de los resultados del conjunto de testeo, los mismos oscilan en un valor cercano a un *F-score* de 0,55.

A través del análisis del gráfico, se buscó aquel valor de hiperparámetro en el que se localice un aumento local del valor de *F-score* en el conjunto de testeo y un punto en el que se viera que la curva se encontrara en descenso. Uno de estos puntos se observa en la cantidad de 30 *minsplit*.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.8 donde se observa que el valor de *F-Score* asciende a 0,65 así como también el valor de *accuracy*; es decir, 0,05 puntos por encima del valor obtenido en la misma métrica pero en el modelo de árboles de decisión modelado con 3 meses de información.

Métrica	**Valor**
Accuracy	0,65
F-score	0,65
Precision	0,60
Recall	0,58
Kappa	0,50

Tabla 3.8: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de árboles de decisión realizado con variables del primer y segundo trimestre.

3.4.3.3. Modelo de regresión logística con $alpha = 0$

Continuando con el análisis utilizando las variables de los primeros seis meses de cada crónica y luego de aplicar la técnica de validación cruzada junto con el algoritmo de regresión logística con $alpha = 0$, se han analizado sus resultados correspondientes a la métrica de *F-score*.

En la Figura 3.15 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

En este caso se ha acotado el valor de *lambda* al rango comprendido entre 0,003 y 0,008 ya que era aquel en el que se observaron los valores más altos de la métrica.

Mencionado esto, se puede observar la evolución del valor de la métrica de *F-score* en la medida que varía el valor de *lambda* tanto para la curva de resultados del conjunto de entrenamiento como para la de testeo. Para el caso del conjunto de entrenamiento, se observa que el *F-score* ronda un valor cercano a 0,73 durante todo el intervalo analizado, a excepción de determinados casos en los cuales fluctúa. Asimismo, en el caso de los resultados del conjunto de testeo, los mismos oscilan en un valor cercano a un *F-score* de 0,71.

Sin embargo, se observa que el punto en el que *lambda* tiene un valor de 0,006 el

Figura 3.15: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a cero y la métrica de *F-score* utilizando las variables correspondientes al primer y segundo trimestre.

conjunto de entrenamiento tiene un valor de 0,73 mientras que el conjunto de testeo presenta un máximo local en 0,72. Por lo tanto, ese fue el valor de *lambda* elegido para entrenar el modelo con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.9 donde se observa que el valor de *F-Score* asciende a 0,70 mientras que el de *accuracy* se encuentra en 0,68; es decir, 0,09 puntos por encima del valor obtenido en la misma métrica pero en el modelo de regresión logística con *alpha* igual a cero modelado con 3 meses de información.

3.4.3.4. Modelo de regresión logística con $alpha = 1$

Finalmente, luego de aplicar la técnica de validación cruzada junto con el algoritmo de regresión logística pero con $alpha = 1$, se han analizado sus resultados correspondientes a la métrica de *F-score*.

Métrica	Valor
Accuracy	0,68
F-score	0,70
Precision	0,68
Recall	0,67
Kappa	0,57

Tabla 3.9: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística con *alpha* igual a cero modelado con variables del primer y segundo trimestre.

En la Figura 3.16 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

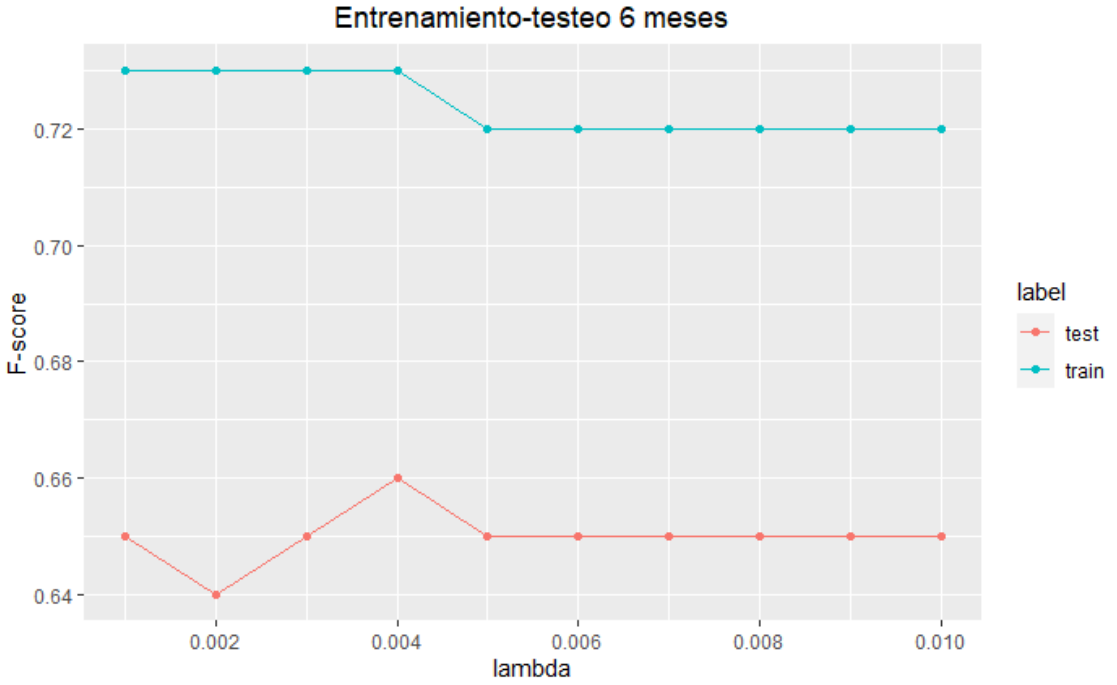

Figura 3.16: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a 1 y la métrica de *F-score* utilizando las variables correspondientes al primer y segundo trimestre.

En este caso se ha acotado el valor de *lambda* al rango comprendido entre 0,001 y 0,010 ya que era aquel en el que se observaron los valores más altos de la métrica.

Mencionado esto, se puede observar que la evolución del valor de la métrica de *F-score* se mantiene estable en la medida que varía el valor de *lambda* tanto para la curva de resultados del conjunto de entrenamiento como para la de testeo. Sin embargo, se observa

un descenso local en la curva de entrenamiento en el valor de *lambda* igual a 0,005 y aquel fue el elegido para entrenar el modelo con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.10 donde se observa que el valor de *F-Score* asciende a 0,69 mientras que el de *accuracy* se encuentra en 0,67; es decir, 0,14 puntos por encima del valor obtenido en la misma métrica pero en el modelo de regresión logística con *alpha* igual a uno modelado con 3 meses de información.

Métrica	**Valor**
Accuracy	0,67
F-score	0,69
Precision	0,67
Recall	0,65
Kappa	0,54

Tabla 3.10: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística con *alpha* igual a uno modelado con variables del primer y segundo trimestre.

3.4.3.5. Resumen de resultados de los modelos con variables del primer y segundo trimestre

Para finalizar el análisis de los modelos que utilizan las variables correspondientes al primer y segundo trimestre, en la Tabla 3.11 se procedió a mostrar los resultados de los cuatro modelos para facilitar su comparación.

Métrica	**KNN**	**Árboles**	**Regresión 0**	**Regresión 1**
Accuracy	0,65	0,65	0,68	0,67
F-score	0,64	0,65	0,70	0,69
Precision	0,63	0,60	0,68	0,67
Recall	0,57	0,58	0,67	0,65
Kappa	0,50	0,50	0,57	0,54

Tabla 3.11: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para los cuatro modelos realizados sobre el conjunto de datos correspondientes a las variables del primer y segundo trimestre del año hidrológico.

En la misma se observa que, a diferencia de los modelos analizados con variables del primer trimestre, los modelos de regresión logística son los que presentan mejores resultados.

3.4.4. Modelos con variables del tercer trimestre

A continuación se elaboraron 4 nuevos modelos que adicionaron las variables del tercer trimestre en su confección. Es decir que se utilizaron las siguientes variables:

- Mediana del primer trimestre
- Mediana del segundo trimestre
- Mediana del tercer trimestre
- Máximo del primer trimestre
- Máximo del segundo trimestre
- Máximo del tercer trimestre
- Mínimo del primer trimestre
- Mínimo del segundo trimestre
- Mínimo del tercer trimestre

3.4.4.1. Modelo de K vecinos más cercanos

Después de aplicar la técnica de validación cruzada junto con el algoritmo de K vecinos más cercanos, se han obtenido resultados que han sido analizados utilizando la métrica de *F-score*.

En la Figura 3.17 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

Primeramente, se puede observar la evolución del valor de la métrica de *F-score* en la medida que aumenta la cantidad de vecinos utilizados. Para el caso del conjunto de entrenamiento, se observa una caída del valor de la métrica estabilizándose en un valor que ronda el intervalo de entre 0,80 y 0,85 de la métrica de *F-score*. Por otro lado, para el caso de los resultados obtenidos en el conjunto de testeo, se observa inicialmente un ascenso

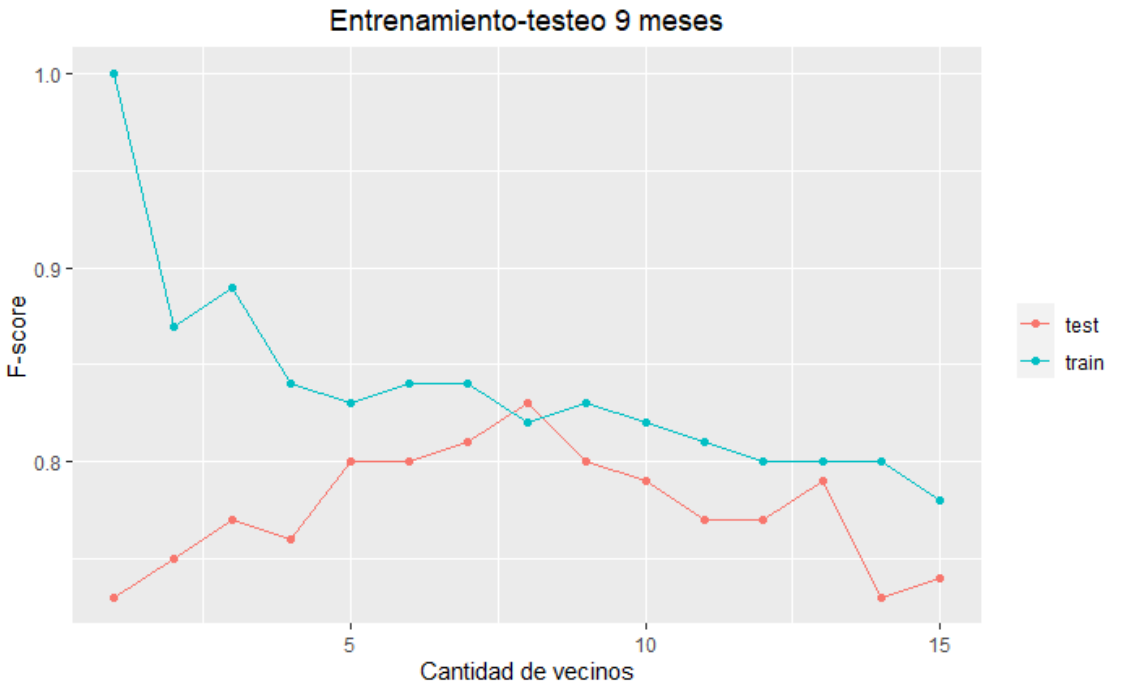

Figura 3.17: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de K vecinos más cercanos y la métrica de *F-score* utilizando las variables correspondientes al primer, segundo y tercer trimestre.

en el valor de la métrica en la medida que aumenta el número de vecinos estabilizándose en valores cercanos al intervalo comprendido entre 0,75 y 0,80 de *F-score*.

Seguidamente, a través del análisis del gráfico, se busca aquel valor de hiperparámetro en el que se localice un aumento del valor de *F-score* en el conjunto de testeo y una caída local en el conjunto de entrenamiento. Respondiendo a dicho requerimiento, existe un caso cuando la cantidad de vecinos es igual a ocho en el que no sólo sucede lo planteado sino que, además, el valor en el conjunto de testeo es máyor que en el conjunto de entrenamiento.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.12 donde se observa que tanto el valor de *F-Score* como el de *accuracy* ascienden a 0,77; es decir, más de 0,12 puntos por encima del valor obtenido en la misma métrica pero en el modelo de K vecinos modelado con 6 meses de información.

Métrica	Valor
Accuracy	0,77
F-score	0,77
Precision	0,73
Recall	0,72
Kappa	0,67

Tabla 3.12: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de K vecinos más cercanos realizado con variables del primer, segundo y tercer trimestre.

3.4.4.2. Modelo de árboles de decisión

Después de aplicar la técnica de validación cruzada junto con el algoritmo de árboles de decisión, se han analizado sus resultados correspondientes a la métrica de *F-score*. Al igual que en el caso de los modelos que incluyen variables de 6 meses, en este caso se han obtenido resultados hasta el valor de 45 *minsplit*.

En la Figura 3.18 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

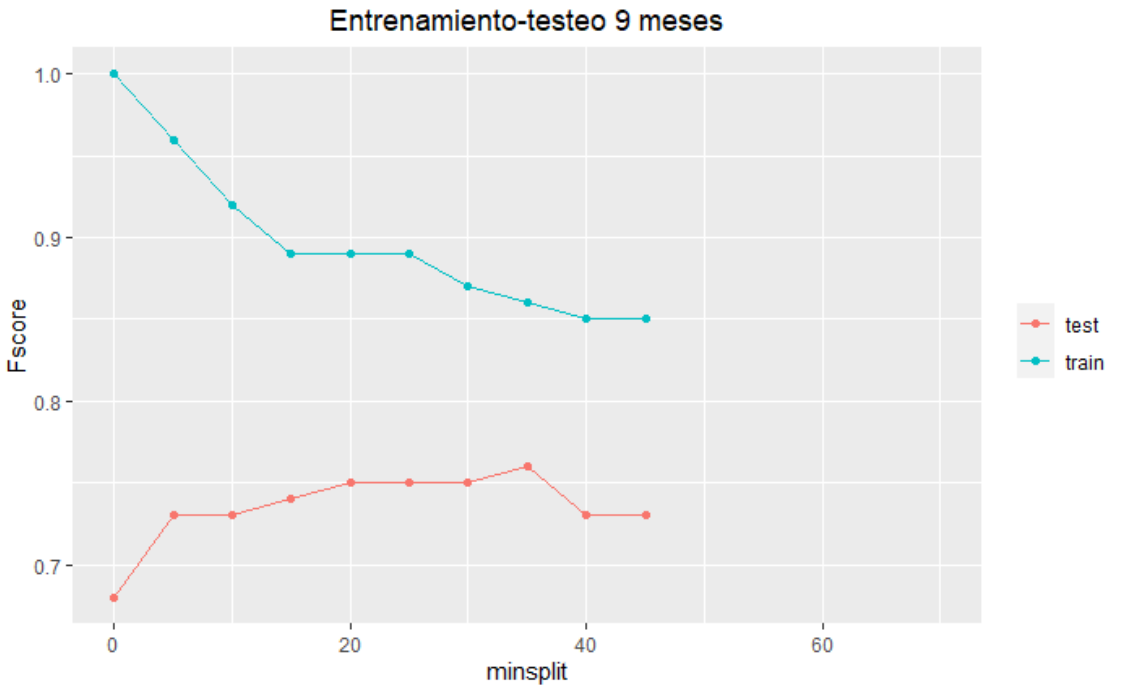

Figura 3.18: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de árboles de decisión y la métrica de *F-score* utilizando las variables correspondientes al primer, segundo y tercer trimestre.

Primeramente, se puede observar la evolución del valor de la métrica de *F-score* en la medida que aumenta el valor de *minsplit*. Para el caso del conjunto de entrenamiento, se observa una caída pronunciada hasta 15 *minsplit*. A partir de ese valor, el descenso comienza a ser más suave. Asimismo, en el caso de los resultados del conjunto de testeo, los mismos oscilan en un valor cercano a un *F-score* de 0,75.

A través del análisis del gráfico, se buscó aquel valor de hiperparámetro en el que se localice un aumento local del valor de *F-score* en el conjunto de testeo y un punto en el que se viera que la curva se encontrara en descenso. Uno de estos puntos se observa en la cantidad de 35 *minsplit*.

Una vez definido el valor del hiperparámetro, se procedió a entrenar el modelo con dicho valor y con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.13 donde se observa que el valor de *F-Score* asciende a 0,76 y el de *accuracy* a 0,75; es decir, aproximadamente 0,10 puntos por encima del valor obtenido en la misma métrica pero en el modelo de árboles de decisión modelado con 6 meses de información.

Métrica	**Valor**
Accuracy	0,75
F-score	0,76
Precision	0,71
Recall	0,72
Kappa	0,66

Tabla 3.13: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de árboles de decisión realizado con variables del primer, segundo y tercer trimestre.

3.4.4.3. Modelo de regresión logística con $alpha = 0$

Continuando con el análisis utilizando las variables de los primeros nueve meses de cada crónica y luego de aplicar la técnica de validación cruzada junto con el algoritmo de regresión logística con $alpha = 0$, se han analizado sus resultados correspondientes a la métrica de *F-score*.

En la Figura 3.19 se puede observar el gráfico en el que se presentan las curvas de

entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

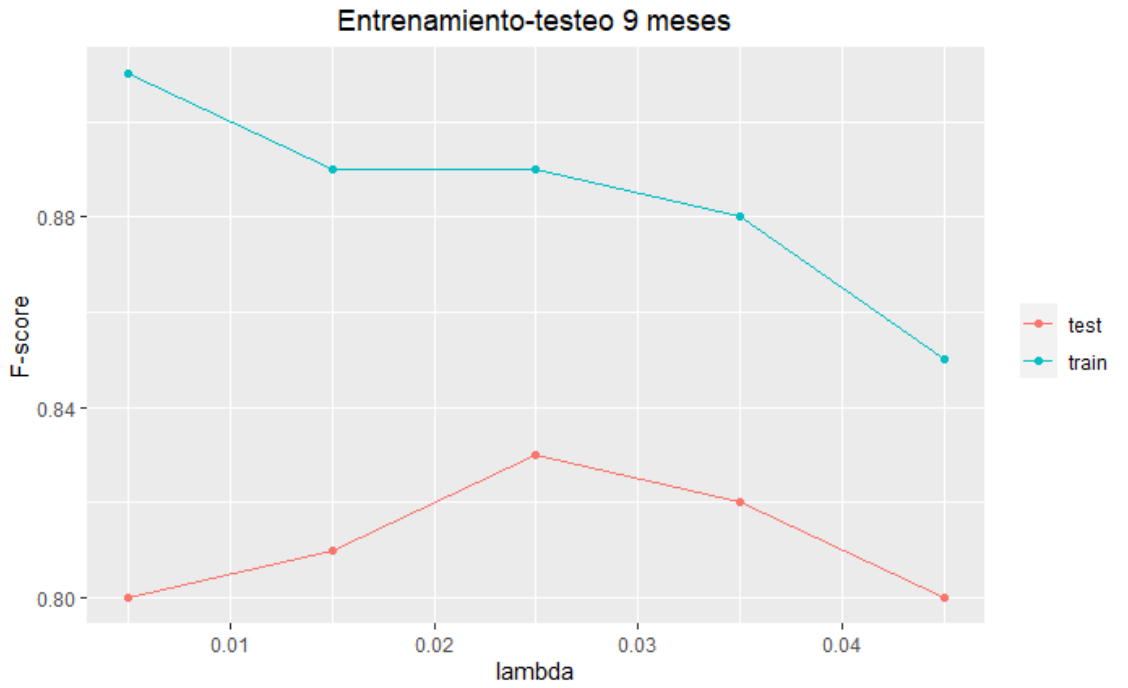

Figura 3.19: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a cero y la métrica de *F-score* utilizando las variables correspondientes al primer, segundo y tercer trimestre.

En este caso se ha acotado el valor de *lambda* al rango comprendido entre 0,005 y 0,045 ya que era aquel en el que se observaron los valores más altos de la métrica.

Mencionado esto, se puede observar la evolución del valor de la métrica de *F-score* en la medida que varía el valor de *lambda* tanto para la curva de resultados del conjunto de entrenamiento como para la de testeo. Para el caso del conjunto de entrenamiento, se observa que el *F-score* desciende durante todo el intervalo en la medida que asciende el valor de *lambda*. Asimismo, en el caso de los resultados del conjunto de testeo, los mismos ascienden hasta llegar a su punto máximo cuando *lambda* es igual a 0,025. Por lo tanto, ese fue el valor de *lambda* elegido para entrenar el modelo con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.14 donde se observa que tanto el valor de *F-Score* como el de

accuracy ascienden a 0,85; es decir, 0,15 puntos por encima del valor obtenido en la misma métrica pero en el modelo de regresión logística con *alpha* igual a cero modelado con 6 meses de información.

Métrica	**Valor**
Accuracy	0,85
F-score	0,85
Precision	0,83
Recall	0,81
Kappa	0,79

Tabla 3.14: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística con *alpha* igual a cero modelado con variables del primer, segundo y tercer trimestre.

3.4.4.4. Modelo de regresión logística con $alpha = 1$

Finalmente, luego de aplicar la técnica de validación cruzada junto con el algoritmo de regresión logística pero con $alpha = 1$, se han analizado sus resultados correspondientes a la métrica de *F-score*.

En la Figura 3.20 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

En este caso se ha acotado el valor de *lambda* al rango comprendido entre 0,01 y 0,10 ya que era aquel en el que se observaron los valores más altos de la métrica.

Mencionado esto, se puede observar que la evolución del valor de la métrica de *F-score* en el conjunto de entrenamiento desciende en la medida que aumenta el valor de *lambda*. Por el contrario, la curva de testeo presenta oscilaciones con máximos locales en ciertos valores de *lambda* como es el caso de *lambda* igual a 0,08. Fue este último el valor elegido con el cual entrenar el modelo con el set de entrenamiento completo para luego testear en el set de testeo.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.15 donde se observa que el valor de *F-Score* asciende a 0,82 mientras que el de *accuracy* se encuentra en 0,83; es decir, más de 0,13 puntos por encima del valor obtenido en la misma métrica pero en el modelo de regresión logística con *alpha* igual a

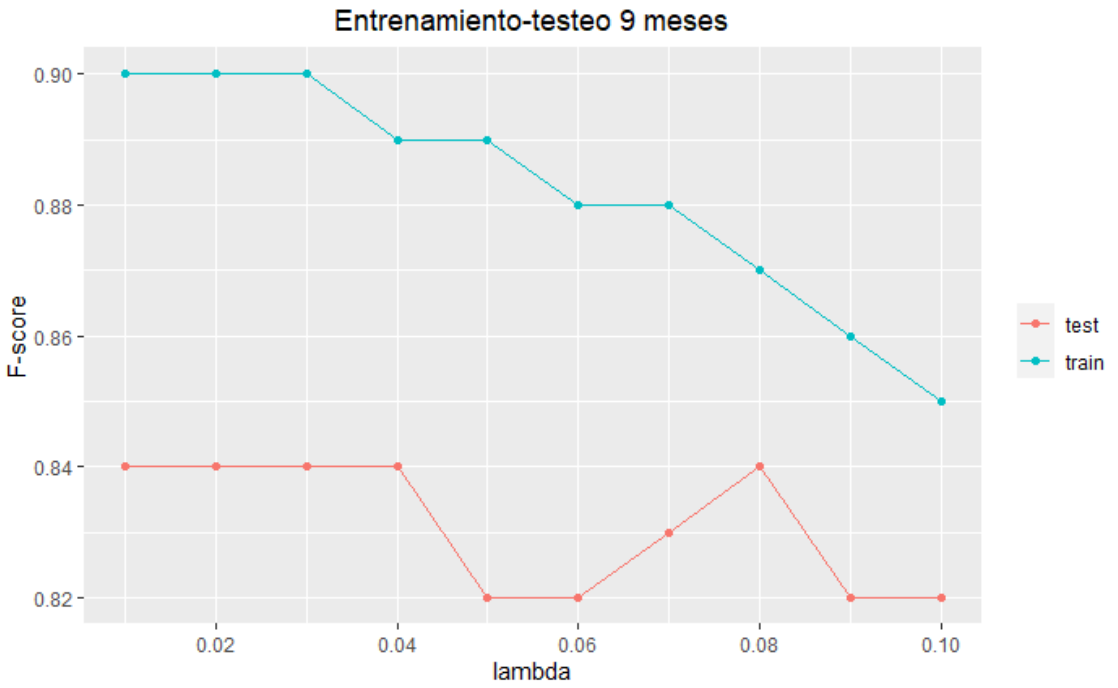

Figura 3.20: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a 1 y la métrica de *F-score* utilizando las variables correspondientes al primer, segundo y tercer trimestre.

uno modelado con 6 meses de información.

Métrica	**Valor**
Accuracy	0,83
F-score	0,82
Precision	0,82
Recall	0,76
Kappa	0,76

Tabla 3.15: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística con *alpha* igual a uno modelado con variables del primer, segundo y tercer trimestre.

3.4.4.5. Resumen de resultados de los modelos con variables del primer al tercer trimestre

Para finalizar el análisis de los modelos que utilizan las variables correspondientes a los tres primeros trimestres de cada crónica, en la Tabla 3.16 se procedió a mostrar los resultados de los cuatro modelos para facilitar su comparación.

En la misma se observa que, al igual que en los modelos analizados con variables de

Métrica	KNN	Árboles	Regresión 0	Regresión 1
Accuracy	0,77	0,75	0,85	0,83
F-score	0,77	0,76	0,85	0,82
Precision	0,73	0,71	0,83	0,82
Recall	0,72	0,72	0,81	0,76
Kappa	0,67	0,66	0,79	0,76

Tabla 3.16: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para los cuatro modelos realizados sobre el conjunto de datos correspondientes a las variables del primer y segundo trimestre del año hidrológico.

los primeros dos meses del año, los modelos de regresión logística son los que presentan mejores resultados, siendo el que utiliza un *alpha* igual a cero el que muestra ser el ganador.

3.4.5. Modelos adicionando los datos del índice del NIÑO

Tal como ha sido mencionado, para finalizar el análisis se prosiguió a adicionar a los modelos los datos de las oscilaciones climáticas del ENSO (El Niño-Southern Oscillation).

Dichos modelos fueron realizados únicamente sobre los modelos de regresión logística con *alpha* igual a cero debido a que en las secciones previas se probó que fue aquel el modelo que arrojó mejores resultados para la mayoría de los conjuntos de datos. A pesar de que no fue así para las variables del primer trimestre, es relevante resaltar que no resulta significativa la diferencia en los resultados de las métricas entre el modelo de árboles de decisión y el de regresión logística.

Asimismo, debido a que, tal como fue demostrado por (Meis et al., 2021), los efectos del ENSO en la cuenca del Plata se observan con un desfasaje de dos estaciones (seis meses), los modelos que se muestran en las subsiguientes secciones son aquellos que consideran únicamente el primer y el segundo trimestre del año. Asimismo, para cada uno de ellos se utilizó el trimestre en curso y el trimestre anterior de los valores del Niño.

Por último, es de relevancia aclarar que para los presentes modelos únicamente se utilizó la información del NIÑO 1+2 debido a la alta correlación que existe entre dichos datos y las restantes variables con información del NIÑO tales como el NIÑO 3, 3.4, SOI y PDO.

3.4.5.1. Modelo con variables del primer trimestre

Para la presente sección se elaboró un único modelo utilizando la técnica de regresión logística con *alpha* igualando a cero; es decir con penalización del tipo Ridge. Para ello, fueron utilizadas las siguientes variables:

- Mediana del primer trimestre
- Máximo del primer trimestre
- Mínimo del primer trimestre
- Media del NIÑO 1+2 del primer trimestre
- Media del NIÑO 1+2 del cuarto trimestre del año hidrológico anterior

Conservando la forma de presentar los resultados de las secciones previas y luego de aplicar la técnica de validación cruzada junto con el mencionado algoritmo de regresión logística, se han analizado sus resultados correspondientes a la métrica de *Accuracy*.

En la Figura 3.21 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

En este caso se ha acotado el valor de *lambda* al rango comprendido entre 0 y 0,15 ya que era aquel en el que se observaron los valores más altos de la métrica.

Mencionado esto, se puede observar la evolución del valor de la métrica de *Accuracy* en la medida que varía el valor de *lambda* tanto para la curva de resultados del conjunto de entrenamiento como para la de testeo. Para el caso del conjunto de entrenamiento, se observa que el *Accuracy* ronda un valor cercano a 0,57 durante todo el intervalo analizado, a excepción del rango de *lambda* comprendido entre 0,10 y 0,12 que muestra un valor mayor. Por otro lado, en el caso de los resultados del conjunto de testeo, los mismos comienzan con un valor de *Accuracy* de 0,65 hasta estabilizarse en 0,59 cuando el conjunto de testeo presenta su valor máximo.

Por lo tanto, fue el valor de *lambda* igual a 0,10 el elegido para entrenar el modelo con el set de entrenamiento completo para luego testear en el set de testeo.

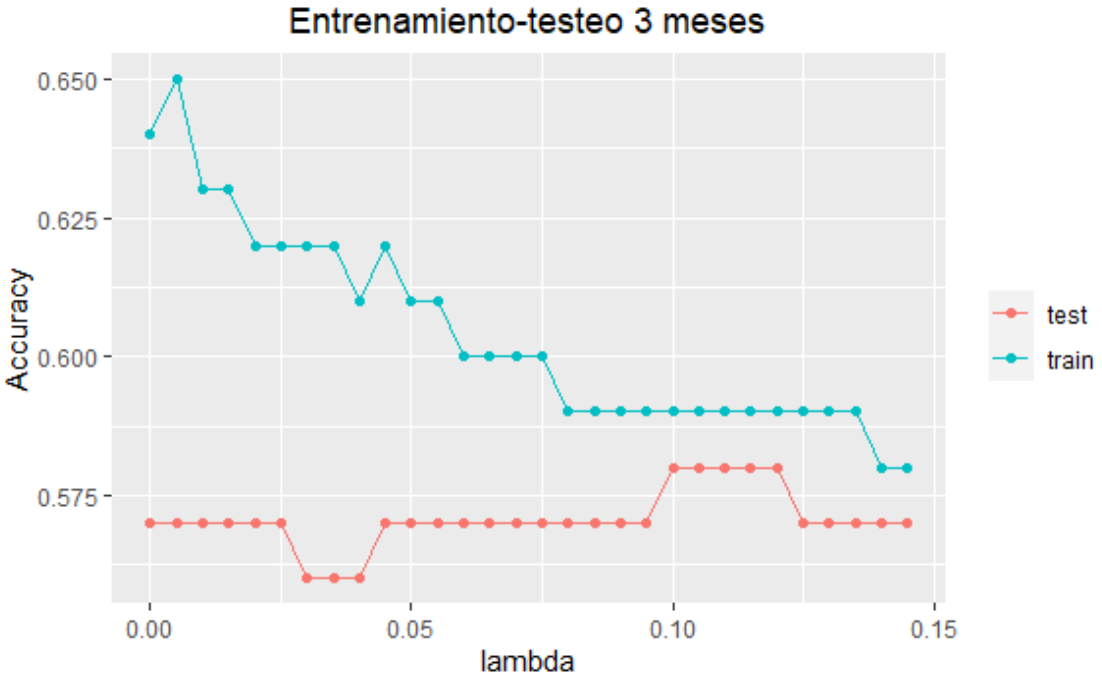

Figura 3.21: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a cero y la métrica de *Accuracy* utilizando las variables correspondientes al primer trimestre y las variables del Niño del primer trimestre del año hidrológico bajo análisis y del último trimestre del año anterior.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.17 donde se observa que el valor de *accuracy* se encuentra en 0,50; es decir, 0,07 puntos por debajo del valor obtenido en la misma métrica pero en el modelo de regresión logística con *alpha* igual a cero modelado con 3 meses de información sin las variables del Niño.

Métrica	**Valor**
Accuracy	0,50
F-score	Indefinido
Precision	Indefinido
Recall	0,38
Kappa	0,25

Tabla 3.17: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística con *alpha* igual a cero modelado con variables del primer trimestre y las variables del Niño del primer trimestre y del cuarto trimestre del año hidrológico anterior.

Es de interés resaltar que el resultado del *Accuracy* presentado en la Tabla 3.17 se encuentra significativamente por debajo de los valores observados en la curva de testeo de

la Figura 3.21.

Para finalizar el análisis se llevó adelante un estudio de importancia de variables que se puede observar en la Figura 3.22 donde se aprecia con claridad que las variables relacionadas con el fenómeno del Niño son las que menos efecto tienen en su capacidad de predicción del modelo. Asimismo, la variable más relevante es aquella que representa los valores máximos del primer trimestre.

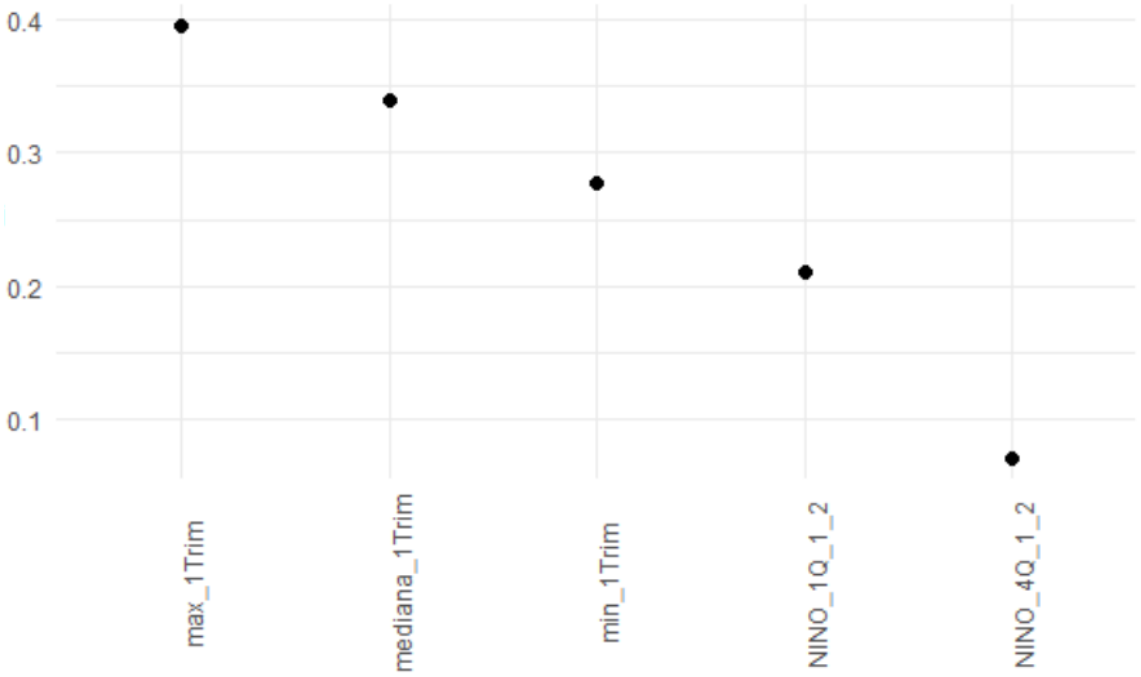

Figura 3.22: Gráfico de importancia de variables del modelo realizado con la técnica de regresión logística con *alpha* igual a cero utilizando las variables correspondientes al primer trimestre y las variables del Niño del primer trimestre del año hidrológico bajo análisis y del último trimestre del año anterior.

3.4.5.2. Modelo con variables del segundo trimestre

En este caso se elaboró un único modelo utilizando la técnica de regresión logística con *alpha* igualando a cero; es decir con penalización del tipo Ridge. Para ello, fueron utilizadas las siguientes variables:

- Mediana del primer trimestre
- Máximo del primer trimestre

- Mínimo del primer trimestre
- Mediana del segundo trimestre
- Máximo del segundo trimestre
- Mínimo del segundo trimestre
- Media del NIÑO 1+2 del primer trimestre
- Media del NIÑO 1+2 del segundo trimestre
- Media del NIÑO 1+2 del cuarto trimestre del año hidrológico anterior

Luego de aplicar la técnica de validación cruzada junto con el mencionado algoritmo de regresión logística, se han analizado sus resultados correspondientes a la métrica de *F-score*.

En la Figura 3.23 se puede observar el gráfico en el que se presentan las curvas de entrenamiento y testeo tras la aplicación de la técnica de validación cruzada. A partir de esta Figura se pueden analizar sus conclusiones.

En este caso se ha acotado el valor de *lambda* al rango comprendido entre 0 y 0,20 ya que era aquel en el que se observaron los valores más altos de la métrica.

Mencionado esto, se puede observar la evolución del valor de la métrica de *F-score* en la medida que varía el valor de *lambda* tanto para la curva de resultados del conjunto de entrenamiento como para la de testeo. Para el caso del conjunto de testeo, se observa que el *F-score* ronda un valor cercano a 0,68 al inicio del intervalo pero que comienza el descenso a partir del valor de *lambda* cercano a 0,07 y desciende más abruptamente a partir de 0,12. Por otro lado, en el caso de los resultados del conjunto de entrenamiento, los mismos comienzan con un valor de *F-score* superior a 0,76 y descienden durante todo el intervalo analizado con un descenso más pronunciado cercano al valor de *lambda* de 0,07.

Por lo tanto, fue el valor de *lambda* igual a 0,10 el elegido para entrenar el modelo con el set de entrenamiento completo para luego testear en el set de testeo.

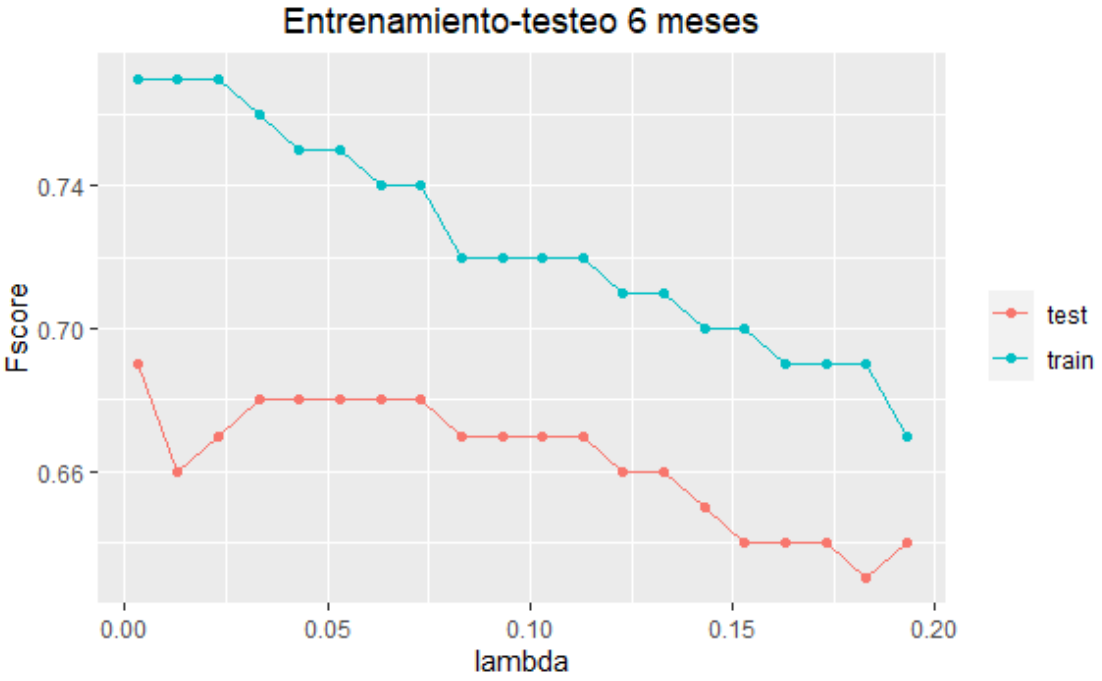

Figura 3.23: Gráfico de entrenamiento-testeo para el modelo realizado con la técnica de regresión logística con *alpha* igual a cero y la métrica de *F-score* utilizando las variables correspondientes al primer y segundo trimestre y las variables del Niño del primer y segundo trimestre del año hidrológico bajo análisis y del último trimestre del año anterior.

Luego de llevado a cabo dicho procedimiento, se observaron los resultados que se muestran en la Tabla 3.18 donde se observa que el valor de *F-score* se encuentra en 0,64; es decir, 0,06 puntos por debajo del valor obtenido en la misma métrica pero en el modelo de regresión logística con *alpha* igual a cero modelado con seis meses de información sin las variables del Niño.

Métrica	**Valor**
Accuracy	0,65
F-score	0,64
Precision	0,65
Recall	0,56
Kappa	0,50

Tabla 3.18: Resultado del cálculo de las principales métricas sobre el conjunto de testeo para el modelo de regresión logística con *alpha* igual a cero modelado con variables del primer y segundo trimestre y las variables del Niño del primer y segundo trimestre y del cuarto trimestre del año hidrológico anterior.

Al igual que en el caso del análisis anterior, el resultado del *F-score* presentado en la Tabla 3.18 se encuentra significativamente por debajo de los valores observados en la curva

de testeo de la Figura 3.23.

Finalmente, al realizar el estudio de importancia de variables que se puede observar en la Figura 3.24 se pudo ver nuevamente que las variables relacionadas con el fenómeno del Niño son las que menos efecto tienen en su capacidad de predicción del modelo.

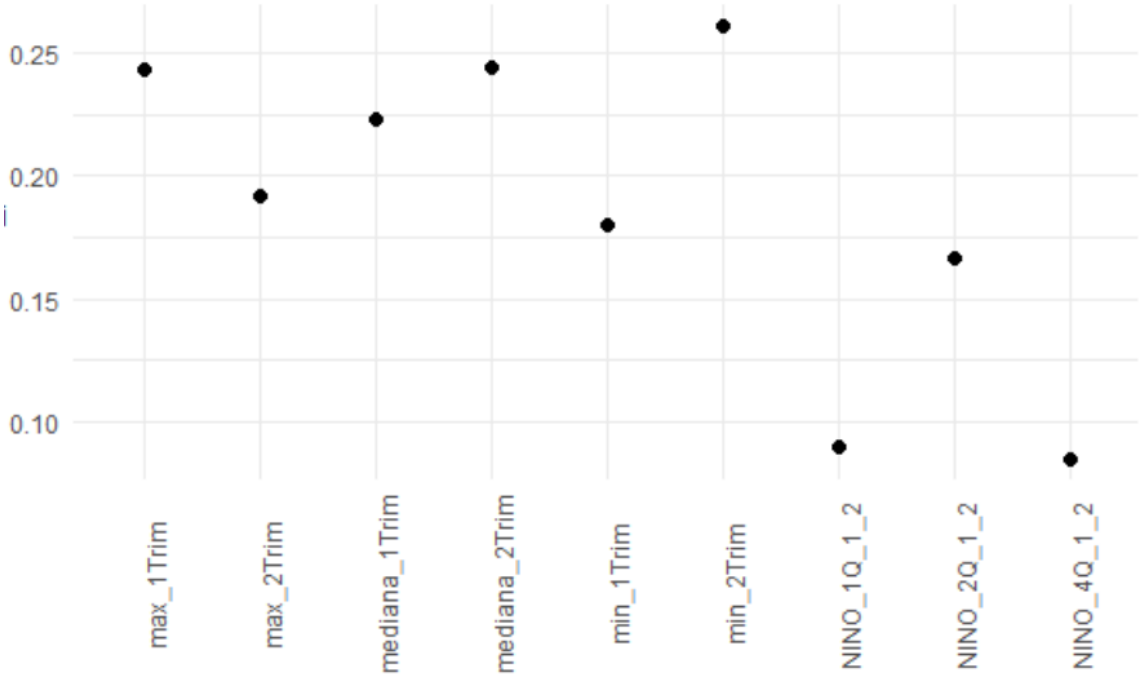

Figura 3.24: Gráfico de importancia de variables del modelo realizado con la técnica de regresión logística con *alpha* igual a cero utilizando las variables correspondientes al primer y segundo trimestre y las variables del Niño del primer y segundo trimestre del año hidrológico bajo análisis y del último trimestre del año anterior.

3.4.5.3. Empeoramiento de resultados al adicionar las variables del ENSO

En el presente apartado se buscará explicar el motivo por el cual se observa un empeoramiento en los resultados de la aplicación de los modelos al adicionar las variables relacionadas con el fenómenos del Niño.

Para tal fin, el primer paso que se realizó fue el de graficar los valores de anomalías de temperatura sobre el nivel del mar pero discriminado por los clústeres en los que originalmente se separó cada año hidrológico para cada uno de los ríos bajo análisis.

Adicionalmente, se distinguió con diferentes colores cada crónica según si correspondía a aquellas del conjunto de entrenamiento o de testeo. Este paso se realizó debido a la

notoria caída en los resultados tras el procedimiento de testeo que fue mencionada en la sección anterior. Estos gráficos a los que se hace referencia pueden ser observados en las Figuras 3.25 para el caso del río Paraná, 3.26 para el río Uruguay y finalmente la Figura 3.27 para el río Limay.

Lo que se busca obtener a través de la observación de los mismos es si tanto los valores de anomalías en la TSM de cada cluster del conjunto de entrenamiento como del conjunto de testeo tienen un comportamiento similar. En el caso de que así sea, se podría afirmar que esa información ayudaría a entrenar el modelo. Si no fuera de tal manera y las crónicas se comportaran de diferente modo, entonces la incorporación de la nueva información adicionaría ruido a la base de datos.

Es importante aclarar que los gráficos comienzan desde el mes de septiembre del año anterior debido a que, como fue comentado en las secciones previas, se toma a partir de los tres meses previos a la crónica bajo análisis de información del Niño.

A partir del análisis detallado de los resultados de la aplicación de los modelos, se arribó a la conclusión de que la mayor confusión del modelo con variables que representan las anomalías de la TSM en comparación con el modelo sin dichos datos es que el modelo concluye que hay más crónicas que tienen únicamente el primer trimestre con relativo exceso de humedad cuando, en efecto, todo ese año es húmedo. Ello quiere decir que los resultados muestran que se interpreta que las crónicas son del cluster 2 cuando verdaderamente son del cluster 1. Entonces podríamos suponer que el problema lo deberíamos ver en esos dos clústeres. En efecto, al observar las crónicas de los ríos Paraná y Uruguay de las Figuras 3.25 y 3.26 se puede comprobar que los datos de las crónicas de color celeste que son aquellas que corresponden al conjunto de testeo presentan, en ciertos casos, características que son propias del cluster 2 que tienen valores de temperatura superficial del mar anómalamente positivos.

Adicionalmente, se observó en los resultados una peor detección de las crónicas totalmente secas del río Limay. En su correlato con los nombres de los clústeres, eso quiere decir que tiene una peor detección del cluster 3 y, al observar las crónicas, se puede observar que vuelve a ocurrir que existen años hidrológicos con características anómalas para dicho

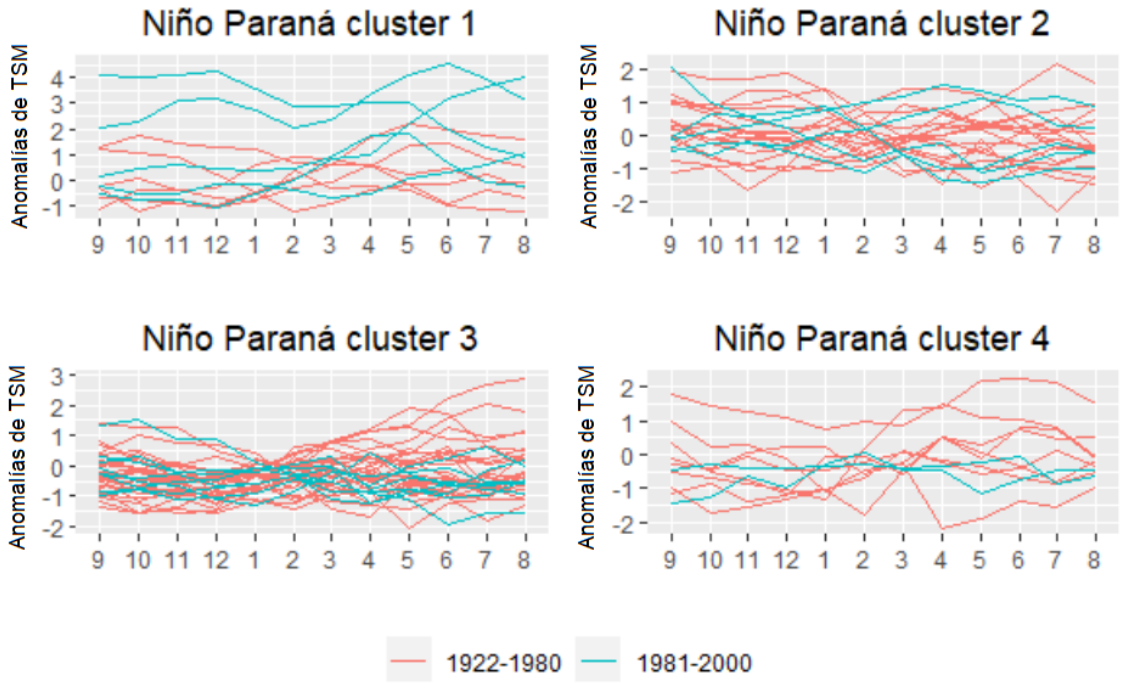

Figura 3.25: Gráficos correspondientes a los datos del ENSO para cada crónica hidrológica analizada separada por cluster y con distinción por color según el conjunto de la base de datos a la que corresponde para el caso del río Paraná.

cluster. En este caso particular, esas crónicas son las correspondientes a los años 1998 y 1983.

Por lo tanto, al observar los casos de los tres ríos, se puede inferir que el motivo por el cual la incorporación de los datos del fenómeno del Niño empeoran la capacidad de pronosticar del modelo es porque las características del set de entrenamiento presentan características diferentes con respecto a las del set de testeo.

3.5. Conclusiones

Después de analizar los resultados de los cuatro modelos desarrollados para cada conjunto de datos con el objetivo de pronosticar la crónica a la que pertenece cada año hidrológico, se puede arribar a las conclusiones que se describen a continuación.

En primer lugar, para la información correspondiente a los primeros tres trimestres, se encontró que el método de árboles de decisión proporciona los mejores resultados en términos de su *Accuracy*, aunque no logra ofrecer resultados para todas las métricas eva-

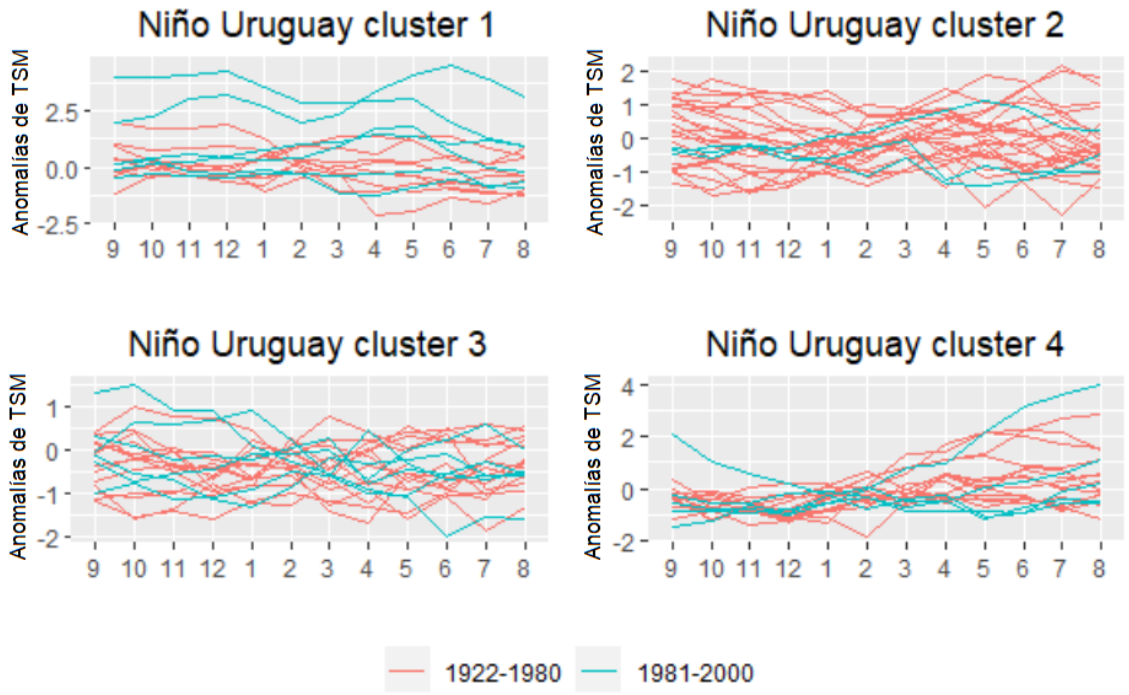

Figura 3.26: Gráficos correspondientes a los datos del ENSO para cada crónica hidrológica analizada separada por cluster y con distinción por color según el conjunto de la base de datos a la que corresponde para el caso del río Uruguay.

luadas. Sin embargo, se destaca un accuracy de 0,60, lo cual indica un desempeño notable. La regresión logística con *alpha* igual a cero también muestra buenos resultados, con un *Accuracy* de 0,57, que se encuentra muy próxima al obtenido con el método de árboles de decisión.

Al considerar las variables de los primeros seis meses, se observa una destacable mejoría en el desempeño del modelo de regresión logística con *alpha* igual a cero.

En el caso del análisis de datos de los nueve primeros meses de cada crónica, se logra confirmar que la regresión logística con *alpha* igual a cero sigue siendo la opción más efectiva.

Resulta de importancia aclarar que, aunque se utilizó el método de *k-means* para la segmentación por clústeres, se encontró que el método de KNN no resulta ser la forma más eficiente de hacer pronósticos.

Al observar los resultados de los modelos, comparándolos de acuerdo a la cantidad de trimestres de información utilizada, sin distinguir el algoritmo empleado, se puede llegar

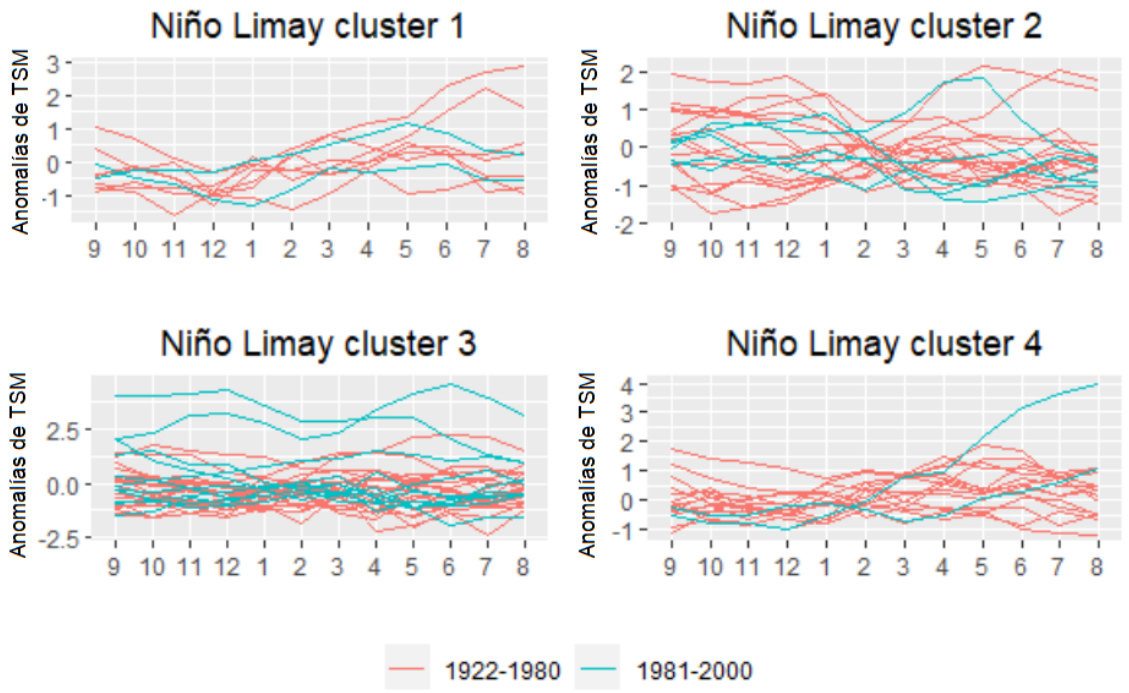

Figura 3.27: Gráficos correspondientes a los datos del ENSO para cada crónica hidrológica analizada separada por cluster y con distinción por color según el conjunto de la base de datos a la que corresponde para el caso del río Limay.

a algunas conclusiones que permiten interpretar mejor los modelos y que se encuentran en línea con el análisis de componentes principales realizado en el capítulo anterior.

En primer lugar, los modelos que únicamente utilizan variables del primer trimestre ya permiten obtener resultados muy satisfactorios debido a que es la información del primer trimestre la que define en gran medida a la segunda componente de las dos componentes principales analizadas.

Una vez incorporada la información del segundo trimestre, fue posible observar una leve mejoría aunque no es ésta tan significativa como cuando se adiciona la información del tercer trimestre debido a que es éste en el que cada crónica suele presentar sus valores máximos y, por ende, el trimestre que define a la primera componente.

Por último, al incorporar las variables relacionadas con el fenómeno del Niño, fue detectado un deterioro en los resultados. Este fenómeno podría atribuirse a diferencias en las características de las crónicas del Niño entre el conjunto de entrenamiento y el conjunto de prueba.

A partir de estas conclusiones y tras haber seleccionado los mejores modelos para cada trimestre analizado, se probaron los modelos utilizando los algoritmos y los hiperparámetros seleccionados en el conjunto de datos denominado *holdout*. Esta aplicación arrojó los siguientes resultados:

- **Modelo de árboles de decisión con variables del primer trimestre:** se obtuvo un *Accuracy* de 0,57 cuando en el conjunto de testeo había sido de 0,60.
- **Modelo de regresión logística con *alpha* igual a cero con variables del primer y segundo trimestre:** se obtuvo un *F-score* de 0,68 cuando en el conjunto de testeo había sido de 0,70.
- **Modelo de regresión logística con *alpha* igual a cero con variables del primer, segundo y tercer trimestre:** se obtuvo un *F-score* de 0,77 cuando en el conjunto de testeo había sido de 0,85.

4. CONCLUSIONES GENERALES

En la presente tesis de maestría se profundizó el estudio de la cuenca del Plata y la cuenca de los ríos Limay, Neuquén y Negro a partir del análisis de patrones temporales de las series de caudal con el objetivo de agrupar las crónicas hidrológicas de acuerdo a patrones comunes para finalmente utilizar dicha información con el fin de pronosticar el comportamiento futuro de las cuencas.

En el estudio exploratorio de los datos que se llevó adelante en el primer capítulo del presente trabajo, se estudió cada uno de los ríos bajo análisis (Paraná, Uruguay y Limay) de forma tal de tratar los datos faltantes en caso de ser necesario para luego poder describir estadísticamente el régimen de cada río. Fue en este proceso que se encontraron cambios en las tendencias históricas de las cuencas siendo particular el caso del río Paraná a partir de la década de 1970.

Adicionalmente, en dicha instancia se buscó caracterizar el régimen hidrológico de cada río para cada crónica o año hidrológico y así encontrar casos de correlación entre los distintos meses, trimestres o medidas estadísticas descriptivas dentro de las crónicas hidrológicas. Entre ellos se puede mencionar la alta correlación en las crónicas del río Paraná entre los datos del segundo y tercer trimestre con los valores máximos de las mismas. En el caso del río Uruguay, se observó una alta correlación entre los valores máximos, media y mediana y los valores del tercer y cuarto trimestre mientras que el río Limay presentó una alta correlación entre la mediana y los valores del segundo y tercer trimestre.

Fue en esta instancia que también se logró hallar correspondencias temporales entre los regímenes de los diferentes ríos. Algunas de las conclusiones a las que se arribó fueron que el río Paraná revela una estación húmeda de diciembre a mayo, una estación seca en septiembre y extremos de caudal máximo de junio a agosto, coincidiendo con eventos de El Niño. Por el contrario, lo que define al ciclo anual para el río Uruguay es que la estación húmeda es entre mayo y noviembre que coincide con el período seco del río Paraná mientras

que la estación seca es entre enero y febrero, también en contraposición con lo que sucede en la misma época en el río Paraná. Por último, en lo que respecta a las características del ciclo anual del río Limay, el período seco es entre diciembre y abril mientras que el húmedo es entre junio y noviembre, en línea con las características del ciclo del río Uruguay.

Es de relevancia aclarar que en esta instancia se halló que los tres ríos presentan la característica común de que sus valores máximos de caudal presentan una amplia variabilidad mientras que los mínimos muestran ser más estables a través del tiempo.

Seguidamente, con esta información analizada, fue aplicada una técnica de agrupamiento denominada *K-means* con el objetivo de hallar patrones comunes en las crónicas hidrológicas. Para tal fin, en primera instancia fue realizado un análisis de componentes principales con el objetivo de colaborar con la interpretación de los resultados resumiendo las variables utilizadas. Este análisis permitió distinguir dos componentes a destacar. La primera de ellas asociada a la naturaleza de las crónicas en términos de humedad y sequedad asociando los valores máximos con los del segundo y tercer trimestre tal como fue concluido en el análisis de correlación. Por otro lado, a través de la confección de la segunda componente se logró capturar la variabilidad entre las estaciones, destacando las diferencias entre las cargas del primer y cuarto trimestre.

Como resultado de la aplicación de la técnica de agrupamiento elegida, se obtuvieron cuatro clústeres que son posibles de describir a través de la caracterización de las componentes principales confeccionadas ya que logran separar las crónicas más secas, las húmedas, las que tienen un primer trimestre húmedo y, finalmente, las que tienen un cuarto trimestre húmedo.

Una vez aplicada la técnica de agrupamiento, se caracterizaron patrones comunes de crónicas hidrológicas y se observaron sus características en términos cronológicos así como también estadísticos. Asimismo, se pudieron sacar conclusiones relacionadas con patrones comunes a los tres ríos bajo análisis. Una de las principales conclusiones obtenidas fue la existencia de años en los que los tres ríos presentan crónicas secas simultáneamente. Por fuera de dicha conclusión se pueden obtener patrones comunes pero, como conclusión, no se ha observado un patrón significativo de compensación entre las cuencas en términos de

humedad de las crónicas.

Por último, este trabajo hizo énfasis en la confección de modelos probabilísticos que tuvieran el objetivo de predecir características futuras de las crónicas hidrológicas en curso. Para tal fin fueron confeccionados cuatro modelos por cada grupo de datos que correspondían a diferentes trimestres del año. Como resultado de la aplicación de los modelos de KNN, árboles de decisión y regresión logística (con penalización Ridge y Lasso) se llegó a la conclusión de que el modelo de árboles de decisión es el más exitoso para el pronóstico a partir de variables del primer trimestre del año mientras que la regresión logística con penalización Ridge es la que mejores resultados arroja en los demás casos.

Para finalizar, el trabajo estudió la influencia de oscilaciones climáticas, particularmente el ENSO, y la posibilidad de que las mismas puedan colaborar en el pronóstico del comportamiento de las crónicas de los ríos. En relación a dicho análisis se arribó a la conclusión de que éstas no aportan información que sirva para mejorar el rendimiento de los modelos de pronóstico.

En base a los resultados obtenidos tanto en la posibilidad de agrupar las crónicas como en la confección de los modelos de pronóstico, ambos análisis podrían ser fácilmente replicables en otras cuencas del país. Ello permitiría, por un lado, un uso más eficiente de los recursos hídricos al conocer el comportamiento y las correspondencias o compensaciones temporales entre las cuencas. Por otro lado, con la confección de nuevos modelos de pronóstico que incluyan otras cuencas hidrológicas, se podría trabajar en una mirada anticipatoria y, consecuentemente, también un mejor manejo de los recursos hídricos.

Asimismo, el presente análisis abre también la posibilidad de continuar investigando otras oscilaciones climáticas que puedan estar afectando el comportamiento de las crónicas de los ríos o incluso el mismo efecto del ENSO pero analizando en profundidad el período de tiempo utilizado como entrenamiento del modelo.

Apéndice

Apéndice A

ANÁLISIS DE ESTACIONARIEDAD DE LAS SERIES TEMPORALES DE CAUDAL

A.1. Resumen

A partir de los datos obtenidos de las series temporales de los tres ríos bajo análisis, en el presente anexo fue desarrollado un análisis de estacionariedad de las mismas mediante diversos tests con el objetivo de probar si su distribución y sus parámetros varían en el tiempo.

A.2. Datos

Los datos utilizados en el presente Anexo corresponden a aquellos presentados en la Sección 1 que corresponden a la información centenaria del caudal de los ríos Paraná, Uruguay y Limay.

A.3. Metodología

De acuerdo a los autores (Montgomery et al., 2015), se determina la estacionariedad de una serie temporal de acuerdo a sus propiedades estadísticas en el tiempo. En otras palabras, una serie temporal es estacionaria si exhibe una distribución de probabilidad constante a través del tiempo.

Por lo tanto, la estacionariedad de una serie temporal puede determinarse tomando muestras arbitrarias del proceso en diferentes momentos en el tiempo y observando el comportamiento general de la serie temporal. Si exhibe un comportamiento de similares características en su distribución, entonces se puede proceder con los esfuerzos de modelado bajo la suposición de estacionariedad.

Teniendo esto en consideración, la técnica utilizada para lograr el objetivo de estudiar

la estacionariedad de las series de tiempo fue el análisis de diferentes tests, tanto a nivel mensual como trimestral.

A continuación se explican las dos técnicas elegidas para tal fin.

A.3.1. Test de Dickey Fuller aumentado

El primer test utilizado para evaluar la estacionariedad de las series se trató del test de Dickey Fuller en su versión aumentada. La versión original es aquella presentada por (Dickey y Fuller, 1979).

El test de Dickey-Fuller aumentado (ADF por sus siglas en inglés) es una prueba estadística utilizada para evaluar la presencia de raíces unitarias en una serie temporal, lo que indica la no estacionariedad de la serie. La presencia de raíces unitarias sugiere que la serie temporal tiene una tendencia determinística y no exhibe un comportamiento estacionario en el tiempo.

La prueba de ADF se formula como una regresión lineal en la que la variable dependiente es la diferencia entre observaciones consecutivas de la serie temporal (Δy_t), y la variable independiente es la propia serie temporal en rezagos previos (y_{t-1}, y_{t-2},... , y_{t-p}), junto con otras posibles variables como las tendencias lineales o cuadráticas. La formulación general del modelo de regresión es:

$$\Delta y_t = \alpha + \beta y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \cdots + \gamma_p \Delta y_{t-p} + \epsilon_t \tag{A.1}$$

Donde:

- Δy_t representa la diferencia entre dos observaciones consecutivas.
- y_{t-1}, y_{t-2},... , y_{t-p} son los valores de la serie en rezagos previos.
- α es el término constante.
- β es el coeficiente asociado al valor de la serie en el rezago anterior.
- γ_1, γ_2, ..., γ_p son los coeficientes asociados a las diferencias de la serie en rezagos previos.

- ϵ_t es el término de error.

La hipótesis nula del test de ADF es que la serie temporal tiene una raíz unitaria, lo que implica que es no estacionaria. La hipótesis alternativa es que la serie no tiene raíz unitaria y, por lo tanto, es estacionaria. El test de ADF compara el estadístico de prueba calculado con valores críticos de una distribución específica para determinar si se rechaza la hipótesis nula.

En resumen, el test de Dickey-Fuller aumentado es una herramienta importante para determinar si una serie temporal es estacionaria o no, lo que proporciona información crucial para el análisis y modelado de datos temporales. Asimismo, la diferencia principal entre el test aumentado y no es que el primero permite incluir términos adicionales en el modelo de regresión para capturar la posible presencia de tendencia y estacionariedad.

En este caso, se empleó el software estadístico R (versión 4.1.1) y, en particular, se utilizó el paquete *tseries* (Trapletti y Hornik, 2021) bajo la función *adf.test* para aplicar el mencionado algoritmo.

A.3.2. Phillips-Perron

Adicionalmente, se aplicó el test de Phillips-Perron (Phillips y Perron, 1988) que es una extensión del test de Dickey-Fuller y que también se utiliza para evaluar si una serie temporal es estacionaria o no. Al igual que el test de Dickey-Fuller, el test de Phillips-Perron se basa en un modelo de regresión que incluye términos autorregresivos y una tendencia determinista. En otras palabras, el modelo analiza a partir de la dependencia de observaciones pasadas y de un crecimiento o decrecimiento persistente.

Las hipótesis del test se presentan a continuación: Hipótesis nula:

- Hipótesis Nula (H_0): La serie temporal tiene una raíz unitaria, lo que implica que no es estacionaria.

- Hipótesis Alternativa (H_1): La serie temporal es estacionaria, es decir, no tiene raíces unitarias.

De acuerdo al procedimiento general del test, en primer lugar se ajusta un modelo

autorregresivo con corrección de tendencia a la serie temporal. A continuación, se calcula un estadístico de prueba basado en el coeficiente de la raíz unitaria en el modelo ajustado. Finalmente, se compara el valor del estadístico de prueba con un valor crítico para determinar si se rechaza la hipótesis nula de no estacionariedad.

En este caso, se empleó el software estadístico R (versión 4.1.1) y, en particular, se utilizó el paquete *tseries* (Trapletti y Hornik, 2021) bajo la función *pp.test* para aplicar el mencionado algoritmo.

A.4. Resultados

En la presente sección se presentan los resultados de la aplicación de los mencionados tests para cada uno de los ríos bajo análisis.

Debido al cambio de tendencia desarrollado en la Sección 1.3.1.3 a partir de la década de 1970, se procedió a aplicar el test dividiendo la base de datos centenaria en las siguientes dos:

- Crónicas comprendidas entre 1922 y 1973
- Crónicas comprendidas entre 1974 y 2021

De esta manera, se puede estudiar la estacionariedad de cada una de las series teniendo en cuenta el cambio de tendencia previamente analizado.

A.4.1. Río Paraná

Para mostrar los resultados a los que se arribó, fueron creadas cuatro tablas que muestran el p-valor de cada test llevado adelante sobre las bases de datos.

En una primera instancia, la Tabla A.1 muestra el resultado del p-valor para cada test aplicado para el período comprendido entre 1922 y 1973 para cada uno de los meses del año.

Debido a que en todos los casos el p-valor se encuentra por debajo del valor de referencia de 0,05, se puede decir que existe evidencia estadísticamente significativa para rechazar la hipótesis nula y, por ende, se rechaza la hipótesis de que la serie no sea estacionaria. Esta

conclusión se observa tanto para el test de Dickey Fuller aumentado como para el test de Phillips-Perron.

Mes	Dickey Fuller aumentado	Phillips-Perron (PP)
Enero	0,01	0,01
Febrero	0,01	0,01
Marzo	0,01	0,01
Abril	0,01	0,01
Mayo	0,01	0,01
Junio	0,01	0,01
Julio	0,01	0,01
Agosto	0,01	0,01
Septiembre	0,01	0,01
Octubre	0,01	0,01
Noviembre	0,01	0,01
Diciembre	0,01	0,01

Tabla A.1: Resultado del p-valor de cada mes para los tests de estacionariedad utilizados para el caso de la serie temporal del río Paraná comprendida entre los años 1922 y 1973.

Asimismo, para el mismo período de tiempo también se llevó adelante el test pero para los valores trimestrales de la misma serie de tiempo. Los resultados de la aplicación del test se muestran en la Tabla A.2 y arrojan la misma conclusión: que se rechaza la hipótesis de que la serie no sea estacionaria.

Mes	Dickey Fuller aumentado	Phillips-Perron (PP)
1° trimestre	0,01	0,01
2° trimestre	0,01	0,01
3° trimestre	0,01	0,01
4° trimestre	0,01	0,01

Tabla A.2: Resultado del p-valor de cada trimestre para los tests de estacionariedad utilizados para el caso de la serie temporal del río Paraná comprendida entre los años 1922 y 1973.

Por otro lado, se aplicaron los mencionados tests en la base de datos correspondiente al período comprendido entre 1974 y 2021, tanto en su versión mensual (Tabla A.3) como en su versión trimestral (Tabla A.4) y se arribó al mismo resultado lo cual significa que existe evidencia estadísticamente significativa para rechazar la hipótesis nula y, por ende, se rechaza la hipótesis de que la serie no sea estacionaria.

A.4.2. Río Uruguay y Río Limay

En el caso de la aplicación de los tests de estacionariedad aplicados a las bases de datos de iguales características pero de los ríos Uruguay y Limay, se arribó a los mismos

Mes	Dickey Fuller aumentado	Phillips-Perron (PP)
Enero	0,01	0,01
Febrero	0,01	0,01
Marzo	0,01	0,01
Abril	0,01	0,01
Mayo	0,01	0,01
Junio	0,01	0,01
Julio	0,01	0,01
Agosto	0,01	0,01
Septiembre	0,01	0,01
Octubre	0,01	0,01
Noviembre	0,01	0,01
Diciembre	0,01	0,01

Tabla A.3: Resultado del p-valor de cada mes para los tests de estacionariedad utilizados para el caso de la serie temporal del río Paraná comprendida entre los años 1974 y 2021.

Mes	Dickey Fuller aumentado	Phillips-Perron (PP)
1° trimestre	0,01	0,01
2° trimestre	0,01	0,01
3° trimestre	0,01	0,01
4° trimestre	0,01	0,01

Tabla A.4: Resultado del p-valor de cada trimestre para los tests de estacionariedad utilizados para el caso de la serie temporal del río Paraná comprendida entre los años 1974 y 2021.

resultados de p-valor para todos los casos y, por ende a las mismas conclusiones analíticas.

A.5. Conclusiones

Luego de analizar los resultados de los tests de estacionariedad aplicados sobre las bases de datos de los tres ríos y tanto para el primer como el segundo período, se puede arribar a la conclusión de que las series de tiempo exhiben una distribución de probabilidad constante. En otras palabras, se observa un comportamiento de similares características en su distribución a lo largo del tiempo y, por lo tanto, existe evidencia estadísticamente significativa para rechazar la hipótesis de que las series de tiempo son no estacionarias.

Bibliografía

Abelen, S., Seitz, F., Abarca-del Rio, R., y Güntner, A. (2015). Droughts and floods in the la plata basin in soil moisture data and grace. Remote Sensing, 7(6):7324–7349.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Andrews, D. F. (1972). Plots of high-dimensional data. Biometrics, pages 125–136.

Antico, A., Torres, M. E., y Diaz, H. F. (2016). Contributions of different time scales to extreme paraná floods. Climate Dynamics, 46(11):3785–3792.

Barra, D. E. (2019). Determinación del riesgo por inundación en el valle del río limay, tramo arroyito-confluencia. B.S. thesis, Universidad Nacional del Comahue. Facultad de Ciencias del Ambiente y la Salud.

Barros, V., Clarke, R., y Dias, P. S. (2006). Climate change in the la plata basin. Publication of the Inter-American Institute for Global Change Research (IAI), São José dos Campos, Brazil, 5.

Berbery, E. H. y Barros, V. R. (2002). The hydrologic cycle of the la plata basin in south america. Journal of Hydrometeorology, 3(6):630–645.

Blanco, P. S. (2022). Frecuencia e intensidad de extremos de caudal del río paraná en corrientes-argentina (1910-2021). In XXIII Jornadas de Investigación, Enseñanza y Extensión de la Geografía 14 y 15 de noviembre de 2022 Ensenada, Argentina. Universidad Nacional de La Plata. Facultad de Humanidades y Ciencias de la ....

Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., y Clark, M. P. (2020). Future streamflow regime changes in the united states: assessment using functional classification. Hydrology and Earth System Sciences, 24(8):3951–3966.

Camilloni, I. y Barros, V. (2000). The parana river response to el nino 1982–83 and 1997–98 events. Journal of Hydrometeorology, 1(5):412–430.

Camilloni, I. A. y Barros, V. R. (2003). Extreme discharge events in the paraná river and their climate forcing. Journal of Hydrology, 278(1-4):94–106.

Chan, Débora, B. C. I. y Rey, A. A. (2019). Análisis inteligente de datos con lenguaje R. edUTecNe - – Editorial de la Universidad Tecnológica Nacional, Sarmiento 440, Piso 6 (C1041AAJ) Buenos Aires, República Argentina.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. Econometrica: Journal of the Econometric Society, pages 591–605.

De Boor, C. y De Boor, C. (1978). A practical guide to splines, volume 27. springer-verlag New York.

del Carmen Paris, M. y Zucarelli, G. V. (2004). Regionalización de caudales. propuesta metodológica para la identificación de regiones homogéneas. Tecnología y ciencias del agua, 19(4):5–19.

Dickey, D. A. y Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, 74(366a):427–431.

Garbarini, E., Skansi, M., Gonzalez, M., y Rolla, A. (2016). Enso influence over precipitation in argentina. Environmental Research Journal, 10(4).

Genta, J., Perez-Iribarren, G., y Mechoso, C. R. (1998). A recent increasing trend in the streamflow of rivers in southeastern south america. Journal of Climate, 11(11):2858–2862.

González, M., Garbarini, E., y Rolla, A. (2017a). Meteorological drought indices: Rainfall prediction in argentina, ch. 29 in handbook of drought and water scarcity, vol. 1: Principles of drought and water scarcity, ed. by eslamian s. and eslamian f., francis and taylor.

González, M., Garbarini, E., y Romero, P. (2015). Rainfall patterns and the relation to

atmospheric circulation in northern patagonia (argentina). Advances in environmental research, 41(1):85–100.

González, M., Romero, P., y Garbarini, E. (2017b). Droughts and floods in northern argentinean patagonia. THE ANDES, page 4.

González, M. H., Losano, F., y Eslamian, S. (2021). Rainwater harvesting reduction impact on hydroelectric energy in a rgentina. Handbook of Water Harvesting and Conservation: Case Studies and Application Examples, pages 251–260.

Grimm, A. M., Almeida, A. S., Beneti, C. A. A., y Leite, E. A. (2020). The combined effect of climate oscillations in producing extremes: the 2020 drought in southern brazil. RBRH, 25.

Gulizia, C. y Camilloni, I. (2020). Relationship between rainfall and streamflow in the la plata basin: annual cycles, interdecadal and multidecadal variability. Atmósfera.

Hastie, T., Tibshirani, R., Friedman, J. H., y Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). An introduction to statistical learning, volume 112. Springer.

Jerome Friedman, Trevor Hastie, R. T. N. S. y Narasimhan, B. (2021). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R package version 4.1-2.

Korsic, S. A. T., Notarnicola, C., Quirno, M. U., y Cara, L. (2023). Assessing a data-driven approach for monthly runoff prediction in a mountain basin of the central andes of argentina. Environmental Challenges, 10:100680.

Krepper, C. M., García, N. O., y Jones, P. D. (2008). Low-frequency response of the upper paraná basin. International Journal of Climatology: A Journal of the Royal Meteorological Society, 28(3):351–360.

Lauro, C., Vich, A. I., y Moreiras, S. M. (2018). Regional flood frequency analysis in the central-western river basins of argentina. River Research and Applications, 34(7):721–733.

Lauro, C., Vich, A. I., y Moreiras, S. M. (2019). Streamflow variability and its relationship with climate indices in western rivers of argentina. Hydrological Sciences Journal, 64(5):607–619.

Lauro, C., Vich, A. I. J., Moreiras, S. M., Bastidas Mejía, L. B., Otta, S. A., y Vaccarino Pasquali, E. L. B. (2021). Regionalización del caudal máximo anual en cuencas del sistema hidrográfico del río colorado, argentina.

Liu, Y., Gupta, H., Springer, E., y Wagener, T. (2008). Linking science with environmental decision making: Experiences from an integrated modeling approach to supporting sustainable water resources management. Environmental Modelling & Software, 23(7):846–858.

Mansor, N. S., Ahmad, N., y Heryansyah, A. (2019). Assessing dynamic-time-warping dissimilarity measures in regionalization of river discharges. Malaysian Journal of Science, 38(Sp2):14–22.

Meis, M. (2019). Coherencia regional entre el clima y el régimen hidrológico en distintas cuencas hıdricas de la Argentina. Implementación y comparación de pronósticos estadısticos de caudal. PhD thesis, Universidad de Buenos Aires.

Meis, M. y Llano, M. P. (2018). Modelado estadístico del caudal mensual en la baja cuenca del plata. Meteorologica, 43(2):63–77.

Meis, M. y Llano, M. P. (2019). Hydrostatistical study of the paraná and uruguay rivers. International Journal of River Basin Management, 17(1):1–12.

Meis, M., Llano, M. P., y Rodriguez, D. (2021). Quantifying and modelling the enso phenomenon and extreme discharge events relation in the la plata basin. Hydrological Sciences Journal, 66(1):75–89.

Meis, M., Llano, M. P., y Rodriguez, D. (2022). A statistical tool for a hydrometeorological forecast in the lower la plata basin. International Journal of River Basin Management, (just-accepted):1–38.

Mihailović, D. T., Nikolić-Đorić, E., Malinović-Milićević, S., Singh, V. P., Mihailović, A., Stošić, T., Stošić, B., y Drešković, N. (2019). The choice of an appropriate information dissimilarity measure for hierarchical clustering of river streamflow time series, based on calculated lyapunov exponent and kolmogorov measures. Entropy, 21(2):215.

Montgomery, D. C., Jennings, C. L., y Kulahci, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons.

Moraes, O., McCormick, N., Maetens, W., Magni, D., Masante, D., Mazzeschi, M., y Seluchi, M. (2021). The 2019-2021 extreme drought episode in la plata basin.

Phillips, P. C. y Perron, P. (1988). Testing for a unit root in time series regression. biometrika, 75(2):335–346.

R Core Team (2021a). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2021b). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Romero, P. E. y González, M. H. (2016). Relación entre caudales y precipitación en algunas cuencas de la patagonia norte. Revista de Geología Aplicada a la Ingeniería y al Ambiente, (36):7–13.

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Vicente-Serrano, S. M., Wehner, M., y Zhou, B. (2021). 11 chapter 11: Weather and climate extreme events in a changing climate.

Seoane, R. y López, P. (2007). Assessing the effects of climate change on the hydrological regime of the limay river basin. GeoJournal, 70(4):251–256.

Sillitoe, P. (2021). The Anthroposcene of Weather and Climate: Ethnographic Contributions to the Climate Change Debate. Berghahn Books.

Tan, S. y Kumar (2014). Introduction to Data Mining. Pearson, Edinburgh Gate, Harlow, Essex, UK.

Therneau, T., Atkinson, B., Ripley, B., y Others (2021). rpart: Recursive Partitioning and Regression Trees. `https://cran.r-project.org/package=rpart`.

Trapletti, A. y Hornik, K. (2021). tseries: Time Series Analysis and Computational Finance. R package version 3.0-10.

Vanelli, F. M. y Kobiyama, M. (2021). How can socio-hydrology contribute to natural disaster risk reduction? Hydrological Sciences Journal, 66(12):1758–1766.

Venables, W. N. y Ripley, B. D. (2002a). class: Functions for Classification. R package version 7.3-20.

Venables, W. N. y Ripley, B. D. (2002b). knn: k-Nearest Neighbour Classification. R package version 7.3-7.

Vu, Q. V. (2011). ggbiplot: A ggplot2 based biplot. `https://github.com/vqv/ggbiplot`.

Weisstein, E. W. (2006). Correlation coefficient. https://mathworld. wolfram. com/.

Zabala, P. L., Aravena, J., y Jurio, E. (2021). Urbanización de áreas ribereñas del rio limay en neuquén y plottier. Boletín Geográfico, 43(1).

Zeileis, A., Grothendieck, G., Ryan, J. A., y Ulrich, J. M. (2022). zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations). R package version 1.8-9.

Zucarelli, G. V. (2017). Regionalización hidrológica con métodos estadísticos multivariados.