UNIVERSIDAD NACIONAL DE ASUNCIÓN
Facultad de Ciencias Exactas y Naturales
Dirección de Postgrado
Maestría en Elaboración, Gestión y Evaluación de Proyectos de Investigación Científica

TÉCNICAS MULTIVARIADAS APLICADAS AL ESTUDIO DE LA CONCENTRACIÓN DE IONES EN AGUAS DEL EMBALSE DE YACYRETÁ

JUAN IGNACIO MERELES AQUINO

Tesis presentada a la Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Asunción, como requisito para la obtención del Grado de Magíster en Elaboración, Gestión y Evaluación de Proyectos de Investigación Científica

San Lorenzo - Paraguay
Julio - 2018

UNIVERSIDAD NACIONAL DE ASUNCIÓN
Facultad de Ciencias Exactas y Naturales
Dirección de Postgrado
Maestría en Elaboración, Gestión y Evaluación de Proyectos de Investigación Científica

TÉCNICAS MULTIVARIADAS APLICADAS AL ESTUDIO DE LA CONCENTRACIÓN DE IONES EN AGUAS DEL EMBALSE DE YACYRETÁ

JUAN IGNACIO MERELES AQUINO

Orientadora: **Prof. Dra. MARÍA CRISTINA MARTÍN**

Tesis presentada a la Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Asunción, como requisito para la obtención del Grado de Magíster en Elaboración, Gestión y Evaluación de Proyectos de Investigación Científica

San Lorenzo - Paraguay
Julio - 2018

Datos Internacionales de Catalogación en la Publicación (CIP)
DE LA BIBLIOTECA E INTERNET DE LA FACEN - UNA

Mereles Aquino, Juan Ignacio

Técnicas Multivariadas aplicadas al estudio de la concentración de iones en aguas del Embalse de Yacyretá/ Juan Ignacio Mereles Aquino. - - San Lorenzo: FACEN, 2018.

i-xi, 73 h.; 30 cm.

Incluye anexos y bibliografías

Tesis (Magíster en Elaboración, Gestión y Evaluación de Proyectos de Investigación Científica). – UNA. Facultad de Ciencias Exactas y Naturales. Dirección de Postgrado, 2018.

1. Análisis multivariado 2. Análisis de componentes principales (ACP) 3. Análisis de conglomerados (AC) 4. Análisis multivariado no paramétrico de la Varianza 5. Embalse Yacyretá 6. Modelo MANOVA 7. IONES 8. Calidad de agua 9. Tesis y disertaciones académicas I. Título.

519.5028/M542t

TÉCNICAS MULTIVARIADAS APLICADAS AL ESTUDIO DE LA CONCENTRACIÓN DE IONES EN AGUAS DEL EMBALSE DE YACYRETÁ

Juan Ignacio Mereles Aquino

Tesis presentada a la Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Asunción, como requisito para la obtención del Título de Magíster en Elaboración, Gestión y Evaluación de Proyectos de Investigación Científica.

Fecha de aprobación: 25 de julio de 2018

MESA EXAMINADORA

MIEMBROS:

Prof. Dra. María Cristina Martín
Universidad Nacional de la Pampa, Argentina

Prof. Dr. Hugo Eduardo Cerecetto Meyer
Universidad de la República, Uruguay

Prof. Dr. Javier Alcides Galeano Sánchez
Universidad Nacional de Asunción, Paraguay

Prof. Dr. Fernando José Méndez Gaona
Universidad Nacional de Asunción, Paraguay

Prof. Mg. Carlos Anibal Peris Castiglioni
Universidad Nacional de Asunción, Paraguay

Prof. MSc. Viviana Isabel Díaz Escobar
Universidad Nacional de Asunción, Paraguay

Aprobado y catalogado por la Dirección de Postgrado de la Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Asunción, en fecha 18 octubre de 2018.

Prof. MSc. Viviana Isabel Díaz Escobar
Director de Postgrado, FACEN - UNA

A mis queridos padres:
Regina Aquino Ibarra e Isidro Mereles González,

A mi amada esposa e hija:
Karim Pineda Rodríguez y Josephine Mereles Pineda.

AGRADECIMIENTOS

A Dios, por darme la oportunidad de extender mi conocimiento, y fortalecerme día tras día para culminar con éxito esta etapa de mi vida.

A mi familia, que en todo momento me ha incentivado para sobresalir en la vida y en especial en este desafiante trabajo.

A la tutora Dra. María Cristina Martín, que a pesar de las adversidades que se presentaban, supo conllevar este trabajo a buen camino orientándome y apoyándome en todo.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y a la Facultad de Ciencias Exactas y Naturales (FACEN) que conjuntamente llevaron a cabo este valioso Postgrado.

A la Dirección de Investigación y a la Coordinación de la Maestría, Prof. Dr. Fernando Méndez Gaona y colaboradores quienes estuvieron en los momentos justos para las orientaciones académicas y administrativas.

A la Entidad Binacional Yacyretá (EBY) por permitirme acceder a los datos que fueron utilizados en este trabajo.

A la invaluable ayuda de los profesores MSc. Teófilo Alberto Díaz, MSc. Hugo Rojas y Lic. Liz Centurión.

TÉCNICAS MULTIVARIADAS APLICADAS AL ESTUDIO DE LA CONCENTRACIÓN DE IONES EN AGUAS DEL EMBALSE DE YACYRETÁ

Autor: Juan Ignacio Mereles Aquino
Orientador: Prof. Dra. María Cristina Martín

RESUMEN

El estudio de las concentraciones de iones en el Embalse de Yacyretá es interesante para conocer el comportamiento de las mismas a través del espacio y del tiempo y determinar de esa manera la calidad del agua del Embalse. Se debe recurrir a Métodos del Análisis de Datos Multivariados para el estudio conjunto de las concentraciones de iones como el Bicarbonato, Calcio, Sodio, Potasio, Magnesio, Cloruro y Sulfato que determinan mayoritariamente la salinidad de las aguas, por las interacciones o estructuras que estas puedan tener entre sí y que no pueden ser detectadas por las técnicas estadísticas univariadas. La importancia de los métodos multivariados radica en que realizan el análisis sobre los datos tomándolos de manera simultánea y teniendo en cuenta la interdependencia, dependencia o correlaciones que estos puedan tener. Los datos analizados corresponden a mediciones realizadas mensualmente, desde febrero de 2001 hasta noviembre de 2010 en tres Estaciones de Muestreo del Embalse de Yacyretá; Puerto Candelaria, Canal de los Jesuitas e Itá Ibaté que son considerados como Entrada, Centro y Salida del Sistema Yacyretá. Los métodos multivariados utilizados en este trabajo son el Análisis de Componentes Principales (ACP), el Análisis de Conglomerados (AC) o de Agrupamiento como técnicas exploratorias multivariadas y el Análisis Multivariado de la Varianza no paramétrico (PERMANOVA) como técnica inferencial. Con el ACP y el AC se logra agrupar los siete iones en tres grupos similares en cada Estación de Muestreo del Embalse, sugiriendo homogeneidad de las concentraciones iónicas, estación por estación y, además, logrando una reducción de la dimensión del problema en cuestión. Con el PERMANOVA se concluye que, las concentraciones medias de los iones permanecen significativamente invariantes en las tres Estaciones de Muestreo en el periodo estudiado.

Palabras Clave: Análisis de Componentes Principales, Análisis de Conglomerados, Análisis Multivariado no paramétrico de la Varianza, Embalse de Yacyretá, iones.

MULTIVARIATE TECHNIQUES APPLIED TO THE STUDY OF IONS CONCENTRATION IN YACYRETÁ RESERVOIR WATER

Author: Juan Ignacio Mereles Aquino
Advisor: Prof. Dra. María Cristina Martín

SUMMARY

The study of the ion concentrations in the Yacyretá Reservoir it is interesting to know their behavior through space and time and thus determine the water quality of the reservoir. The Multivariate Data Analysis Methods should be used for the joint study of ion concentrations such as Bicarbonate, Calcium, Sodium, Potassium, Magnesium, Chloride and Sulfate, which mainly determine the salinity of the water, because of the interactions or structures that these may have with each other and that can not be detected by univariate statistical techniques . The importance of multivariate methods is that they perform the analysis on the data simultaneously and taking into account the interdependence, dependence or correlations that these data may have. The analyzed data correspond to measurements carried out monthly, from February 2001 to November 2010, in three Sampling Stations of the Yacyretá Reservoir: Puerto Candelaria, Canal de los Jesuitas and Itá Ibaté that are considered as Entry, Center and Exit of the Yacyretá System. The multivariate methods used in this work are Principal Component Analysis (PCA), Cluster Analysis (AC) as multivariate exploratory techniques and Multivariate Analysis of Nonparametric Variance (PERMANOVA) as inferential technique. With the PCA and the AC it is achieved group the seven ions into three similar groups in each Reservoir Sampling Station, suggesting homogeneity of the ionic concentrations station by station and, in addition, achieving a reduction of the dimension of the problem in question. With the PERMANOVA it is concluded that the average concentrations of the ions remain significantly invariant in the three Sampling Stations in the studied period.

Keywords: Principal Component Analysis, Cluster Analysis, Nonparametric Multivariate Analysis of Variance, Yacyretá Reservoir, ionic.

ÍNDICE

Página

LISTA DE FIGURAS

Página

LISTA DE TABLAS

Página

1. INTRODUCCIÓN

La creación de embalses altera el comportamiento de la circulación de las aguas, provocando cambios sustanciales en las concentraciones de compuestos químicos que son arrastrados hacia las profundidades. Un aspecto importante que refleja, en parte, la calidad del agua es la salinidad que esta pueda tener. Particularmente, existen ciertos iones que determinan esa salinidad en las aguas y variaciones significativas en sus concentraciones podrían afectar el ecosistema y los usos habituales que de las aguas dependen.

En este trabajo se estudian algunos iones principales como; *Bicarbonato, Cloruro, Sulfato, Calcio, Magnesio, Sodio y Potasio* que definen, en parte, la calidad del agua del *Embalse de Yacyretá* en el periodo 2001-2010, utilizando técnicas estadísticas de Análisis de Datos Multivariados. Estas técnicas permiten analizar los datos de manera simultánea en cuanto a las variables que se están midiendo sobre los individuos u objetos de estudio, considerando de esta manera las interdependencias o dependencias entre ellas.

La concentración de iones en el agua puede variar dependiendo de muchos factores, entre ellos el espacio y el tiempo. Por esta razón, es importante identificar si existen diferencias significativas en las concentraciones de los siete iones en los tres puntos de muestreos considerados (*Entrada, Centro y Salida*) del Embalse de Yacyretá, que los expertos consideran de interés para el estudio de la calidad del agua, tomando simultáneamente estos iones.

Así también, se intenta mostrar la utilidad que las técnicas estadísticas de Análisis de Datos Multivariados tienen para el estudio de la calidad del agua, específicamente en la salinidad del Embalse de Yacyretá a través del estudio de los siete iones principales, anteriormente citados.

1.1. Planteamiento del Problema

Desde la creación del Embalse de Yacyretá se han realizado varios estudios relacionados a la calidad de sus aguas, desde trabajos de tesis de grado y postgrado a investigaciones financiadas por la propia Entidad Binacional Yacyretá (EBY). La necesidad de conocer si la creación del Embalse de Yacyretá incide o no en la calidad del agua motiva a los investigadores a realizar diferentes tipos de estudios utilizando distintas técnicas y procedimientos, y encontrar aquellos factores que más intervienen o inciden en la calidad del agua.

La Facultad de Ciencias Exactas y Naturales (FACEN) de la Universidad Nacional de Asunción (UNA), gracias a un Convenio Marco firmado con la EBY en el año 1993, realiza, desde entonces, tareas de tomas de muestras de embalses y sub-embalses de Yacyretá con el fin de estudiar la calidad del agua. En este sentido, se realizan monitoreos periódicos recolectando datos relacionados a las variables hidroquímicas o fisicoquímicas que definen la calidad del agua.

La evolución y la dinámica de las concentraciones de los iones, que definen la salinidad de las aguas del Embalse de Yacyretá, pueden ocasionar cambios en la biodiversidad que en ella hay, y en el proceso de la potabilidad de la misma. Un enfoque multivariado para el estudio de la salinidad, que define en parte la calidad del agua del Embalse, permitirá obtener resultados representativos, encontrar estructuras en los datos y establecer relaciones en el espacio y el tiempo, ya que se toman mediciones de manera conjunta de las distintas variables, en este caso la de los iones, sobre cada objeto de estudio.

1.2. Justificación

El agua es fuente de desarrollo para los seres vivos (Sierra, 2011) y materia prima para la generación de energía, por lo que estudiar la calidad de la misma se ha convertido en uno de los temas más importantes en los últimos años. Por estos motivos, el análisis de los iones del agua es interesante para detectar el comportamiento o estructuras que éstas tienen conforme se generan cambios o modificaciones en sus usos. Resulta, entonces, muy importante estudiar la concentración de los iones en aguas del Complejo Hidroeléctrico de Yacyretá-Apipé (conocido simplemente como Yacyretá).

En el presente trabajo se estudia el comportamiento simultáneo de las con-

centraciones de siete iones en aguas del mencionado Embalse, con el soporte de técnicas Exploratorias e Inferenciales de Análisis de Datos Multivariados, como el Análisis de Cluster o de Conglomerados, el de Componentes Principales y el Análisis Multivariado de la Varianza no paramétrico (MANOVA no paramétrico o MANOVA Permutacional), y así explorar si las diferencias entre las concentraciones de los iones de cada uno de los grupos definidos, en este caso, de los tres diferentes puntos de muestreo del Embalse (Entrada, Centro y Salida) son estadísticamente significativas, durante un periodo comprendido entre 2001 y 2010.

La importancia de evaluar la calidad del agua del embalse (en cuanto a iones se refiere) e identificar si existen anomalías durante el trascurso de los años en los tres puntos de muestreo precisa de técnicas multivariadas. Se busca, a su vez, establecer la utilidad que estas técnicas de Análisis de Datos Multivariados tienen para el estudio de la calidad del agua del Embalse de Yacyretá y extenderlo al estudio de muchas otras variables que determinan la mencionada calidad.

1.3. Objetivos

1.3.1. Objetivo General

Estudiar la concentración de iones en tres puntos de muestreo: Puerto Candelaria, Canal de los Jesuitas e Itá Ibaté, considerados como Entrada, Centro y Salida del Sistema Yacyretá, en el periodo 2001-2010, utilizando técnicas de Análisis de Datos Multivariados.

1.3.2. Objetivos Específicos

i) Describir exploratoriamente, de manera univariada y multivariada, la concentración de iones en las aguas del Embalse de Yacyretá.
ii) Estudiar el comportamiento conjunto de los iones de interés procurando establecer grupos que definan esta conducta.
iii) Establecer si existen diferencias significativas de la concentración de los siete iones en los tres puntos de muestreo del Embalse de Yacyretá.

1.4. Hipótesis

La concentración de iones (Bicarbonato, Cloruro, Sulfato, Calcio, Magnesio, Sodio, Potasio) en aguas del Embalse de Yacyretá se mantiene invariable en los tres puntos de muestreo (Entrada, Centro y Salida), durante el periodo 2001-2010.

2. MARCO TEÓRICO

2.1. Métodos para el Análisis de Datos Multivariados

Los métodos del Análisis de Datos Multivariados consisten, básicamente, en analizar grandes volúmenes de datos tomando simultáneamente las variables medidas sobre cada objeto de análisis (Johnson y Wichern, 1992; Comas *et al.*, 1998; Peña, 2002; Pérez, 2004; Cuadras, 2014). Estos métodos fueron considerados relevantes para la investigación científica con la aparición de los ordenadores y el desarrollo de paquetes estadísticos, ya que con estos se solucionaba el problema de procesar y analizar datos voluminosos que manualmente son casi imposibles de realizar.

Es así que, especialmente desde las últimas dos décadas del siglo pasado, los métodos de Análisis de Datos Multivariados se han venido utilizando en casi todas las ramas del conocimiento, incluyendo las ciencias ambientales, y muy en especial en el estudio de la calidad del agua porque permiten estudiar las relaciones existentes entre las variables y objetos de investigación, estableciendo relaciones funcionales multivariadas, clasificando y resumiendo información a partir de la combinación simple de variables con una mínima pérdida de la información original, entre otras ventajas importantes (Peña, 2002; Pérez, 2004; Ávila *et al.*, 2015).

2.2. Técnicas exploratorias del Análisis Multivariado

En todo análisis estadístico la exploración de los datos es relevante para conocer la estructura general de los mismos, si existen o no datos influyentes o atípicos y ver las posibles relaciones existentes entre los objetos o variables de estudio (Johnson & Wichern, 1992; Pérez, 2004). Los análisis exploratorios pueden ser de carácter numérico o de carácter gráfico. Sin embargo, ambos proporcionan información útil para describir la distribución de los datos.

Cuando se dispone de una gran cantidad de datos los métodos multivariados juegan un papel trascendental en el descubrimiento de estructuras, agrupaciones y/o relaciones de dependencia entre las variables en estudio. A continuación se mencionan las características y propiedades principales con que cuenta cada método multivariado seleccionado y utilizado en el presente trabajo.

2.2.1. Caras de Chernoff

Las *caras de Chernoff* son representaciones gráficas que muestran, mediante rasgos faciales, el comportamiento de cada variable en los individuos u objetos que intervienen en el estudio. Es decir, se pueden apreciar mediante "rostros" del ser humano características multidimensionales de cada una de las observaciones y de esta manera hacerlos comparables unos con otros. Cada variable representa una característica del rostro (ojos, cejas, cabellos, boca, nariz, etc.) cuyas asignaciones se realizan de forma arbitraria, generalmente.

Este método es utilizado para aprovechar la capacidad humana para reconocer y diferenciar los rostros humanos mediante ciertas características del mismo (Hair *et al.*, 1999). La cantidad máxima de rasgos faciales o variables que puede soportar un rostro es de aproximadamente 20. Es un método descriptivo bastante útil ya que permite tener una visualización de ciertas agrupaciones mediante rasgos similares.

2.2.2. Gráfico de Estrellas

Las mediciones realizadas sobre ciertas observaciones u objetos de estudio también pueden ser representadas mediante *rayos* en un gráfico llamado de *estrellas*, que al igual que las caras de Chernoff, tiene el objetivo de resumir información mediante las características que posee cada objeto de estudio. Para cada uno de los objetos se generan "estrellas" que tienen rayos cuyas longitudes son proporcionales a los valores que toma cada variable analizada sobre el objeto de estudio.

Mediante los gráficos de estrellas se pueden realizar agrupaciones de las observaciones u objetos según la similitud que posean las estrellas generadas. Son sumamente útiles para la descripción de datos multivariados.

2.2.3. Matriz de diagramas de dispersión

Consiste básicamente en una matriz gráfica donde en cada componente se disponen las dispersiones de puntos entre pares de variables. Ayuda a descubrir si las relaciones entre las variables son o no lineales y de esta manera decidir si el uso de la matriz de varianzas/covarianzas es adecuado o no. Proporciona, además, una representación útil para la detección de valores atípicos en las relaciones bivariantes (Peña, 2002).

2.2.4. Análisis de Componentes Principales

Cuando se analiza una gran cantidad de datos, tanto de observaciones como de variables, es necesario recurrir a métodos que permitan, de una forma eficiente, representar y/o agrupar las variables en unas pocas mediante posibles combinaciones sencillas e interpretables (Johnson y Wichern, 1992; Peña, 2002; Pérez, 2004). Esto se puede obtener mediante el Análisis de Componentes Principales (ACP) que cae dentro del grupo de métodos multivariados de interdependencia muy utilizado en casi cualquier disciplina.

El ACP es una técnica cuyo principal objetivo es la de reducir la dimensión (números de variables) de un conjunto grande (Ávila *et al.*, 2015). El método consiste en construir nuevas variables llamadas "componentes principales" a partir de combinaciones lineales de las variables originales, es decir, si en un análisis intervienen p variables el método trata de obtener $r<p$ componentes principales que expliquen la mayor proporción de la variabilidad total original de los datos. Estas componentes obtenidas tienen la particularidad de que están incorreladas (geométricamente son perpendiculares), lo que permite una mejor interpretación de los resultados obtenidos (Peña, 2002), aunque en general, las interpretaciones de las componentes obtenidas son algo difíciles y requiere del conocimiento del investigador acerca del tema estudiado (Cayuela, 2011).

El ACP tiene sentido cuando las variables originales tienen un cierto grado de correlación, lo cual indica redundancia de información. Para conocer estas correlaciones se utilizan algunos gráficos como la matriz de diagramas de dispersión y algunos indicadores numéricos como lo son el determinante de la matriz de correlaciones y el índice de Kaiser-Meyer-Olkin (KMO), que según los valores que presenten se puede hablar o no de correlación entre las variables en estudio.

Para la extracción de las componentes principales se utiliza la matriz de

varianzas-covarianzas o la matriz de correlaciones dependiendo de cómo se presentan los datos. Podría decirse, aunque no es una regla general, que se prefiere utilizar la matriz de varianzas-covarianzas cuando las variables están representadas en las mismas unidades de medida atendiendo además que no existan variables cuyas mediciones sean muy diferentes al resto, ya que esto podría ocasionar ciertas distorsiones y tendencia en los resultados y, cuando ocurre lo contrario, se prefiere utilizar la matriz de correlaciones evitando así el problema de las unidades de medida.

Según Peña (2002) el problema del ACP puede darse mediante tres enfoques:

- **Enfoque descriptivo**, en donde se busca un subespacio que cuente con menor dimensión que el espacio original de datos, proyectando los puntos sobre él sin modificar la estructura existente.

- **Enfoque estadístico**, que consiste en representar los puntos en un espacio más pequeño sin perder mucha información, esto es, perdiendo una mínima proporción de la variabilidad total original.

- **Enfoque geométrico**, que sugiere la orientación de la nube de puntos para la maximización de la varianza según el eje mayor de la elipse o elipsoide sobre la dirección de la proyección del eje.

• Obtención o cálculo de las Componentes Principales

La primera componente principal es obtenida a través de la combinación lineal de las p variables originales con la característica de que esta primera componente principal tendrá la mayor varianza entre todas. Matemáticamente esta componente principal se expresa como sigue (Peña, 2002):

$$Z_{1i} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{2.1}$$

donde los valores a_{1i}, para $i = 1, 2, ..., p$, son los pesos de cada una de las variables originales X_i en la primera componente principal, sujeta a la restricción de que la suma de sus cuadrados resulte la unidad.

Matricialmente se denota así:

$$\boldsymbol{Z_1 = X a_1} \tag{2.2}$$

siendo $\boldsymbol{X} = [X_1, X_2, ..., X_p]$ la matriz de datos multivariados de dimensión $n \times p$.

La segunda componente principal también resulta de la combinación lineal de las p variables originales cuya varianza es la segunda mayor y que representa parte de la varianza no captada por la primera componente principal:

$$Z_{2i} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \tag{2.3}$$

donde los valores a_{2i}, para $i = 1, 2, ..., p$, son los pesos de cada una de las variables originales X_i en la segunda componente principal, sujeta a la restricción de que la suma de sus cuadrados resulte la unidad y que además esta componente sea ortogonal a la primera, o lo que es lo mismo, la segunda componente esté incorrelada con la primera.

Matricialmente, esta segunda componente se puede formular como:

$$\boldsymbol{Z_2 = X a_2} \tag{2.4}$$

donde $\boldsymbol{X}$ representa la misma matriz indicada para la primera componente principal.

El proceso continua de esta manera hasta la obtención de la p-ésima componente principal, manteniendo siempre las restricciones sobre los coeficientes de las componentes principales además de la no correlación entre cada par de componentes principales.

Por lo tanto, genéricamente, se establece la notación matricial de una componente principal:

$$\boldsymbol{Z_p = X a_p} \tag{2.5}$$

• Propiedades de las componentes principales obtenidas

Las componentes principales cuentan con unas ciertas propiedades interesantes. Siguiendo a Peña (2002), estas propiedades son:

i. La variabilidad inicial de los datos se conserva, es decir, la suma de las varianzas de cada componente principal coincide con la suma de las varianzas de las variables originales.

ii. Cada componente tiene una varianza asociada que resulta del cociente entre su propia varianza y la suma de las varianzas de todas las componentes principales.

iii. Entre cada componente principal y las variables originales existe una covarianza igual al producto entre cada varianza de un componente con el vector de coeficientes del mismo componente, es decir, el producto entre cada autovector con su correspondiente autovalor.

iv. La correlación entre un componente principal y una variable original es igual al producto entre el coeficiente de la variable original y el cociente entre la desviación típica de la componente principal con la desviación típica de la variable original.

v. Las r componentes principales $(r < p)$ proporcionan la predicción lineal óptima con r variables del conjunto de variables X.

vi. Si se divide cada componente principal por su desviación típica se obtiene la estandarización multivariada de los datos originales.

• **Criterios sobre el número de Componentes Principales a retener**

Como el objetivo principal del ACP es la de reducir la dimensión de trabajo, es preferible que pocas componentes principales expliquen la mayor parte de la variabilidad total de los datos. Es por eso que es importante establecer algunos criterios que puedan dar a conocer el número adecuado de componentes principales a retener. Se mencionan algunos de los criterios más utilizados según Peña (2002), Pérez (2004) y Cuadras (2014):

i. **Criterio del gráfico de autovalores frente al número de componentes principales (Gráfico de Sedimentación):** Se representan, en un gráfico de dos dimensiones, las raíces características ubicadas en el eje de ordenadas y en el eje de abscisas las componentes principales en orden decreciente. Se van uniendo los puntos mediante segmentos formando así una figura parecida al declive de una montaña. Se seleccionan aquellas componentes principales que cuenten con autovalores distintas entre sí y se descartan aquellas que cuenten con similares autovalores. En esencia, lo que se busca en el gráfico es una especie de "codo" a partir de cuales los segmentos se vuelen aproximadamente constantes, permitiendo desechar aquellas componentes que tienen autovalores aproximados.

ii. **Criterio del porcentaje de variabilidad explicada por algunos componentes:** Se seleccionan algunas componentes que determinen un porcentaje de variabilidad acumulada de al menos 70 % u 80 %, aunque estos porcentajes pueden variar dependiendo del área en donde se está investigando.

iii. **Criterio de Kaiser o Promedio de autovalores:** Establece que debe seleccionarse aquellas componentes principales que tengan un autovalor superior al promedio de los mismos. Este valor promedio es la unidad cuando se utiliza la matriz de correlaciones para la extracción de las componentes. Por lo tanto, se seleccionan aquellas componentes principales que tengan un valor igual o superior a 1, aunque si existen algunas componentes que tengan valores inferiores pero muy próximos a 1 y explican una cierta proporción de la variabilidad total, también son seleccionadas.

• Representación gráfica de las Componentes Principales

Se pueden realizar representaciones gráficas en dos dimensiones de pares de componentes principales observando así las direcciones de las correlaciones entre cada una de las variables originales y los pares de componentes representados. Esto se hace con las componentes principales retenidas con algunos de los criterios mencionados anteriormente. Se prefieren gráficas en dos dimensiones ya que facilitan la interpretación y visualización de las correlaciones y dan un indicio de agrupación de variables mediante esas correlaciones.

• Rotación de las componentes principales

Cuando las componentes principales no pueden ser interpretadas directamente, ya sea por carecer de sentido o porque no están muy bien definidas, se recurre a la rotación de las mismas. Es decir, si en principio más de una componente principal está altamente correlacionada con las mismas variables, es generalmente inválida la interpretación de esas componentes principales. Para salvar esto se realizan rotaciones de algún tipo que generen componentes que sean interpretables en sentido teórico, de manera que cada una esté correlacionada altamente solo con algunas variables y pobremente con las demás.

La rotación se realiza con la cantidad de componentes principales retenidas anteriormente con la solución inicial (sin la rotación), por esta razón la proporción de variabilidad explicada inicialmente no cambia al menos que se desee incorporar más variables a la solución inicial (de la Garza *et al.*, 2013). Al rotar las

componentes principales es equivalente a decir, gráficamente, que los ejes de cada componente en el círculo de correlaciones es girada de tal forma que los ejes nuevos se acerquen lo más posible a las variables analizadas (Pérez, 2004; de la Garza *et al.*, 2013).

• Métodos de Rotación de Componentes

Básicamente existen dos formas de realizar la rotación de componentes, una de ellas es mediante rotaciones ortogonales y la otra mediante rotaciones oblicuas. Cada una de estas alternativas conlleva a métodos diseñados para conseguir el mismo objetivo principal, interpretabilidad correcta de las componentes principales, pero siguiendo en algunos casos medios o caminos diferentes.

Existen muchos métodos de rotación *ortogonal* y *oblicua*, sin embargo, se describe solo y brevemente la "rotación varimax" que es uno de los procedimientos más utilizados por su fácil interpretación.

Rotación Varimax: El término "varimax" proviene del hecho de que la varianza es maximizada (de la Garza *et al.*, 2013). Específicamente, este criterio tiene como objetivo la maximización de la varianza de los coeficientes que definen los efectos de cada componente sobre las variables observadas (Peña, 2002). Se utiliza para identificar a un grupo de variables con una sola componente principal y de esta manera se facilita la interpretación de cada una de las componentes. Es uno de los métodos de rotación ortogonal más utilizados.

Por otra parte, hay un número considerable de métodos de rotaciones oblicuas como el *quartimin*, *oblimax*, *promax*, *binormamin*, entre otros más. Sin embargo, los métodos de rotación oblicua presentan, en algunas situaciones, inconvenientes ya que las componentes o factores están correladas dificultando la interpretación independiente de cada componente principal (Peña, 2002).

2.2.5. Análisis de Conglomerados o de Agrupamientos

El Análisis de Conglomerados (AC), también llamado Análisis de Agrupamiento o Análisis de *Cluster*, es un conjunto de técnicas multivariadas que permite clasificar observaciones u objetos en grupos homogéneos no definidos a priori considerando las características con que cuentan estos objetos (Kaufman y Rousseeuw, 1990, 1992; Hair *et al.*, 1999; Castro *et al.*, 2012). Sin embargo, el AC

también puede utilizarse para la agrupación de variables, siendo éste uno de los objetivos de su uso en este trabajo, por lo que las definiciones y características dadas están concentradas en, justamente, la agrupación de variables.

Por otro lado, la técnica puede llevarse a cabo con variables métricas o no métricas (dicotómicas por lo general) a partir de algunos criterios de clasificación según sea la situación. El AC se divide básicamente en dos métodos para la generación de los grupos: *clusters jerárquicos* y *clusters no jerárquicos*, cada uno de ellos con algunas características particulares y diferenciales.

Los grupos son conformados utilizando medidas de distancias o de similitudes que expresan el grado de semejanza, proximidad o relación entre las variables u objetos en estudio.

Al igual que el ACP, el AC es un método exploratorio ya que no requiere de la suposición de un modelo estadístico que genere los datos, solo se basa en métodos matemáticos para la conformación de los grupos.

• Medidas de distancia y similitud

Existen muchas medidas de distancia y de similitud para evaluar el parecido entre los elementos u objetos de estudio. Entre las más comunes se mencionan las siguientes:

Medidas de distancia	
Distancia euclidiana	$d_{ij} = \left[\sum_{k=1}^{r}(X_{ik} - X_{jk})^2\right]^{1/2}$
Distancia euclidiana al cuadrado	$d_{ij}^2 = \sum_{k=1}^{r}(X_{ik} - X_{jk})^2$
Distancia de Chebychev	$d_{ij} = m\acute{a}x\lvert X_{ik} - X_{jk}\rvert$
Distancia de Mahalanobis	$d_{ij} = \sqrt{(X_i - X_j)'\Sigma^{-1}(X_i - X_j)}$
Distancia de Manhattan	$d_{ij} = \sum_{k=1}^{r}\lvert(X_{ik} - X_{jk})\rvert$
Distancia de Minkowski	$d_{ij} = \left[\sum_{k=1}^{r}\lvert X_{ik} - X_{jk}\rvert^{\lambda}\right]^{1/\lambda}$ $\lambda \geq 1$

Por otro lado, para poder medir distancias o grados de asociación entre pares de variables que son de tipo continuo, generalmente, se utilizan covarianzas y correlaciones, los cuales solo consideran las relaciones lineales entre las variables.

Además, para que no haya una dependencia de las unidades de medida, es preferible la estandarización de las mismas (Peña, 2002). La medida presente en la literatura del AC más utilizada para agrupar variables cuantitativas está definida por:

$$d_{ij} = 1 - |r_{ij}| \tag{2.6}$$

donde r_{ij} es el elemento que ocupa la posición ij de la matriz de correlaciones y se calcula como sigue:

$$r_{ij} = \frac{\sum_{k=1}^{r}(X_{ik} - \overline{X}_k)(X_{jk} - \overline{X}_k)}{\left[\sum_{k=1}^{r}(X_{ik} - \overline{X}_k)^2 \sum_{k=1}^{r}(X_{jk} - \overline{X}_k)^2\right]^{1/2}} \tag{2.7}$$

En caso de que las variables sean de tipo cualitativas, generalmente dicotómicas o transformadas a dicotómicas, se utilizan medidas de asociación para este tipo de variables. Habitualmente se utiliza la distancia del valor de Chi cuadrado de Pearson χ^2 o el coeficiente de contingencia $1 - \sqrt{\chi^2}$.

• Métodos Jerárquicos de Agrupamiento

En general, en las ciencias naturales y en otros campos, es importante realizar las agrupaciones permitiendo observar como se van configurando los grupos mediante estructuras parecidas a un árbol. Esto contribuye al estudio de los subniveles que se van generando al establecer las agrupaciones y de esa manera tomar las decisiones correctas sobre la agrupación realizada (claro está, si tiene sentido la agrupación que se va construyendo). Los métodos jerárquicos de agrupamientos se dividen en dos:

a. Clusters Jerárquicos Aglomerativos

El método aglomerativo comienza con la suposición de que cada una de las variables a agrupar constituye un grupo, en otras palabras, se considera la misma cantidad de variables como de grupos iniciales. A partir de allí las variables se van agrupando en grupos disjuntos mediante pasos sucesivos siguiendo algún criterio de agrupamiento hasta conseguir en el último paso un solo grupo, producto de la unión de todos los subgrupos formados en los pasos anteriores. En resumen, al principio se cuenta con p grupos que van uniéndose hasta conseguir un solo grupo. Siguiendo a Peña (2002) y a de la Garza *et al.* (2013).

Criterios para la definición de distancias entre grupos para Clusters aglomerativos

Se definen brevemente algunos de los métodos o criterios de unión de grupos o algoritmos de clasificación más comunes y más utilizados en muchos campos del conocimiento. Para el efecto, se considera primeramente que se tienen dos grupos U y V con n_u y n_v elementos, respectivamente. Estos dos grupos se fusionan para formar el grupo UV con $n_u + n_v$ elementos y se calcula la distancia a otro grupo W con n_w elementos con algunos de los siguientes criterios:

i. Encadenamiento simple o distancias mínimas (*single linkage*): La distancia entre los dos nuevos grupos es el *mínimo* de las distancias entre grupos antes de realizarse la agrupación. Si existen valores extremos este método tiende a ser inapropiado. Matemáticamente, la unión se define por:

$$d_{W(UV)} = min(d_{WU}, d_{WV}) \tag{2.8}$$

donde $d_{W(UV)}$ expresa la distancia del grupo W al grupo UV, d_{WU} la distancia del grupo W al grupo U y d_{WV} la distancia del grupo W al grupo V, definidas por alguna de las medidas de distancia anteriormente indicadas.

ii. Encadenamiento completo o distancias máximas (*complete linkage*): La distancia entre los dos nuevos grupos es el *máximo* de las distancias entre grupos antes de realizarse la agrupación. Este criterio tiende a producir grupos esféricos o de igual diámetro. Matemáticamente, la unión se define por:

$$d_{W(UV)} = máx(d_{WU}, d_{WV}) \tag{2.9}$$

iii. Promedio entre grupos (*average linkage*): Considera la distancia entre dos grupos como el promedio de distancia entre todas las combinaciones posibles de parejas de variables. Este es uno de los criterios más utilizados porque genera clusters compactos. Los grupos se fusionan a una distancia dada por:

$$d_{W(UV)} = \frac{\sum_i \sum_j d_{i,j}}{N_{UV} N_W} \tag{2.10}$$

donde $d_{i,j}$ es la distancia entre el objeto i en el grupo UV y el objeto j en el grupo W; N_{UV} es el número de objetos en el grupo UV y N_W es la cantidad de objetos en el grupo W.

iv. Promedio intra-grupos (*withing-groups linkage*): Este criterio combina los grupos considerando que la distancia promedio entre todas las variables en el grupo resultante sea lo mínimo posible. Este criterio es una variante del *average linkage*.

v. Método del centroide (*Centroid method*): Es aplicable generalmente a variables de tipo continuas. La distancia entre dos clusters es en realidad la distancia entre sus centroides o centro de gravedad. Este criterio, al no considerar los valores extremos, es considerado robusto. El criterio de agrupamiento por este método es:

$$d^2_{W(UV)} = \frac{n_u}{n_u + n_v} d^2_{WU} + \frac{n_v}{n_u + n_v} d^2_{WV} - \frac{n_u n_v}{(n_u + n_v)^2} d^2_{UV} \tag{2.11}$$

vi. Método de Ward (*Ward method*): Conocido también como criterio de varianza mínima. Considera dos clusters que al unirlos lleve al incremento menor de la varianza. Tiende a formar grupos compactos pero de igual tamaño. Para la agrupación de los objetos se parte de una medida global de heterogeneidad. Esta medida es la suma de distancias euclídeas al cuadrado entre cada objeto y su grupo, a saber:

$$\boldsymbol{W} = \sum_g \sum_{i \in g} (\boldsymbol{X}_{ig} - \overline{\boldsymbol{X}}_g)'(\boldsymbol{X}_{ig} - \overline{\boldsymbol{X}}_g) \tag{2.12}$$

vii. Método de la Mediana (*median method*): Considera como distancia, existente entre dos clusters o grupos, la mediana de las variables que integran el grupo. Es similar al método del centroide en cuanto a la ponderación de los clusters que se combinan, pero con la diferencia de que no depende del número de variables que hay en cada uno de los clusters. La expresión para agrupar los clusters se define en este caso por:

$$d^2_{W(UV)} = \frac{d_{UW} + d_{VW}}{2} - \frac{d_{UV}}{4} \tag{2.13}$$

b. Clusters Jerárquicos Divisivos

En el método divisivo ocurre lo contrario que en el método aglomerativo, es decir, se parte del supuesto de que hay un solo grupo constituido por las p variables analizadas, y a partir de este se va dividendo basándose en las diferencias existentes hasta que la cantidad final de grupos coincida con la cantidad de variables

considerados. Los métodos divisivos no son tan usados y para su conocimiento se sugiere ver Kauffman y Rousseeuw(1990).

• Selección del número de conglomerados

No existe una regla universal y objetiva para la selección óptima de clusters. Esto se debe a que no se utiliza ningún modelo de probabilidad para el agrupamiento, lo que imposibilita realizar algún tipo de inferencia sobre el número de clusters.

La visualización de las agrupaciones generadas se puede realizar mediante el "árbol de clasificación o dendrograma". Este es un dispositivo gráfico que permite obtener o sugerir, en un principio, la cantidad de conglomerados o clusters a tener en cuenta o a retener gracias a las ramas unidas que contienen a las variables en grupos disjuntos. Es por eso que este gráfico es de gran ayuda para la selección del número de clusters. Sin embargo, hay que tener algunas consideraciones antes de decidir sobre el número de clusters a retener mediante este gráfico, ya que las ramificaciones pueden estar sesgadas por mal agrupamiento.

Por este motivo, es de vital importancia el conocimiento del investigador respecto al área donde está sumergido para la construcción y selección "adecuada" del número de clusters.

• Métodos no Jerárquicos de Agrupamiento

Cuando se realiza el agrupamiento de los elementos u objetos de estudio en grupos separados que configura el propio análisis, estamos frente a cluster no jerárquicos (Pérez, 2004). Este método presupone que se han fijado de antemano la cantidad de grupos o clusters en las que se quiere agrupar los datos.

Puede encontrarse tres divisiones de los métodos no jerárquicos: i) Reasignación, que cuenta con los métodos de k-medias y k-medianas; ii) Búsqueda de la densidad, la cual utiliza aproximaciones probabilísticas y iii) Reductivos, que dispone de los análisis de factores tipo Q.

Uno de los métodos no jerárquicos más utilizados indudablemente es el de "k-medias". Este método puede ser utilizado para realizar el agrupamiento de los

objetos de forma aleatoria o bien mediante criterios propios del investigador (de la Garza *et al.*, 2013). Para verificar si la asignación de los objetos en los clusters es la correcta, se calcula la media de cada grupo de cada una de las variables utilizadas en la segmentación.

2.3. Técnicas inferenciales de Análisis Multivariado

2.3.1. Análisis Multivariado de la Varianza paramétrico

El Análisis Multivariado de la Varianza (MANOVA, por sus siglas en inglés) es una generalización del Análisis de la Varianza (ANOVA). Entonces también, en el MANOVA se estudian las posibles diferencias existentes entre los promedios de cada uno de los grupos analizados. Específicamente, lo que hace el MANOVA es tomar $p > 1$ variables dependientes métricas de forma simultánea, supuestamente relacionadas entre sí, para medir las diferencias entre los grupos formados a partir de variables independientes llamados habitualmente factores. Así, el MANOVA es una técnica multivariante de dependencia, con p variables dependientes métricas y m variables independientes no métricas.

En el MANOVA, como ya se ha mencionado, se puede utilizar por lo menos una variable independiente no métrica. Se presenta a continuación el MANOVA en el caso de que se dispone de un solo factor.

• MANOVA de un factor

La situación en la que se cuenta con p variables dependientes métricas y una sola variable independiente no métrica con K categorías o niveles se denomina "MANOVA de un factor". En ella se estudia la diferencia existente entre los vectores de medias, provenientes de las medias de las variables dependientes en cada uno de los niveles o categorías de la variable independiente. Este modelo puede ser expresado de la siguiente manera:

$$\boldsymbol{y_k} = \boldsymbol{\mu_k} + \boldsymbol{\varepsilon_k} \tag{2.14}$$

con $\boldsymbol{y_k}$ igual a la suma entre el promedio teórico del grupo ($\boldsymbol{\mu_k}$) y término de error aleatorio no controlada en el experimento ($\boldsymbol{\varepsilon_k}$). Esta formulación sugiere que el valor esperado del vector $\boldsymbol{\varepsilon_k}$ debe ser el vector nulo $\mathbf{0}$.

• **Supuestos en un modelo MANOVA de un factor**

Como el MANOVA es una extensión del ANOVA hereda supuestos similares pero considerando notaciones y definiciones multivariadas. Se enumeran a continuación estos supuestos:

i. Las observaciones son independientes entre sí dentro y entre los grupos.

ii. En cada uno de los grupos las matrices de varianzas-covarianzas deben ser iguales, este es el supuesto de homocedasticidad.

iii. Debe existir una correlación significativa entre las variables dependientes y mantenerse esa correlación grupo a grupo.

iv. En cada uno de los grupos se tiene una distribución normal multivariada de las p variables dependientes.

• **Hipótesis a probar en el MANOVA de un factor**

En el ANOVA se prueban hipótesis de igualdad de medias de todos los grupos o categorías del factor. Por su parte, el MANOVA trata de probar la hipótesis de igualdad de vectores de medias de los grupos del factor estudiado. Concretamente, se trata de probar lo siguiente:

Hipótesis Nula (H_0): Los vectores de medias de los K grupos son iguales. Es decir,

$$H_0 : \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{bmatrix} = \begin{bmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{bmatrix} = \cdots = \begin{bmatrix} \mu_{1k} \\ \mu_{2k} \\ \vdots \\ \mu_{pk} \end{bmatrix}$$

Hipótesis Alternativa (H_1): No todos los vectores de medias son iguales, es decir los K grupos no provienen de la misma población.

$$H_1 : \quad Algún \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \\ \vdots \\ \mu_{pi} \end{bmatrix} \neq \begin{bmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{pj} \end{bmatrix} \quad para\ i \neq j$$

• **Descomposición de la variabilidad total**

La descomposición de la variabilidad total (en desviaciones con respecto a la media) se realiza de manera análoga al del ANOVA. Se considera para tal efecto la siguiente relación (Uriel, 1995):

$$\mathbf{T} = \mathbf{F} + \mathbf{W} \tag{2.15}$$

donde $\mathbf{T}$ es la matriz de sumas de cuadrados y productos cruzados total, $\mathbf{F}$ es la matriz de la suma de cuadrados y productos cruzados del factor y $\mathbf{W}$ es la matriz de la suma de cuadrados y productos cruzados del residual.

Esta expresión es fundamental para la construcción de estadísticos para la realización de los contrastes en un MANOVA. Uno de ellos es el estadístico, denominado lambda (Λ) de Wilks para un MANOVA con un factor. Este estadístico considera la relación entre los determinantes de $\mathbf{W}$ y $\mathbf{T}$, concretamente,

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{2.16}$$

donde $|\mathbf{W}|$ y $|\mathbf{T}|$ representan el determinante de $\mathbf{W}$ y $\mathbf{T}$, respectivamente. La distribución exacta del estadístico en (2.16) es complicada de definir o encontrar, es por eso, que se han realizado aproximaciones a distribuciones conocidas como la F de Snedecor o la distribución Chi cuadrado.

A pesar de la complejidad en la interpretación de los resultados en un MANOVA, la potencia del mismo es extraordinaria cuando se cumplen los supuestos subyacentes en los que se basa. Sin embargo, en la práctica, muchos de estos supuestos son difíciles de conseguir, más aún la normalidad multivariante, para el cual el MANOVA es particularmente no robusto (Mardia 1971, citado por Anderson 2001). Alternativamente, ante la imposibilidad de verificar ciertos supuestos, se han diseñado pruebas no parámetricas, tales como el PERMANOVA o en otros casos llamados también MANOVA no paramétrico.

2.3.2. Análisis Multivariado de la Varianza no paramétrico

Como se ha mencionado, los supuestos en un MANOVA clásico son bastante estrictos y generalmente difíciles de cumplir. Así, al aplicar el MANOVA con la violación de algunos de los supuestos puede llevar a resultados erróneos y por ende a conclusiones equivocadas. Es por esto que Anderson (2001) propone un

método no paramétrico para el estudio del análisis multivariante de la varianza utilizando análisis de permutaciones y matrices de distancias o disimilaridad, denominado PERMANOVA.

Siguiendo al mencionado autor (Anderson, 2001 y Anderson, 2014) se destacan algunas de las características principales de este método.

Este método persigue probar la hipótesis de igualdad de los grupos pero utilizando matrices de distancias y realizando la evaluación de la significancia a través de permutaciones, sin la necesidad de suponer normalidad multivariante de los datos. Este análisis divide las matrices de distancia entre las diferentes fuentes de variación para ajustar modelos lineales. Geométricamente, se realiza una partición multivariada del espacio basada en una cierta medida de disimilaridad según el diseño elegido. Por otro lado, el p-valor obtenido mediante las permutaciones y su precisión dependerán del número de permutaciones realizadas, por lo que se aconseja utilizar una buena cantidad de permutaciones, entre más permutaciones se realicen mejor es el análisis.

Las observaciones deben ser intercambiables suponiendo una hipótesis cierta. Esto es, las observaciones deben ser independientes con distribuciones parecidas (variaciones multivariadas parecidas), ya que el PERMANOVA prueba las diferencias en la ubicación, generalmente promedios o centros de gravedad, entre los grupos estudiados. La cantidad de observaciones u objetos puede ser inferior al de variables dependientes analizadas. Es una técnica muy utilizada en diferentes campos por la potencia de la misma, aunque originalmente fue diseñada y motivada para el estudio de datos ecológicos.

• PERMANOVA de un factor

El análisis PERMANOVA de un factor persigue el mismo propósito que el MANOVA clásico de un factor. El método analiza si existe un efecto significativo de los niveles o categorías del factor sobre las múltiples variables dependientes métricas, es decir, se compara la variabilidad entre los grupos frente a la variabilidad dentro de los grupos utilizando un estadístico similar a la F, generalmente llamada “Pseudo F”. La principal diferencia entre un PERMANOVA de un factor y un MANOVA clásico de un factor es que el primero realiza la partición de la variabilidad sobre una matriz de distancias o disimilitudes para comparar geomé-

tricamente las ubicaciones de los grupos y no directamente de los datos originales como lo hace el MANOVA clásico.

• El estadístico Pseudo F

Sea $\mathbf{Y}$ una matriz que tiene p columnas o variables medidas sobre n filas u objetos de estudio. Se define una matriz de distancias o disimilitudes $\mathbf{D} = \{d_{ij}\}$, $i = 1, 2, ..., n;\ j = 1, 2, ..., n$ entre pares de unidades de muestreo. Las medidas de distancia o disimilitud pueden ser de cualquier tipo en un PERMANOVA, entre ellas se encuentra una de las más utilizadas, la semimétrica de Bray-Curtis definida como:

$$d_{ij} = \frac{\sum_{k=1}^{p} |y_{ik} - y_{jk}|}{\sum_{k=1}^{p} (y_{ik} + y_{jk})} \tag{2.17}$$

De la matriz $\mathbf{D}$ se utilizan los elementos que están por debajo de la diagonal principal de cada grupo para el cálculo de las sumas de cuadrados. Estas sumas de cuadrados son entonces sumas de cuadrados de distancias entre pares de objetos (i, j). Al igual que en un MANOVA clásico la variación total se obtiene mediante la adición de dos fuentes de variación, una entre grupos y la otra dentro de los grupos. La notación adoptada es como sigue:

$$SC_T = SC_A + SC_W \tag{2.18}$$

SC_T es la suma total de distancias cuadradas, SC_A es la suma de cuadrados entre los grupos, es decir, es la suma de las distancias cuadradas de los centroides individuales del grupo al centroide general y SC_W es la suma de cuadrados residuales el cual representa la suma de los cuadrados de las distancias a los centroides de objetos u elementos de muestreo individuales a su propio centroide del grupo.

Las sumas de cuadrados total y residual se calculan de la siguiente manera:

$$SC_T = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2 \tag{2.19}$$

$$SC_W = \sum_{\ell=1}^{g} W_\ell \tag{2.20}$$

donde g es el número de grupos del factor, W_ℓ es la suma de cuadrados dentro

del grupo en el grupo ℓ definida por:

$$W_\ell = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\epsilon^{[\ell]} d_{ij}^2}{n_\ell} \tag{2.21}$$

donde:
▪ n_ℓ es la cantidad de elementos en el grupo ℓ, es decir, el tamaño muestral en el grupo ℓ, así $n = \sum_{\ell=1}^{g} n_\ell$, y, ▪ $\epsilon^{[\ell]}$ es un parámetro indicador, esto es, si los objetos (i, j) están en el grupo ℓ entonces $\epsilon^{[\ell]} = 1$, en caso contrario $\epsilon^{[\ell]} = 0$.

Finalmente, la suma de cuadrados entre grupos se obtiene mediante sustracción directa, esto es $SC_A = SC_T - SC_W$.

Mediante las sumas de cuadrados se obtiene el llamado "Pseudo F" análogo a la estadística F de un modelo ANOVA y cuya finalidad también, es observar si existe o no discrepancias entre los grupos:

$$Pseudo\ F = \frac{SC_A/(g-1)}{SC_W/(n-g)} \tag{2.22}$$

• Supuestos de un PERMANOVA

Se supone que existe intercambiabilidad entre las unidades observacionales considerando una hipótesis nula cierta. Esta prueba puede ser sensible a la dispersión multivariada si los tamaños muestrales en cada grupo son muy diferentes, sin embargo, para diseños balanceados o aproximadamente balanceados PERMANOVA es bastante robusta.

• El uso de permutaciones para la obtención del p-valor

La distribución del estadístico Pseudo F no es como una F de Fisher del modelo ANOVA. Esto se debe a que las variables individuales no necesariamente tienen distribución normal y que la distancia utilizada no necesariamente es la euclidiana.

Un p-valor es construido a partir de permutaciones de los objetos entre los diferentes grupos del factor. Este hecho proviene de la hipótesis de permutabilidad de las observaciones en los distintos grupos. Es decir, si las observaciones son permutables se pueden disponer aleatoriamente cada observación en cualquiera de

los grupos teniendo así una mezcla de elementos de diferentes grupos y generando un nuevo valor denotado como F^{π}. Un p-valor para el PERMANOVA se calcula como:

$$p - valor = \frac{(Número\ de\ F^{\pi} \geq Pseudo\ F) + 1}{(Número\ total\ de\ F^{\pi}) + 1} \tag{2.23}$$

Este p-valor es asintótico ya que se obtiene mediante algoritmos de permutación.

2.4. El agua y sus propiedades

El agua, al igual que muchos otros recursos naturales, es imprescindible para el ser humano, sea para el consumo, para el riego o para otras utilidades como la generación de energía, este último fundamental para la calidad de vida de las personas. Es un líquido compuesto principalmente por dos gases, hidrógeno y oxígeno (H_2O). Tiene las características de carecer de olor y de sabor y es un líquido inodoro, todo esto cuando está en su estado puro. Además, es uno de los recursos más abundante del planeta, con un 71 % aproximadamente de la superficie de la corteza terrestre. Sin embargo, de estos, tan solo el 2,75 % aproximadamente corresponde a agua dulce, distribuidos estos a su vez en ríos, lagos, humedales, embalses y otros.

Por otra parte, es sabido que existen algunos parámetros que determinan las características principales del agua dependiendo de la mayor o menor concentración de los mismos, como ser, aspectos fisicoquímicos, microbiológicos, organolépticos y los de radioactividad (Rojas, 2010; Díaz, 2013). Cada uno de estos parámetros definen el olor, sabor, color, y otras características del agua que, según el grado de concentración de los mismos, van desde apto para cualquier uso hasta ser inapropiado para cualquier fin.

2.5. El uso de las aguas y su calidad

Como ya se ha mencionado, el agua cuenta con múltiples usos. Uno de los principales es el aprovechamiento de los caudales de ríos para la construcción de hidroeléctricas con el fin de producir energía eléctrica y contar así con un recurso renovable. Sin embargo, atendiendo al ecosistema de un río, cuando hay una transformación sobre él, queda afectada la biodiversidad que en él existiese, más aún también la calidad del agua. A todo esto hay que agregarle además la actividad humana de las cercanías, como las agrícolas, ganaderas e industriales

en general que, de alguna u otra forma, aportan para el cambio sustancial de la calidad del agua de los ecosistemas naturales. Si bien existen guías internacionales como la de la Organización Mundial de la Salud (OMS, 2006) acerca de la calidad del agua, es sumamente interesante el estudio desde un punto de vista estadístico.

La calidad del agua es definida por muchos factores, entre ellos la salinidad. La salinidad en el agua es consecuencia de concentraciones de ciertos iones mayoritarios como los cationes: Calcio, Magnesio, Potasio y Sodio y los aniones: Bicarbonato, Sulfato y Cloruro. Cada uno de estos iones, dependiendo de sus concentraciones en el agua, pueden propiciar un ambiente de mucha biodiversidad y hacerlo apto para la recreación e, inclusive, para el consumo. Sin embargo, variaciones significativas de estas concentraciones pueden generar hasta un total desequilibrio ecológico.

2.6. Estudios multivariados de la calidad del agua

La calidad del agua es definida por muchos factores que pueden tener interacciones entre ellos y que no pueden ser detectados mediante las técnicas habituales de análisis de datos univariados (Gómez y Peñuela, 2016). Es aquí que los métodos de Análisis de Datos Multivariados reflejan su potencia con relación a los procedimientos univariados. Existen, a nivel internacional, varias aplicaciones de los métodos multivariados al estudio de la calidad del agua. Coletti *et al.* (2010) aplicaron el Análisis Factorial para construir un índice de la calidad del agua, considerando actividades agrícolas existentes en las cercanías del río Das Pedras en Brasil. Por otro lado, se han utilizado combinaciones entre métodos, Análisis de Componentes Principales, Análisis de Clúster y Modelos de Regresión para el estudio de la calidad del agua subterránea, como el realizado en la región de Atlas, en Túnez, por Chenini y Khemiri (2009). Así también, López y Palací (2014) realizaron un estudio de la calidad del agua del Río Júcar (España) durante el periodo 1990 a 2013 con la aplicación del Análisis de Componentes Principales y el Análisis Discriminante.

Valencia, (2007) ha realizado un estudio detallado de la calidad del agua en una cuenca hidrográfica del Río Ebro con la utilización de herramientas estadísticas muy potentes y representativos.

En Paraguay existen escasos trabajos relacionados al estudio de la calidad del agua desde un enfoque multivariado, inclusive de las aguas del Embalse de

Yacyretá, una de las grandes hidroeléctricas de Sudamérica. En la gran mayoría de los casos, en el mencionado embalse, los estudios se realizan utilizando métodos univariados. Rojas (2010) analizó la calidad del agua del Embalse de Yacyretá utilizando el Test de Tendencia Kendall Estacional y el Análisis de Varianza (ANOVA univariado) para observar las diferencias entre las concentraciones de las componentes físico-químicos en el agua del embalse. Por otra parte, Díaz (2013) utilizó un modelo de Series de Tiempo univariado para estudiar la concentración de un componente iónico (Alcalinidad Total) en aguas del Embalse de Yacyretá, en el periodo 2001-2010, con la finalidad de evaluar la evolución espacio-temporal del mencionado ion.

3. METODOLOGÍA

3.1. Materiales

3.1.1. Características del área de estudio

La represa de la central hidroeléctrica Yacyretá está ubicada sobre el Río Paraná compartida por dos países, de este a oeste por Paraguay y de sureste a noreste por Argentina (EBY, 2013). Existe una distancia de 310 km con Asunción y 1.470 km de Buenos Aires, las dos capitales de los dos países mencionados, respectivamente. Exactamente su ubicación geográfica es ($27° 29' 02" - 56° 24' 44"$). La presa cuenta con una extensión total de 65 km distribuidos de la siguiente manera: en la margen derecha 28 km correspondiente a Paraguay, en la margen izquierda 17 km correspondiente a Argentina y los 20 km restantes sobre la Isla Yacyretá. La presa brinda energía eléctrica a los dos países mencionados gracias a sus 20 hidrogeneradores de una potencia instalada total de 3.200 MW. Por otro lado, la presa está a una altura máxima de 83 metros sobre el nivel del mar (Rojas, 2010).

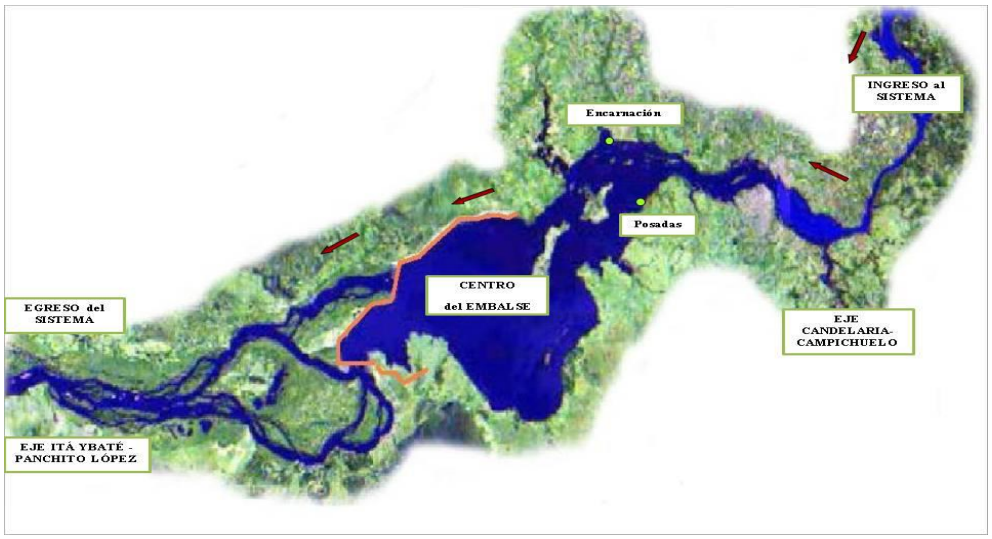

Figura 1: Ubicación de las estaciones de muestreos del Embalse de Yacyretá (Entrada, Centro y Salida).
Fuente: Extraído de Díaz (2013)

Con un caudal medio aproximado de 12.000 m^3/s, el río Paraná, es uno de los más caudalosos de Sudamérica. Además, cuenta con un área de aproximadamente 1.600 km^2. El aprovechamiento de este caudaloso río es sorprendente ya que es una de las centrales hidroeléctricas que produce grandes cantidades de energía.

Para el presente estudio se consideran tres puntos o estaciones principales de muestreo sobre el Embalse de Yacyretá. Habitualmente estos puntos de muestreo se denominan Entrada, Centro y Salida del Embalse (ver Figura 1). La Entrada está ubicada en el eje llamado Candelaria-Campichuelo. Por su parte, el Centro está ubicado aproximadamente a 60 km aguas debajo de la Entrada del sistema. Entre el Centro y la Salida del embalse existe una distancia aproximada de 80 km, este último punto de muestreo esta sobre el eje denominado Itá Ibaté-Panchito López.

3.1.2. Datos

Este trabajo se fundamenta en una investigación exploratoria-inferencial cuantitativa. Los datos utilizados para la aplicación de los métodos corresponden a mediciones mensuales realizadas por la FACEN mediante convenios de estudios sobre la calidad del agua del Embalse de Yacyretá con la EBY. Las mediciones corresponden a parámetros in situ desde febrero del año 2001 a noviembre del 2010. Se considera este periodo por la gran cantidad de datos contenidos en él, no así en los demás periodos. Cabe señalar que las muestras provienen de los tres puntos de muestreos tomadas a una profundidad de 50 cm de la superficie del embalse, excepto en el Centro que fue tomada solo de la superficie del Embalse. Todos los iones están expresados en mg/L.

3.1.3. Variables consideradas

Para la presente investigación se consideran principalmente ocho variables, de las cuales siete son métricas y una categórica (o de factor). Específicamente, siete corresponden a iones: cuatro cationes; Potasio (K^+), Calcio (Ca^{2+}), Magnesio (Mg^{2+}) y Sodio (Na^+) y tres aniones; Bicarbonato (HCO_3^-), Sulfato (SO_4^{2-}) y Cloruro (Cl^-). Y, por otro lado, la variable categórica corresponde a Puntos de Muestreo cuyas tres categorías o niveles son: Entrada, Centro y Salida, del embalse.

3.1.4. Soporte informático para el análisis de los datos

Todos los análisis estadísticos, desde los descriptivos univariados y multivariados hasta los métodos multivariados (exploratorios e inferenciales), fueron realizados íntegramente con el soporte del paquete estadístico R (mediante sintaxis en la consola y mediante rutinas de la interfaz gráfica R-Commander), cuya principal ventaja es la flexibilidad con que cuenta en su lenguaje para el análisis estadístico (Pérez, 2015; R CORE TEAM, 2016). Además, es un software de libre distribución y con colaboradores en todo el mundo que brindan ayuda a la generación de paquetes sofisticados para cálculo y análisis estadístico, especialmente los de métodos multivariados.

Los paquetes o librerías especiales más importantes de R utilizadas en este trabajo son:

- **aplpack:** Para la generación de las *caras de Chernoff* en los tres puntos de muestreo del Embalse.
- **ggplot2:** Para la construcción de los gráficos de cajas y bigotes o los *boxplot* de cada ion en las tres estaciones de muestreo.
- **ade4:** Para realizar el Análisis de Componentes Principales y los círculos de correlaciones correspondientes.
- **psych:** Para la rotación de las componentes principales utilizando el método varimax.
- **cluster:** Para el análisis de conglomerados o de *cluster*.
- **vegan:** Para realizar el Análisis Multivariado de la Varianza no Paramétrico o PERMANOVA con el comando `adonis`.

Así también, se ha utilizado *Calc* de *LibreOffice* para la tabulación de datos y ordenamiento de los resultados.

3.2. Métodos

3.2.1. Análisis descriptivo o exploratorio univariando y multivariado

Se realizan descripciones univariadas de las concentraciones de los siete iones antes mencionados y el análisis de las medidas resúmenes se hace para todo el embalse y comparándolas posteriormente con las obtenidas en cada estación de

muestreo.

Se utilizan los siguientes métodos exploratorios de Datos Multivariados:

- Las matrices de correlaciones, que permiten ver las correlaciones bivariadas entre los siete iones, por un lado considerando todo el embalse y, por otro lado, teniendo en cuenta cada punto muestral.

- Los gráficos de estrellas y las caras de *Chernoff*, para visualizar el comportamiento simultáneo de la concentración de los iones mediante estrellas y rostros, respectivamente, observando la posible discrepancia existente mes a mes.

- Análisis de Componentes Principales (ACP), que permite obtener una reducción de dimensión en cuanto a las variables analizadas. Se generan en primer lugar las componentes principales, para establecer la cantidad de componentes a retener observando la proporción de variabilidad explicada. Si es necesario, se recurre a la rotación ortogonal *varimax* de las componentes retenidas para que se pueda obtener correlaciones altas de algunos iones solo con una componente principal. Se presentan las correlaciones entre cada componente principal y cada ion en una tabla al igual que en círculos de correlaciones.

- Análisis de Conglomerados (AC) o de Agrupamientos, para la configuración de grupos de iones que están en cierta forma asociados en cada punto muestral. El dendrograma, como dispositivo gráfico, permite una visualización de la conformación posible de dichos grupos de iones.

En todos los casos, para la presentación de resultados, se utilizan tablas de resúmenes estadísticos y gráficos representativos que caracterizan parcial o completamente a las variables y a las técnicas univariadas y multivariadas utilizadas como ya se ha mencionado.

3.2.2. PERMANOVA - Análisis Multivariado de la Varianza no paramétrico

Se utiliza la técnica PERMANOVA como alternativa al MANOVA clásico, dada la imposibilidad de verificar el supuesto de normalidad multivariada y otros más para la aplicación de un MANOVA. La finalidad de la utilización de esta técnica radica en medir si existen diferencias significativas de las concentraciones

de los siete iones en los tres puntos de muestreo. Esta técnica compara la igualdad de los grupos mediante la utilización matrices de distancias y de un p-valor obtenido a través de técnicas de permutación.

4. RESULTADOS Y DISCUSIÓN

4.1. Análisis exploratorio de los datos

Antes de aplicar las estadísticas propias del análisis multivariado, se realiza un análisis exploratorio de las variables correspondiente a los siete iones, de manera individual y grupal (univariada y multivariada) para conocer la estructura y el comportamiento que los mismos tienen conforme fueron medidos espacial y temporalmente.

4.1.1. Comportamiento univariado de los iones

De la Tabla 1 y de las Figuras 2 al 8 respecto del comportamiento individual de las variables en estudio, tanto en forma global (es decir, en todo el Embalse) como en las diferentes estaciones de muestreo:

i. En media y en mediana los distintos iones no muestran importantes discrepancias estación por estación, es decir, muestran valores similares en las distintas estaciones de muestreo como en todo el Embalse (Tabla 1).

ii. Las tres estaciones de muestreo, al igual que todo el Embalse, presentan una composición iónica atendiendo a las concentraciones medias encontradas:

$$HCO_3^- > Ca^{2+} > Cl^- > Na^+ > Mg^{2+} > K^+ > SO_4^{2-}$$

Resultados similares en cuanto al ordenamiento promedio de estos iones en el embalse San Roque se encontraron por Rodríguez *et al.* (2001). Puede apreciarse que existe una supremacía absoluta en la concentración promedio del anión Bicarbonato (HCO_3^-) que está muy alejada del segundo mayor componente iónico, en este caso, el Calcio (Ca^{2+}) (Tabla 1).

iii. Las variabilidades promedios de las concentraciones de cada componente presentan algunas pequeñas discrepancias de la Entrada hasta la Salida del Embalse. Individualmente, el anión Bicarbonato es quien presenta mayor

variabilidad, esto puede deberse a que este es el ion mayoritario del Embalse con bastante diferencia con los demás iones en las tres estaciones de muestreo (Figuras 2 al 8).

iv. Los valores de la simetría y curtosis aportan para conocer la forma y distribución de los siete iones en los tres puntos de muestreo:

- Bicarbonato: Este es el único ion que cuenta con sesgo negativo (valores que se acumulan mayormente hacia la derecha de la distribución) en las tres estaciones de muestreo. Además, presenta valores que están más concentrados en el Centro y Salida del Embalse y menos concentrados en la Entrada.
- Calcio: Presenta distribuciones relativamente simétricas en las tres estaciones de muestreo del Embalse. Por otra parte, los valores negativos de las curtosis en los tres puntos de muestreo indican que la forma de la distribución son platicurticas.
- Cloruro: La distribución de este ion, en cada estación de muestreo, es sesgada positivamente. Presenta valores muy concentrados. Además, se puede apreciar valores atípicos en las tres estaciones de muestreo de este ion.
- Sodio: Presenta sesgo positivo (valores que se acumulan mayormente hacia la izquierda de la distribución) en cada punto de muestreo. Sin embargo, existen valores menos concentrados indicados por los valores negativos de la curtosis.
- Magnesio: Cuenta con sesgos negativos y con valores más concentrados en la Entrada y Centro del Embalse, mientras que en la Salida la distribución es sesgada positivamente con valores menos concentrados.
- Potasio: Este ion presenta distribución algo sesgada hacia derecha con valores no muy concentrados en cada estación de muestreo.
- Sulfato: Cuenta con sesgo positivo en las tres estaciones de muestreo y con valores menos concentrados en el Centro y Salida del Embalse, lo contrario ocurre en la Entrada.

Con estas descripciones resultaría difícil que la mayoría de las variables se distribuyan normalmente.

Tabla 1: Estadísticos descriptivos univariantes de las concentraciones en mg/L de los siete iones en los tres puntos de muestreo del Embalse Yacyretá, período 2001-2010.

Estadístico	Puntos de Muestreo	Bicarbonato	Calcio	Cloruro	Sodio	Magnesio	Potasio	Sulfato
N	Entrada	115	115	115	115	115	115	115
	Centro	114	114	114	114	114	114	114
	Salida	115	115	115	115	115	115	115
	Total	344	344	344	344	344	344	344
Media	Entrada	19,46	4,44	3,15	2,32	1,61	1,43	0,99
	Centro	19,49	4,51	3,26	2,32	1,58	1,43	1,06
	Salida	19,60	4,47	3,18	2,34	1,59	1,43	1,02
	Total	19,52	4,47	3,20	2,33	1,59	1,43	1,02
Mediana	Entrada	19,80	4,44	3,15	2,27	1,68	1,42	0,94
	Centro	19,80	4,44	3,20	2,26	1,50	1,43	0,99
	Salida	19,80	4,44	3,09	2,26	1,50	1,44	0,94
	Total	19,80	4,44	3,16	2,26	1,50	1,43	0,96
Desviación Típica	Entrada	1,19	0,49	0,47	0,39	0,21	0,27	0,56
	Centro	1,31	0,44	0,49	0,39	0,23	0,28	0,58
	Salida	1,17	0,44	0,49	0,39	0,21	0,27	0,60
	Total	1,22	0,46	0,48	0,39	0,22	0,27	0,58
Asimetría	Entrada	-0,41	0,01	0,05	0,32	-0,03	0,15	0,90
	Centro	-0,47	-0,20	0,75	0,28	-0,10	0,07	0,70
	Salida	-0,49	-0,19	0,65	0,36	0,12	0,09	0,69
	Total	-0,46	-0,13	0,50	0,32	-0,03	0,10	0,75
Curtosis	Entrada	0,11	-0,23	0,41	-0,90	0,18	-0,84	0,73
	Centro	-0,12	-0,42	1,97	-0,66	1,30	-0,60	-0,17
	Salida	0,04	-0,41	0,48	-0,75	-0,24	-1,03	-0,35
	Total	-0,01	-0,35	1,01	-0,78	0,56	-0,83	-0,01
Mínimo	Entrada	16,00	3,35	1,75	1,70	0,99	0,91	0,30
	Centro	15,80	3,45	2,20	1,54	0,72	0,75	0,30
	Salida	16,40	3,40	2,15	1,67	1,20	0,91	0,30
	Total	15,80	3,35	1,75	1,54	0,72	0,75	0,30
1er. Q	Entrada	18,70	4,04	2,88	1,99	1,48	1,19	0,54
	Centro	18,60	4,14	2,95	1,98	1,46	1,21	0,56
	Salida	18,70	4,04	2,89	2,03	1,46	1,20	0,49
	Total	18,70	4,08	2,90	2,00	1,46	1,20	0,54
2do. Q	Entrada	20,05	4,80	3,41	2,65	1,73	1,64	1,33
	Centro	20,50	4,85	3,52	2,61	1,72	1,65	1,38
	Salida	20,50	4,83	3,45	2,66	1,73	1,66	1,46
	Total	20,50	4,85	3,46	2,64	1,73	1,64	1,35
Máximo	Entrada	22,50	5,60	4,25	3,31	2,21	2,08	2,74
	Centro	22,50	5,30	5,25	3,38	2,21	2,11	2,70
	Salida	22,50	5,30	4,67	3,49	2,21	2,00	2,68
	Total	22,50	5,60	5,25	3,49	2,21	2,11	2,74

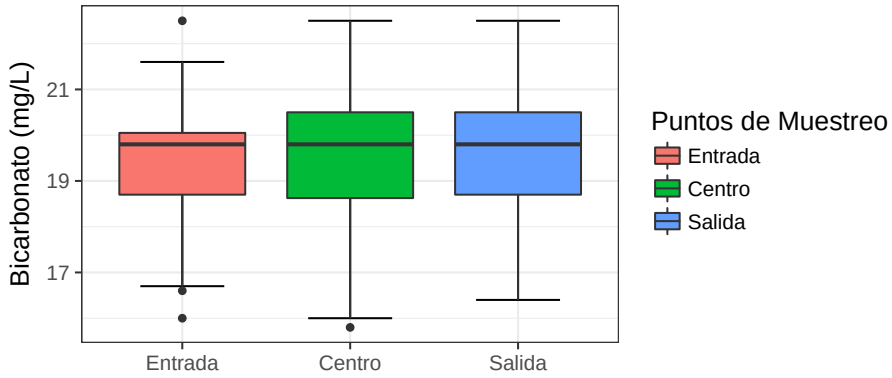

Figura 2: Gráfico de cajas y bigotes de las concentraciones del ion Bicarbonato en los tres puntos de muestreo del Embalse de Yacyretá

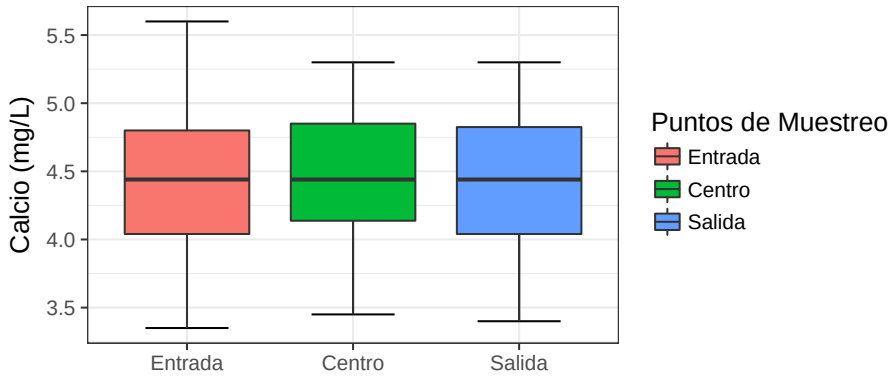

Figura 3: Gráfico de cajas y bigotes de las concentraciones del ion Calcio en los tres puntos de muestreo del Embalse de Yacyretá

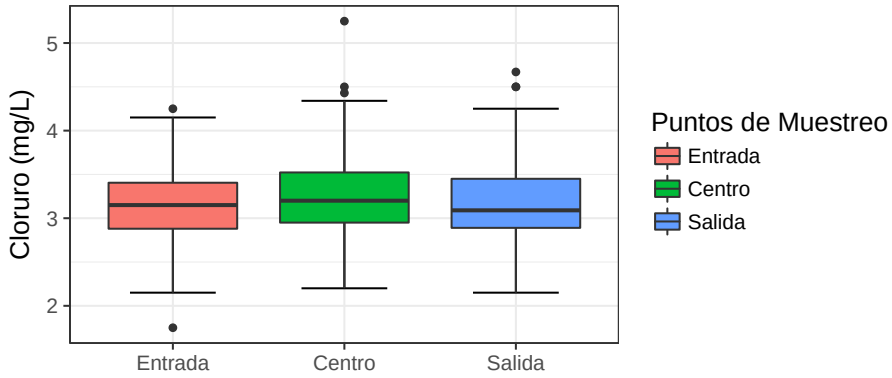

Figura 4: Gráfico de cajas y bigotes de las concentraciones del ion Cloruro en los tres puntos de muestreo del Embalse de Yacyretá

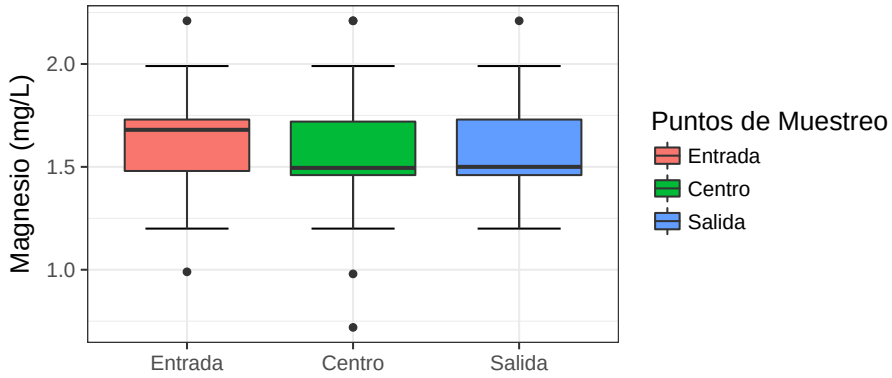

Figura 5: Gráfico de cajas y bigotes de las concentraciones del ion Magnesio en los tres puntos de muestreo del Embalse de Yacyretá

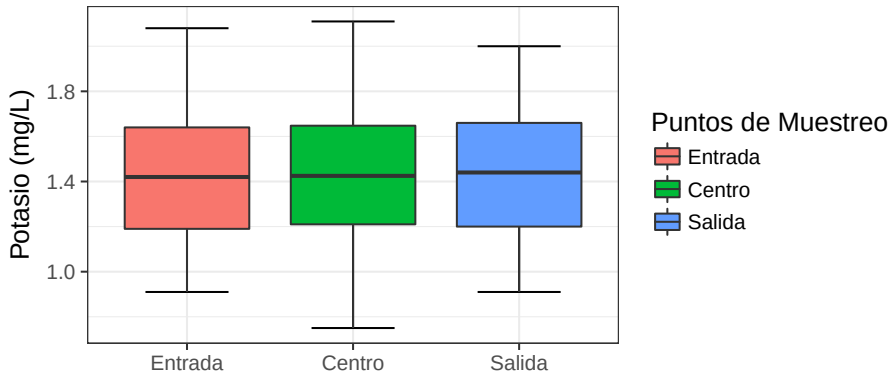

Figura 6: Gráfico de cajas y bigotes de las concentraciones del ion Potasio en los tres puntos de muestreo del Embalse de Yacyretá

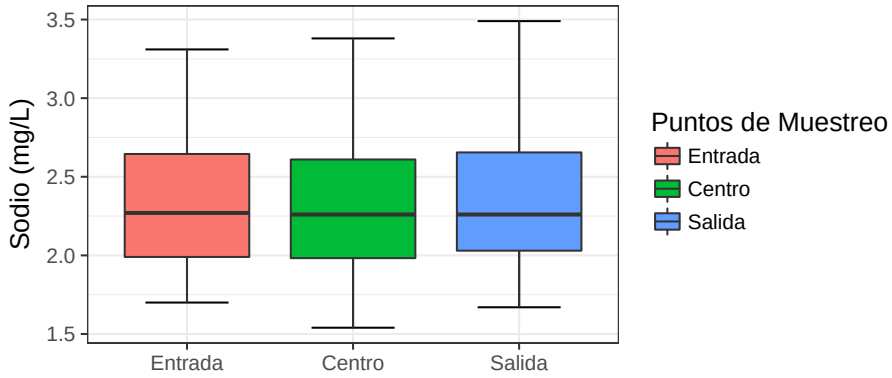

Figura 7: Gráfico de cajas y bigotes de las concentraciones del ion Sodio en los tres puntos de muestreo del Embalse de Yacyretá

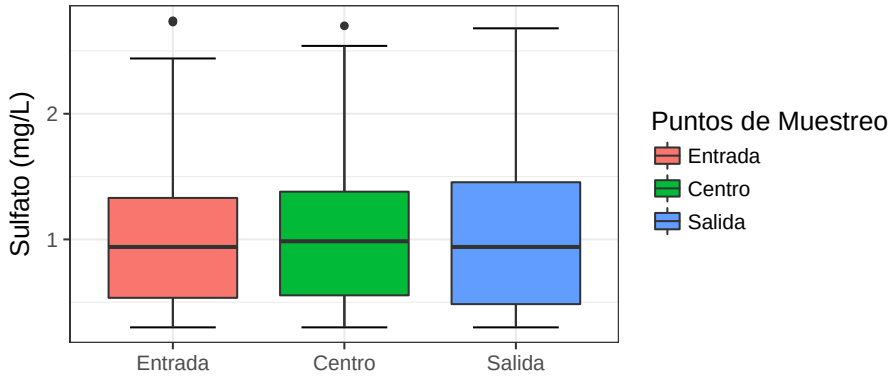

Figura 8: Gráfico de cajas y bigotes de las concentraciones del ion Sulfato en los tres puntos de muestreo del Embalse de Yacyretá

4.1.2. Comportamiento multivariado de los iones

De las Tablas 2 a 5 y de las Figuras 9 a 17 se destacan las siguientes conclusiones respecto del comportamiento global de las variables en estudio en las diferentes estaciones de muestreo (aquí los subíndices E, C y S indican "Entrada", "Centro" y "Salida" del Embalse, respectivamente):

i. De la Matriz de Correlaciones Bivariadas (Tabla 2) para todo el embalse puede observarse que:

   - Existe una correlación alta y significativa (al 5 %) entre Sodio y Potasio ($r = 0,80$, $p < 0,0001$).
     - La correlación puede considerarse moderada entre Potasio y Bicarbonato ($r = 0,56$; $p < 0,0001$), entre Calcio y Bicarbonato ($r = 0,55$; $p < 0,0001$), entre Potasio y Calcio ($r = 0,53$; $p < 0,0001$) y entre Sodio y Bicarbonato ($r = 0,5$; $p < 0,0001$).
   - Correlaciones de leves a bajas se observan en los siguientes pares de iones: Sodio y Calcio ($r = 0,43$; $p < 0,0001$), Sulfato y Cloruro ($r = -0,35$; $p < 0,0001$), Sulfato y Sodio ($r = -0,35$; $p < 0,0001$), Sodio y Cloruro ($r = 0,29$; p< $0,0001$), Magnesio y Cloruro ($r = 0,25$; $p < 0,0001$), Sulfato y Potasio ($r = -0,21$; $p < 0,0001$), Magnesio y Calcio ($r = -0,11$; $p = 0,04$).
   - No se observa una correlación significativa entre los restantes pares de iones.

ii. Realizando un análisis bivariado de las correlaciones por estación de muestreo (Tablas 3, 4, 5), se observa que:

- Los iones Sodio y Potasio mantienen una correlación alta en los tres puntos ($r_E = 0,81$; $r_C = 0,79$; $r_S = 0,80$; $p < 0,0001$).
- Existe una correlación moderada entre los iones Potasio y Bicarbonato ($r_E = 0,57$; $r_C = 0,56$; $r_S = 0,54$; $p < 0,0001$), entre Calcio y Bicarbonato ($r_E = 0,56$; $r_C = 0,51$; $r_S = 0,60$; $p < 0,0001$), entre Potasio y Calcio ($r_E = 0,56$; $r_C = 0,49$; $r_S = 0,54$; $p < 0,0001$) y entre Sodio y Bicarbonato ($r_E = 0,48$; $r_C = 0,53$; $r_S = 0,48$; $p < 0,0001$).
- Leves a bajas correlaciones significativas existen entre Sodio y Calcio ($r_E = 0,46$; $r_C = 0,38$; $r_S = 0,46$; $p < 0,0001$), entre Sulfato y Cloruro ($r_E = -0,40$; $r_C = -0,33$; $r_S = -0,35$; $p < 0,001$, $p = 0,0004$, $p = 0,0001$, respectivamente), entre Sulfato y Sodio ($r_E = -0,31$; $r_C = -0,37$; $r_S = -0,38$; $p = 0,0008$, $p < 0,0001$, $p < 0,0001$, respectivamente) y entre Sodio y Cloruro ($r_E = 0,32$; $r_C = 0,28$; $r_S = 0,26$; $p = 0,0005$, $p = 0,0023$, $p = 0,0049$, respectivamente).
- Existen pares de iones que cuentan con correlaciones significativas en algunos puntos y en otros no: Magnesio y Cloruro, Sulfato y Potasio presentan una correlación baja pero significativa en el Centro y Salida del sistema, no así en la Entrada. Por último, Magnesio y Calcio solo presentan una correlación que es significativa en la Entrada del sistema ($r_E = -0,19$; $p = 0,04$).
- Los histogramas mostrados en las diagonales de las Figuras 9, 10,11 indican que las distribuciones de los siete iones en cada estación de muestreo no se distribuyen como una normal.

iii. La evolución de los iones a lo largo de los casi 10 años tomados para el estudio, en cada uno de los tres puntos de muestreo, es algo cambiante conforme se dieron las mediciones en cada mes. Se puede decir que Bicarbonato es el que presenta una variabilidad mayor que las demás según se observa mes tras mes. Sin embargo, Cloruro y Sulfato también presentan una variabilidad a tener en cuenta. Por otro lado, pareciera que Sodio tiene una tendencia a aumentar en sus concentraciones durante los meses transcurridos. En síntesis, de manera general, podría decirse que los siete iones van siguiendo una tendencia casi constante o con leves aumentos durante los meses en los años estudiados (Figura 28 al 30).

iv. En las Figuras 12 a 17 cada estrella y cada rostro representan las características de los iones en un mes. Estas figuras muestran que, comparando por estación de muestreo, existen algunas pequeñas diferencias entre las concentraciones de los iones en meses iguales, incluso, en algunos casos las discrepancias son insignificantes. Se destacan:

- Al inicio del período de medición febrero-marzo del 2001 las aguas presentan altas concentraciones de sulfato, mientras que a partir de enero de 2008 este anión desaparece o está presente en concentraciones muy bajas en las aguas del Embalse. A partir de 2008 aparecen aguas donde, salvo sulfato, todas las componentes están presentes.
- Respecto del comportamiento de los meses a través de los años:
  • En enero de 2001 a 2004, el cloruro y el magnesio están presentes en mayor medida que los otros componentes; sin embargo, a partir de 2005 (salvo alguna diferencia como en enero de 2008 y enero de 2010) todos las componentes están presentes distribuyéndose en proporciones similares en igual medida en las muestras.
  • En febrero las aguas comienzan teniendo sulfato, se le suma el cloruro, luego un poco de todas y finalmente las muestras no tienen presencia notable de sulfato.

Cabe señalar que los pequeños puntos observados en las gráficas de estrellas indican que en esos meses no se realizaron mediciones de los siete iones.

En las Caras de *Chernoff* cada uno de los iones en estudio ha sido asociado e identificado con una característica facial, tal y como se señala a continuación:

- Bicarbonato: altura de la cara, ancho de ojos y altura de la oreja.
- Calcio: ancho de cara y altura del cabello.
- Cloruro: estructura de la cara y ancho de cabello.
- Magnesio: altura de la boca y estilo de cabello.
- Potasio: ancho de boca y altura de la nariz.
- Sodio: sonrisa y ancho de nariz.
- Sulfato: altura de ojos y ancho de la oreja.

Tabla 2: Matriz de correlación y las significancias estadísticas entre los iones en todo el Embalse de Yacyretá

		Alcalinidad Total	Calcio	Cloruro	Magnesio	Potasio	Sodio	Sulfato
Alcalinidad Total	Correlación	1						
	Sig. Bilateral							
Calcio	Correlación	0,5534	1					
	Sig. Bilateral	<0,0001						
Cloruro	Correlación	-0,0883	-0,0682	1				
	Sig. Bilateral	0,1019	0,2073					
Magnesio	Correlación	0,0487	-0,1112	0,2530	1			
	Sig. Bilateral	0,3680	0,0393	<0,0001				
Potasio	Correlación	0,5557	0,5315	0,0673	0,0100	1		
	Sig. Bilateral	<0,0001	<0,0001	0,2129	0,8530			
Sodio	Correlación	0,4958	0,4332	0,2860	0,0613	0,8006	1	
	Sig. Bilateral	<0,0001	<0,0001	<0,0001	0,2571	<0,0001		
Sulfato	Correlación	-0,0729	-0,0575	-0,3539	-0,0537	-0,2141	-0,3524	1
	Sig. Bilateral	0,1776	0,2876	<0,0001	0,3206	<0,0001	<0,0001	

Tabla 3: Matriz de correlación y las significancias estadísticas entre los iones en la Entrada del Embalse de Yacyretá

		Alcalinidad Total	Calcio	Cloruro	Magnesio	Potasio	Sodio	Sulfato
Alcalinidad Total	Correlación	1						
	Sig. Bilateral							
Calcio	Correlación	0,5637	1					
	Sig. Bilateral	<0,0001						
Cloruro	Correlación	-0,0569	-0,1226	1				
	Sig. Bilateral	0,5457	0,1918					
Magnesio	Correlación	0,0842	-0,1919	0,1357	1			
	Sig. Bilateral	0,3712	0,0400	0,1482				
Potasio	Correlación	0,5673	0,5635	0,1092	-0,0298	1		
	Sig. Bilateral	<0,0001	<0,0001	0,2452	0,7515			
Sodio	Correlación	0,4787	0,4605	0,321	-0,0040	0,8127	1	
	Sig. Bilateral	<0,0001	<0,0001	0,0005	0,9663	<0,0001		
Sulfato	Correlación	0,0122	-0,0361	-0,4079	-0,0376	-0,1492	-0,3069	1
	Sig. Bilateral	0,8972	0,7019	<0,0001	0,6899	0,1115	0,0008	

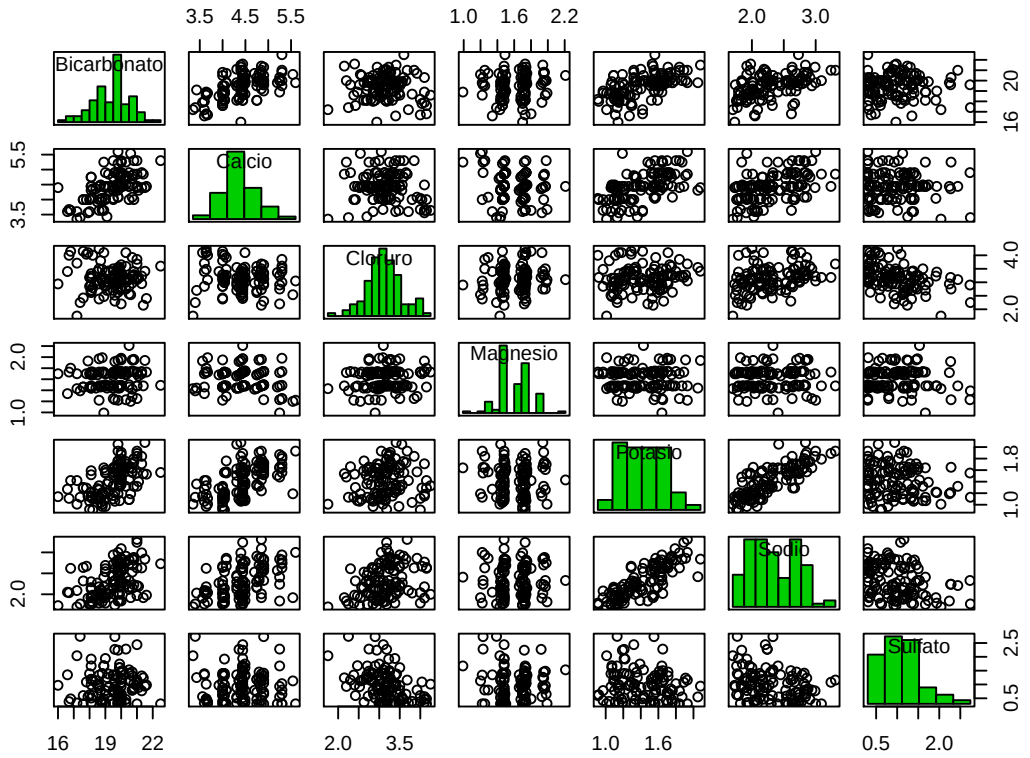

Figura 9: Matriz de diagramas de dispersión entre los siete iones con sus distribuciones en la diagonal principal en la Entrada del Embalse de Yacyretá

Tabla 4: Matriz de correlación y las significancias estadísticas entre los iones en el Centro del Embalse de Yacyretá

		Alcalinidad Total	Calcio	Cloruro	Magnesio	Potasio	Sodio	Sulfato
Alcalinidad Total	Correlación	1						
	Sig. Bilateral							
Calcio	Correlación	0,5137	1					
	Sig. Bilateral	<0,0001						
Cloruro	Correlación	-0,056	-0,0456	1				
	Sig. Bilateral	0,5542	0,6302					
Magnesio	Correlación	0,0660	-0,1088	0,3345	1			
	Sig. Bilateral	0,4854	0,2491	0,0003				
Potasio	Correlación	0,5627	0,4893	0,0155	-0,0016	1		
	Sig. Bilateral	<0,0001	<0,0001	0,8698	0,9869			
Sodio	Correlación	0,5295	0,3773	0,2827	0,1139	0,7893	1	
	Sig. Bilateral	<0,0001	<0,0001	0,0023	0,2274	<0,0001		
Sulfato	Correlación	-0,0936	0,0015	-0,3256	-0,1165	-0,2338	-0,3682	1
	Sig. Bilateral	0,3218	0,9877	0,0004	0,217	0,0123	<0,0001	

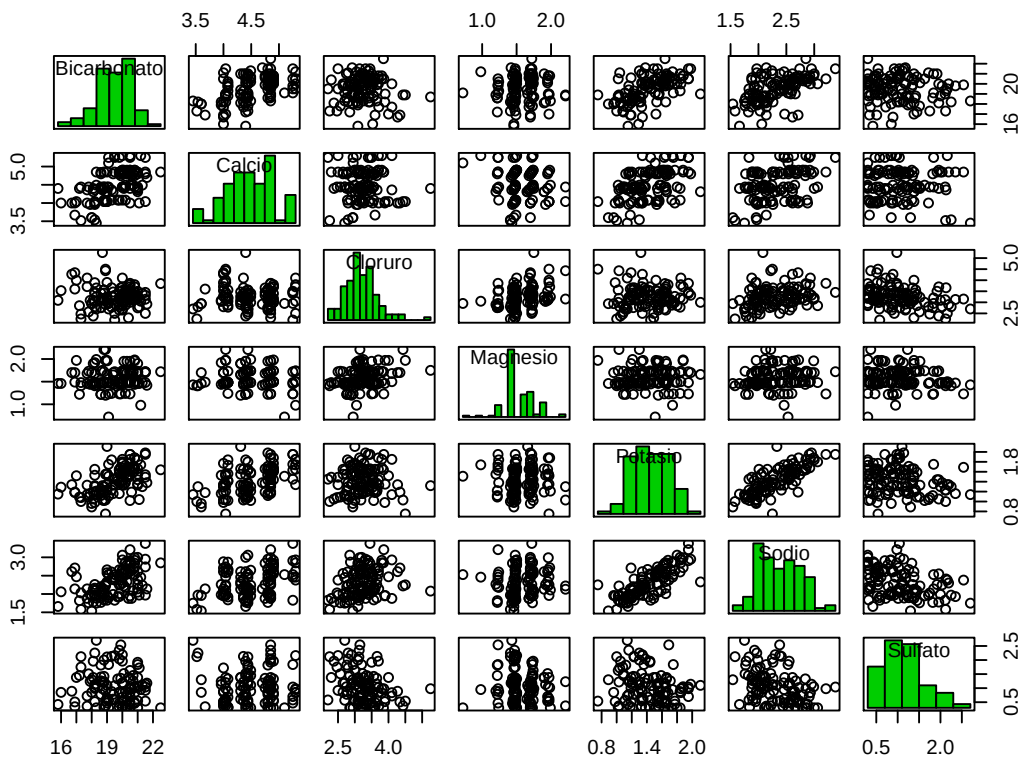

Figura 10: Matriz de diagramas de dispersión entre los siete iones con sus distribuciones en la diagonal principal en el Centro del Embalse de Yacyretá

Tabla 5: Matriz de correlación y las significancias estadísticas entre los iones en la Salida del Embalse de Yacyretá

		Alcalinidad Total	Calcio	Cloruro	Magnesio	Potasio	Sodio	Sulfato
Alcalinidad Total	Correlación	1						
	Sig. Bilateral							
Calcio	Correlación	0,5961	1					
	Sig. Bilateral	<0,0001						
Cloruro	Correlación	-0,1565	-0,054	1				
	Sig. Bilateral	0,0948	0,5666					
Magnesio	Correlación	-0,0044	-0,017	0,2919	1			
	Sig. Bilateral	0,9625	0,8572	0,0015				
Potasio	Correlación	0,5381	0,5423	0,0783	0,0653	1		
	Sig. Bilateral	<0,0001	<0,0001	0,4053	0,4879			
Sodio	Correlación	0,4764	0,4622	0,2608	0,0699	0,8002	1	
	Sig. Bilateral	<0,0001	<0,0001	0,0049	0,4577	<0,0001		
Sulfato	Correlación	-0,1336	-0,1459	-0,3503	0,006	-0,2584	-0,3818	1
	Sig. Bilateral	0,1547	0,1198	0,0001	0,9496	0,0053	<0,0001	

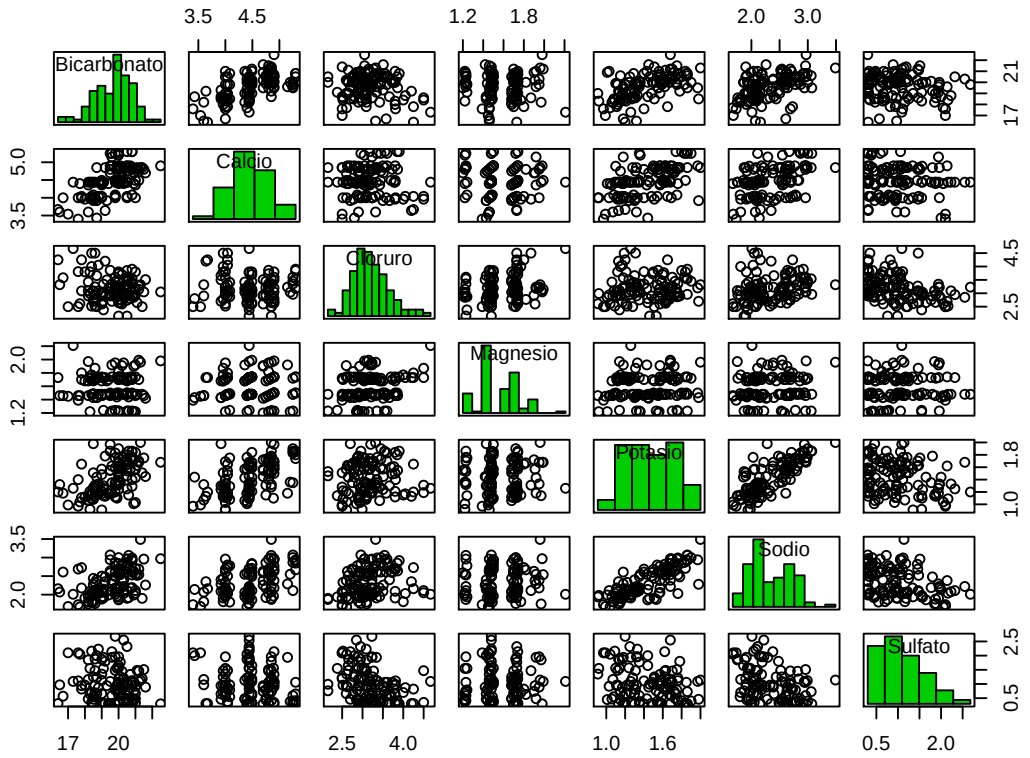

Figura 11: Matriz de diagramas de dispersión entre los siete iones con sus distribuciones en la diagonal principal en la Salida del Embalse de Yacyretá

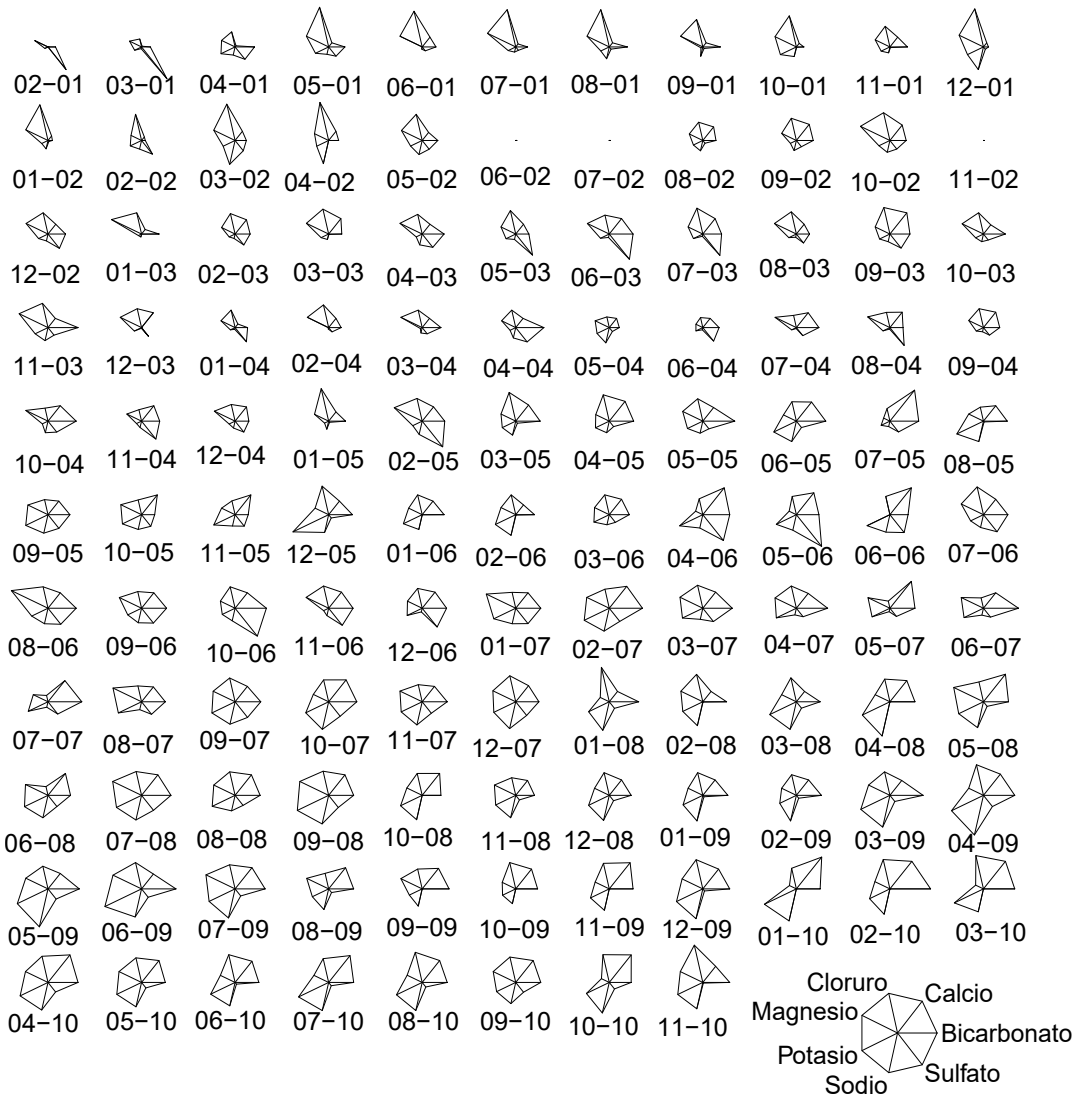

Figura 12: Gráfico de estrellas de las concentraciones de los iones en la Entrada del Embalse de Yacyretá

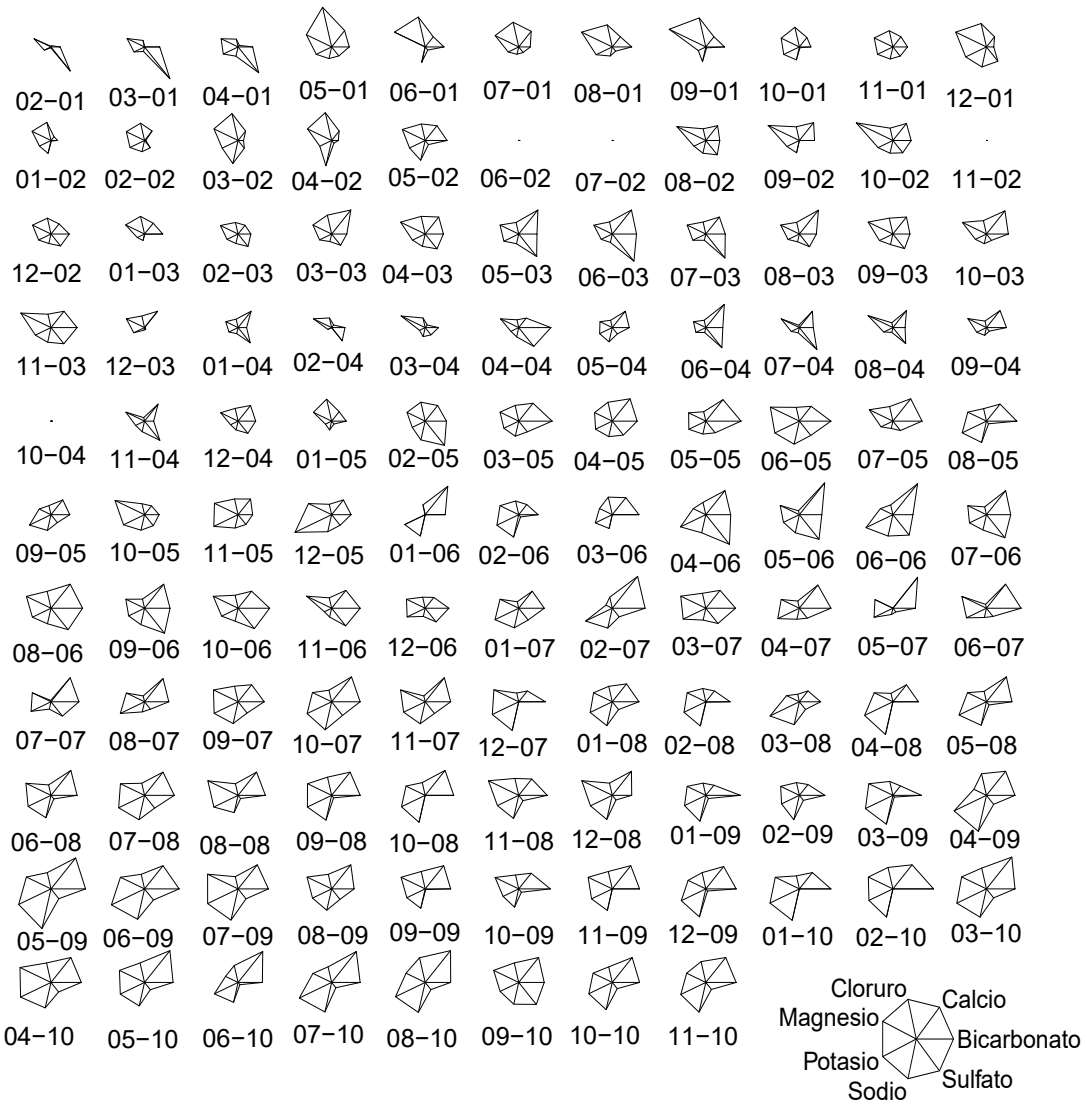

Figura 13: Gráfico de estrellas de las concentraciones de los iones en el Centro del Embalse de Yacyretá

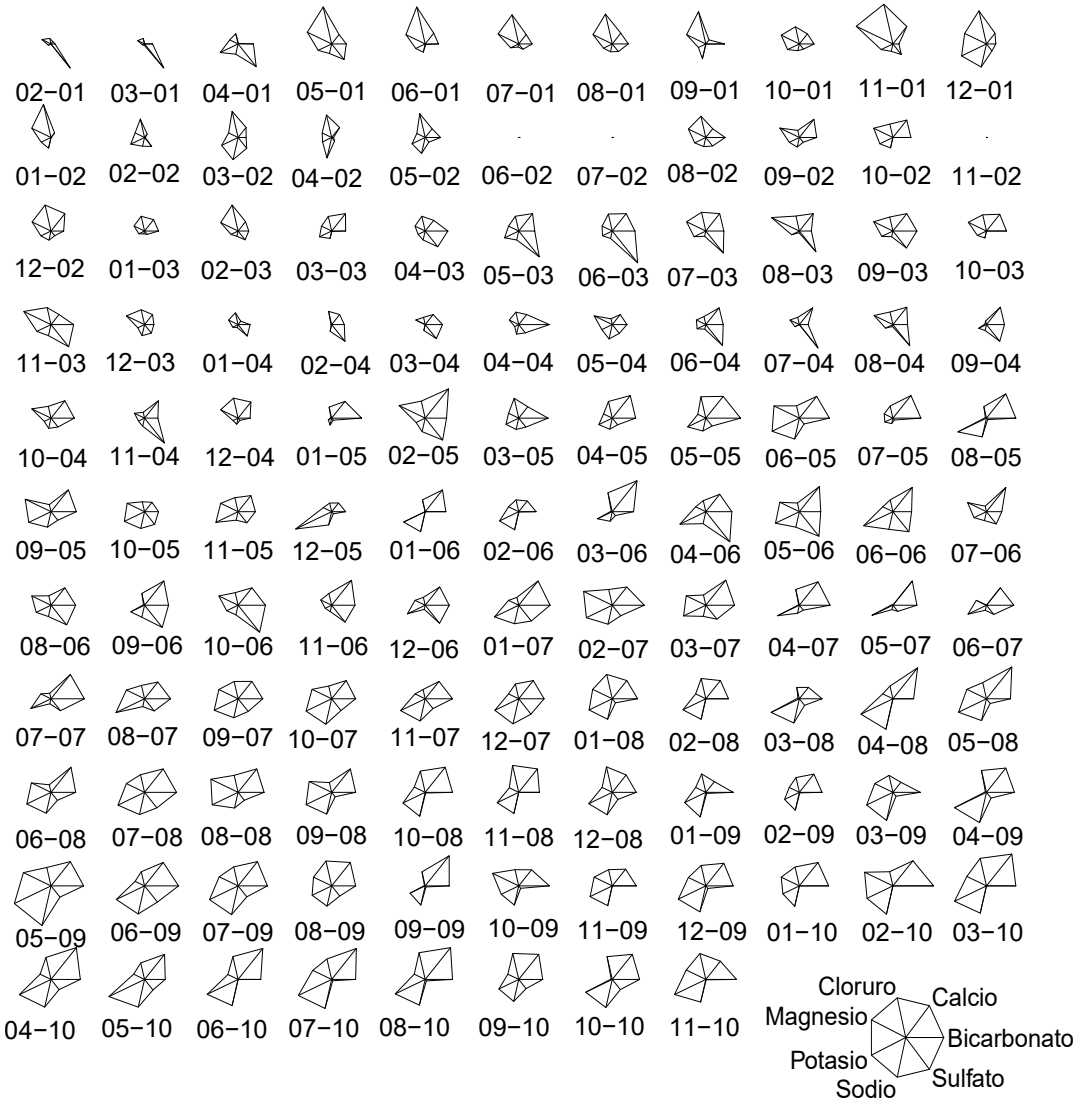

Figura 14: Gráfico de estrellas de las concentraciones de los iones en la Salida del Embalse de Yacyretá

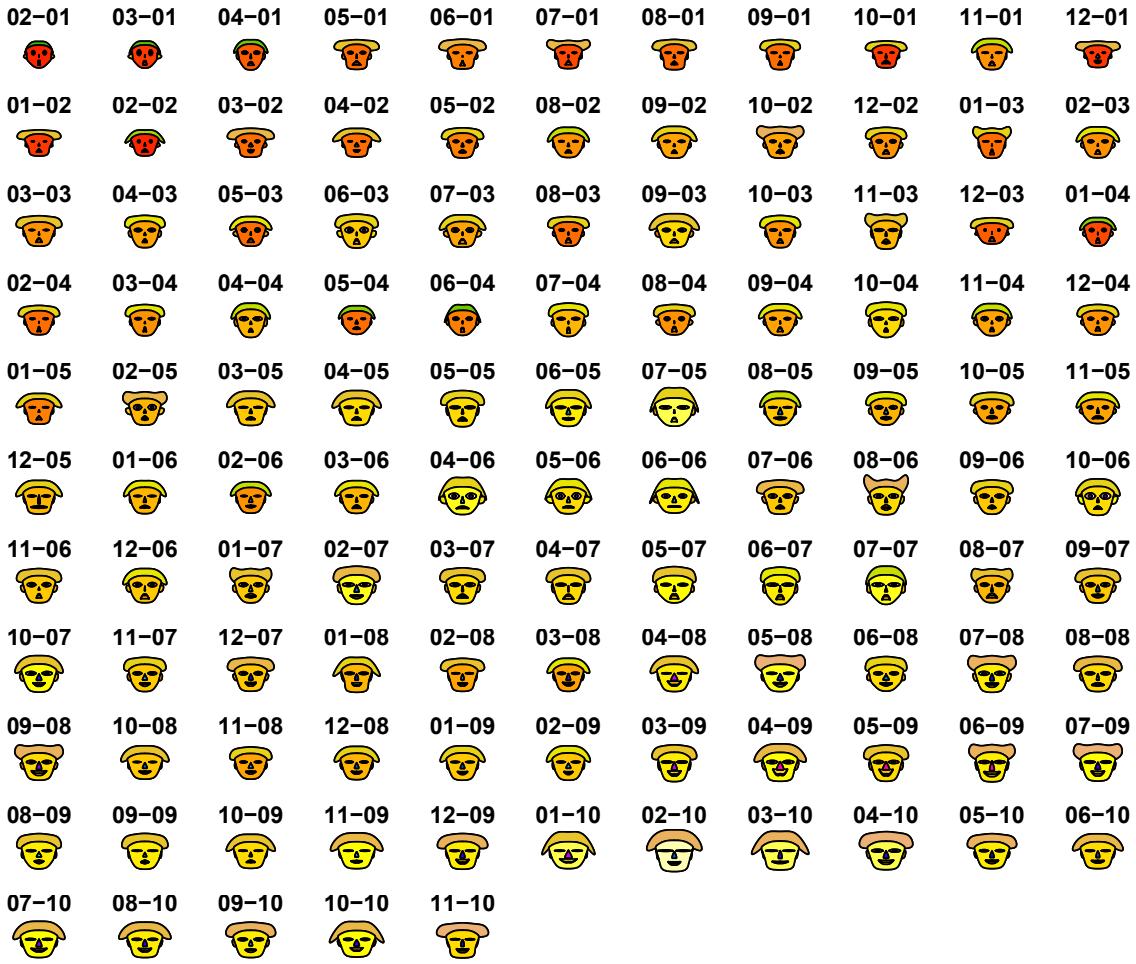

Figura 15: Caras de Chernoff para las concentraciones de los iones en la Entrada del Embalse de Yacyretá

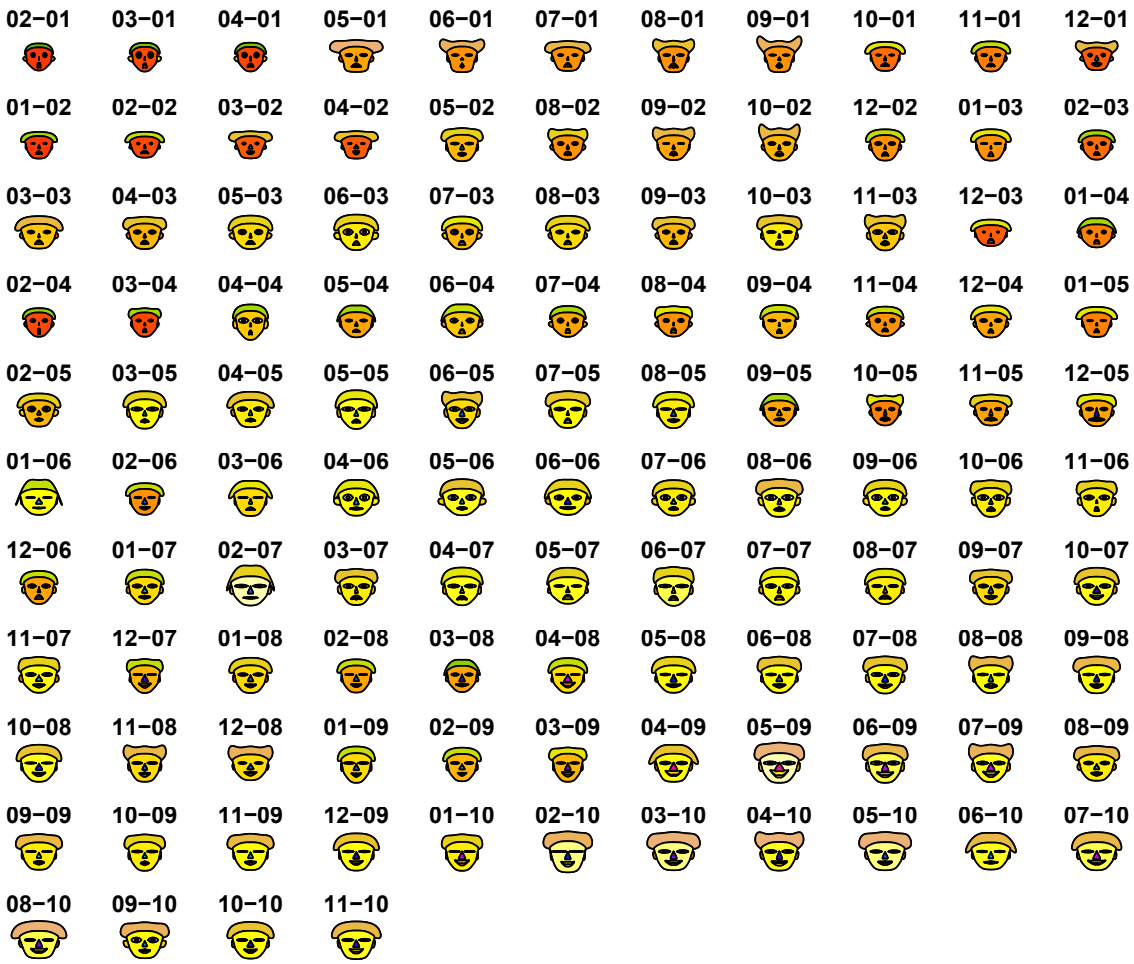

Figura 16: Caras de Chernoff para las concentraciones de los iones en el Centro del Embalse de Yacyretá

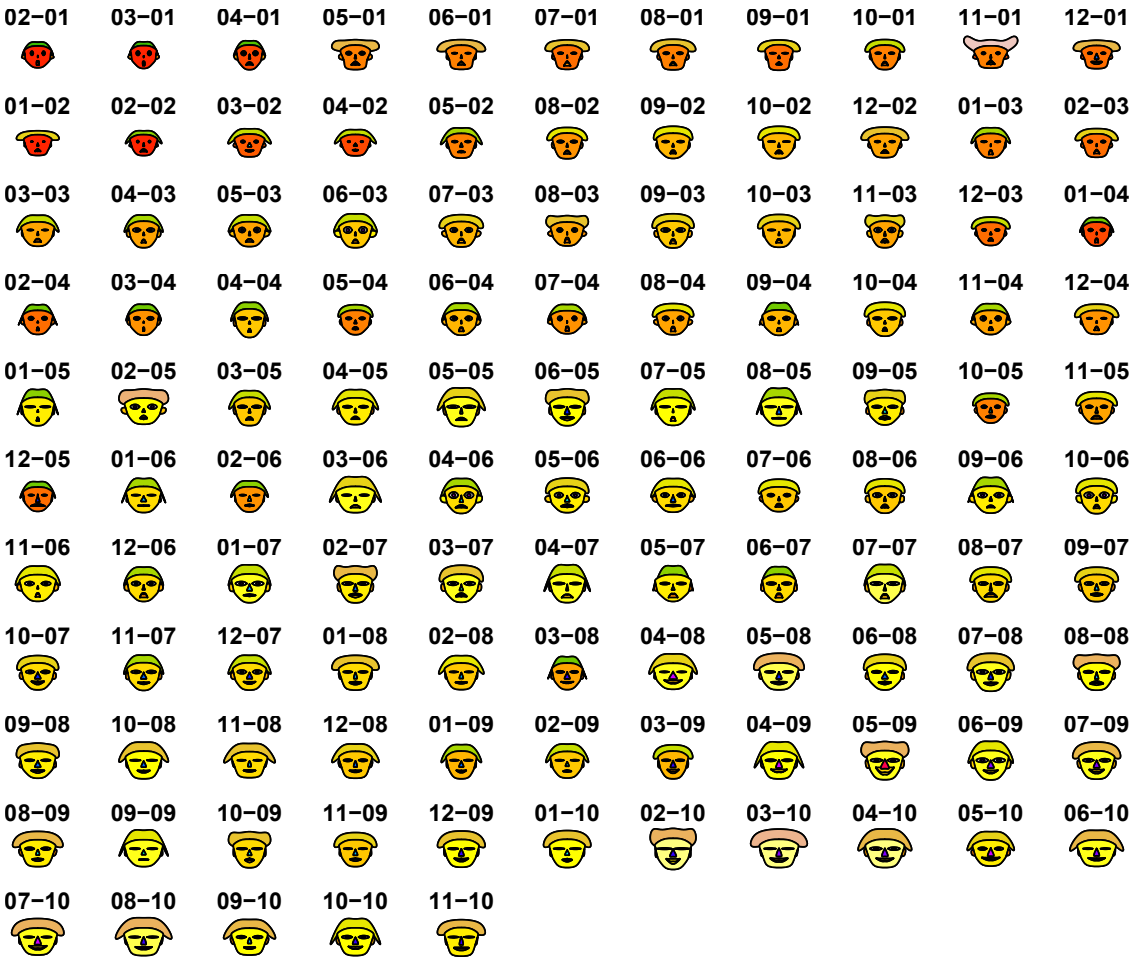

Figura 17: Caras de Chernoff para las concentraciones de los iones en la Salida del Embalse de Yacyretá

4.1.3. Comportamiento de los iones según el Análisis de Componentes Principales

En primer lugar, antes de aplicar el método del Análisis de Componentes Principales (ACP), se procede a realizar algunas consideraciones para tener un indicio de que el método es aplicable.

Se ha mostrado, en la sección previa, a partir de la Matriz de Correlaciones Bivariante entre los siete iones, que algunos de ellos se relacionan de manera significativa. A partir de la misma matriz, se analiza el determinante y el valor del índice de Kaiser-Meyer-Olkin (KMO), para ver si se puede factorizar las variables originales de forma eficiente. En la Tabla 6 se observa que los valores de los determinantes son muy próximos a cero, indicando por tal motivo, posibles redundancia de las variables. Además, los índices KMO son relativamente altos, considerados como "regular o mediocre", pero que están al límite de ser aceptables para la aplicación del ACP. Así entonces, se procede a aplicar el método de Análisis de Componentes Principales con los siete iones considerados en el presente estudio.

Tabla 6: Determinantes e Índice KMO de las Matrices de Correlación de todo el Embalse de Yacyretá y de todas las Estaciones de Muestreo

		Determinante	Índice KMO
Matriz de Correlación	Total Embalse	0,089	0,69
	Entrada	0,074	0,68
	Centro	0,090	0,66
	Salida	0,083	0,69

Para la extracción de las componentes principales en cada uno de los puntos de muestreo se ha utilizado la matriz de correlaciones. En la "Entrada del Embalse" se ha encontrado que con tres componentes principales se explica aproximadamente el 77 % de la variabilidad total de los datos. Se han seleccionado estas tres componentes principales tal como lo recomienda el gráfico de sedimentación (Figura 18) al igual que los valores propios (Tabla 7) cuyos valores son mayores a la unidad hasta la tercera componente y, además, explican valores razonables de proporciones de variabilidad total.

Tabla 7: Resumen del Análisis de Componentes Principales en la Entrada del Embalse de Yacyretá, período 2001-2010

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Varianza	2,79	1,57	1,03	0,61	0,47	0,37	0,15
% de varianza explicada	39,9	22,4	14,8	8,8	6,7	5,3	2,2
% varianza acumulada	39,9	62,3	77,0	85,8	92,5	97,8	100,0

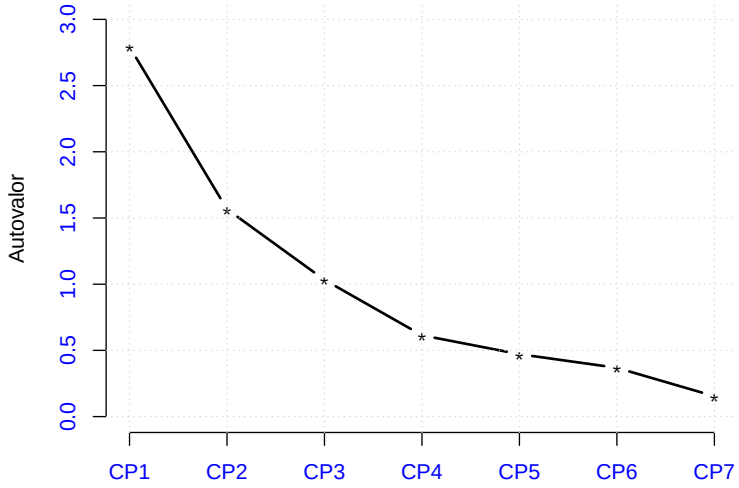

Figura 18: Gráfico de sedimentación en la Entrada del Embalse de Yacyretá

Por otra parte, en la Tabla 8 se puede apreciar las correlaciones entre las tres componentes principales seleccionadas y los siete iones. En ella se observa que existe una clara estructura de los datos ya que se pueden definir correlacionalmente las variables en diferentes componentes principales sin la necesidad de realizar algún tipo de rotación.

La primera componente principal, que explica el 39,9 %, está altamente relacionada de manera positiva con el anión Bicarbonato y los cationes Calcio, Potasio y Sodio. Se puede decir entonces, que la primera componente principal está asociada a la contribución mayoritaria (excepto Magnesio por su ausencia en esta primera componente principal) al Bicarbonato del agua.

La segunda componente principal, que explica el 22,4 %, está altamente relacionada de manera negativa y positiva con los aniones Cloruro y Sulfato, respectivamente. Esta segunda componente principal puede interpretarse como aquella que determina las concentraciones de los contribuyentes pobres al Bicarbonato del agua.

Mientras que, la tercera componente principal, que explica el 14,8 %, está altamente relacionada de manera negativa con el catión Magnesio. Esta tercera componente principal es la que determina entonces la concentración de Magnesio en el agua.

La Figura 19 muestra, de forma visual, estas relaciones conforme a las altas correlaciones entre las componentes principales y cada una de las variables analizadas. Se puede ver como las variables Bicarbonato, Calcio, Potasio y Sodio quedan explicadas por el primer eje principal, mientras que las variables sulfato y cloruro quedan explicadas por el segundo eje principal y, además, se puede notar que Magnesio no está bien explicada por los dos primeros ejes principales (Figura 19 a y b). El tercer eje principal es el que explica altamente la variable Magnesio, no así a los demás iones (Figura 19 c).

Tabla 8: Correlación entre iones y las tres componentes principales retenidas en la Entrada del Embalse de Yacyretá

Matriz de Componentes no rotados			
	CP1	CP2	CP3
Bicarbonato	0,74	0,29	-0,31
Calcio	0,74	0,39	0,13
Cloruro	0,20	-0,82	0,07
Magnesio	-0,04	-0,33	-0,91
Potasio	0,90	0,04	-0,02
Sodio	0,88	-0,22	0,04
Sulfato	-0,28	0,70	-0,28

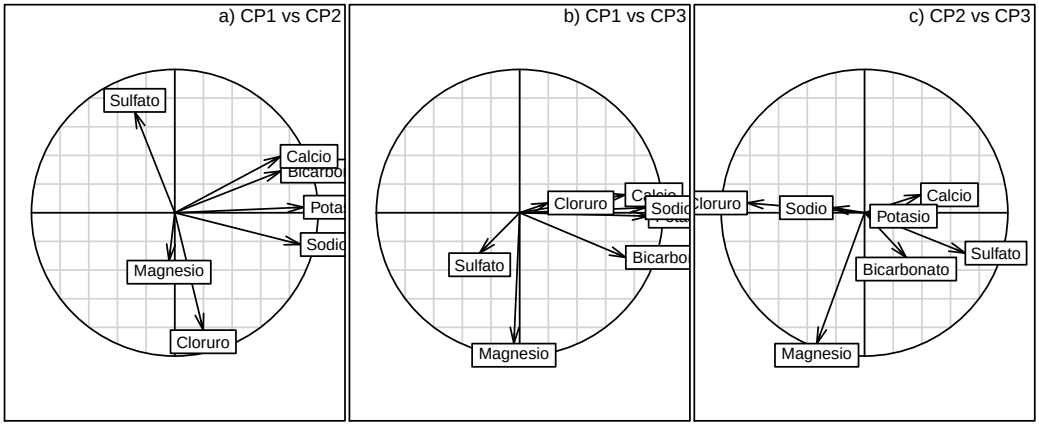

Figura 19: Circulo de correlaciones entre los siete iones y las tres componentes principales retenidas en la Entrada del Embalse de Yacyretá

La Tabla 9 y la Figura 20 sugieren que en el "Centro del Embalse" con tres componentes principales se logra captar aproximadamente el 75 % de la variabilidad total, con lo cual se retienen estas tres componentes principales. Se selecciona también la tercera componente a pesar de que su autovalor asociado es menor que la unidad pero es relativamente cercano a este valor y, además, esta tercera componente explica un porcentaje alto de variabilidad (casi 13 %). En el caso del Centro del sistema, se opta por rotar los tres primeros ejes principales utilizando la rotación ortogonal "varimax" ya que existían altas correlaciones entre algunos iones con más de una componente principal. Al realizar la rotación mencionada se logra que cada uno de los iones tenga altas correlaciones con una sola componente principal, aunque podría decirse que el ion Cloruro mantiene moderadas correlaciones con la segunda y la tercera componente. Sin embargo, se resuelve dejar este ion en la segunda componente principal por tener una leve mayor correlación con esta.

Tabla 9: Resumen del Análisis de Componentes Principales en el Centro del Embalse de Yacyretá, período 2001-2010

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Varianza sin rotar	2,75	1,6	0,9	0,65	0,54	0,41	0,16
% de varianza explicada sin rotar	39,3	22,8	12,8	9,2	7,7	5,9	2,3
% varianza acumulada sin rotar	39,3	62,1	74,9	84,2	91,9	97,7	100
% de varianza explicada CP rotadas	37	20	18				
% varianza acumulada CP rotadas	37	57	75				

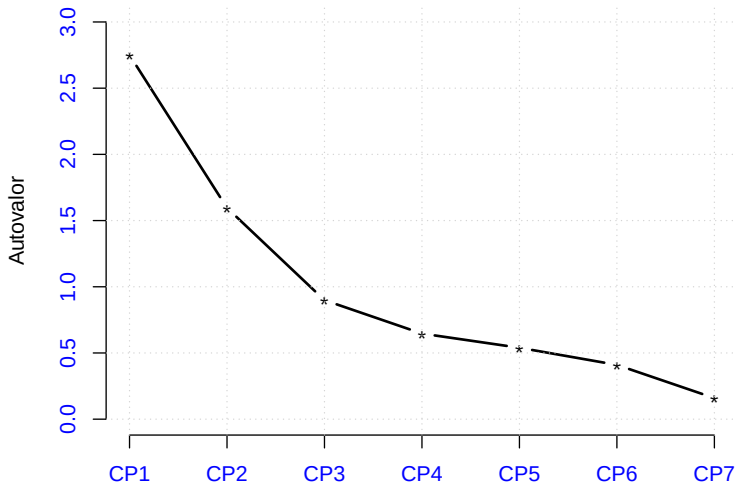

Figura 20: Gráfico de sedimentación en el Centro del Embalse de Yacyretá

La Tabla 10 proporciona las correlaciones de los iones con las tres componentes principales conservadas. Al igual que en la Entrada, se logra una misma clasificación de las variables a diferencia de algunos signos opuestos especialmente en las componentes segunda y tercera. Al realizar la rotación, las proporciones de variabilidad explicada por cada componente varía levemente con relación a las que se obtuvieron sin realizar la rotación, lo cual teóricamente sucede en general. En este sentido, entonces, se tiene que el 37 % aproximadamente de la variabilidad total es explicada por la primera componente principal, que se relaciona de manera significativa con el anión Bicarbonato y los cationes Calcio, Potasio y Sodio todas ellas con correlaciones positivas con esta componente principal. Por otra parte, aproximadamente, el 20 % de la variabilidad total queda explicada por la segunda componente principal, que se relaciona de manera directa con el anión Cloruro e inversamente con el anión Sulfato. Mientras que la tercera componente está altamente relacionada con el ion Magnesio y este componente absorbe, aproximadamente, el 18 % de la variabilidad total.

La interpretación de las tres componentes principales, en el "Centro del Embalse", es la misma que en la "Entrada" por la obtención de la misma clasificación de iones en las tres componentes principales como puede apreciarse en la Tabla 10 y en la Figura 21

Tabla 10: Correlación entre iones y las tres componentes principales retenidas, rotadas y no rotadas en el Centro del Embalse de Yacyretá

	Componentes no rotadas			Componentes rotadas		
	CP1	CP2	CP3	CP1	CP2	CP3
Bicarbonato	0,76	-0,26	0,26	0,83	-0,07	0,09
Calcio	0,65	-0,41	0,14	0,76	-0,14	-0,11
Cloruro	0,2	0,78	0,02	-0,06	0,58	0,56
Magnesio	0,11	0,63	0,69	0,02	0	0,94
Potasio	0,88	-0,14	-0,07	0,85	0,27	-0,06
Sodio	0,89	0,16	-0,1	0,75	0,49	0,12
Sulfato	-0,39	-0,56	0,56	-0,08	-0,88	0,01

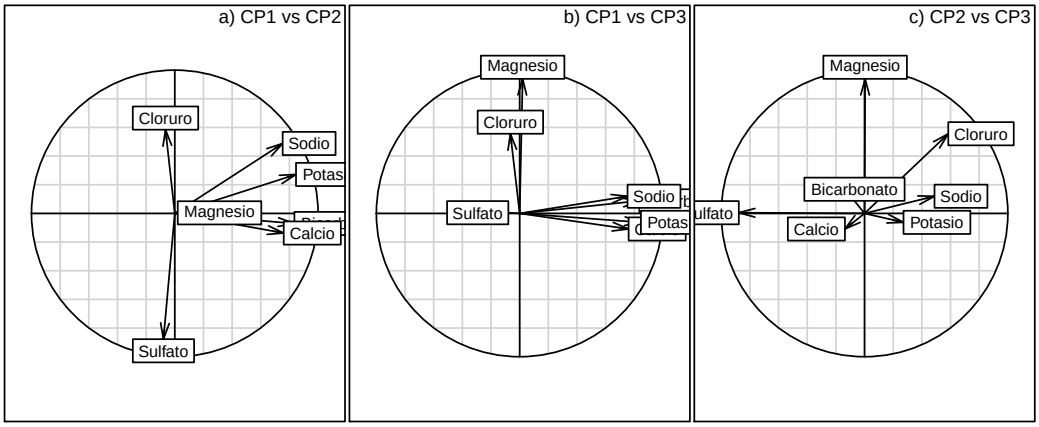

Figura 21: Circulo de correlaciones entre los siete iones y las tres componentes principales retenidas (rotadas) en el Centro del Embalse de Yacyretá

Por último, así como en la "Entrada" y el "Centro" del Embalse, en la "Salida" se seleccionaron tres componentes principales cuyo poder de explicación de la variabilidad total es de, aproximadamente, 76 %. Esto se aprecia en la Tabla 11, lo que implica la retención de tres componentes principales ya que los autovalores asociados son superiores a la unidad, excepto el de la tercera componente principal. Sin embargo, como esta última componente principal explica un alto porcentaje de la variabilidad total (14,1 %) y tiene un autovalor muy próximo a uno, también es retenida. Además, al igual que en el Centro del sistema, se realiza una rotación varimax de las componentes principales para obtener correlaciones más adecuadas de cada ion con las componentes principales retenidas en principio.

Tabla 11: Resumen del Análisis de Componentes Principales en la Salida del Embalse de Yacyretá, período 2001-2010

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Varianza sin rotar	2,85	1,51	0,99	0,59	0,51	0,38	0,17
% de varianza explicada sin rotar	40,8	21,6	14,1	8,4	7,2	5,4	2,5
% varianza acumulada sin rotar	40,8	62,4	76,5	84,9	92,1	97,5	100
% de varianza explicada CP rotadas	38	22	16				
% varianza acumulada CP rotadas	38	60	76				

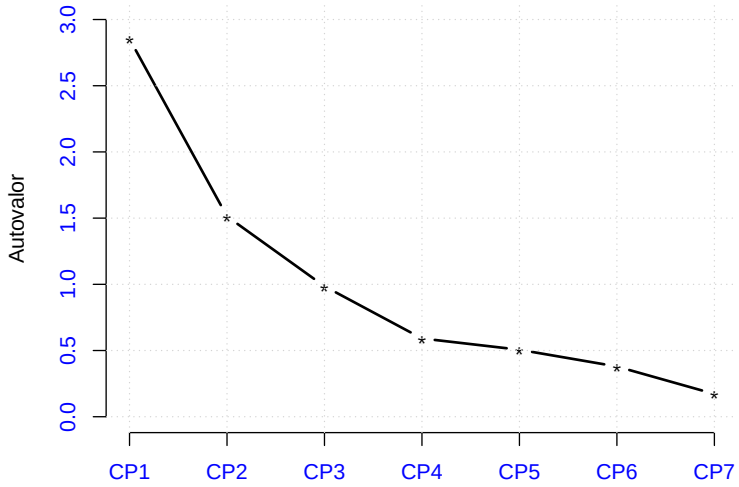

Figura 22: Gráfico de sedimentación en la Salida del Embalse de Yacyretá

En la Tabla 12 se muestran las correlaciones entre los iones con cada componente principal. Como ya se ha mencionado, para el análisis final, se han utilizado las correlaciones que corresponden a las componentes principales rotadas. Así, la primera componente principal explica el 38 % de la variabilidad total y está altamente correlacionada de manera positiva con los iones Bicarbonato, Calcio, Potasio y Sodio. Por otro lado, la segunda componente principal cubre un 22 % de variabilidad total correlacionándose altamente con los aniones Cloruro y Sulfato de manera positiva y negativa, respectivamente. Por último, la tercera componente que explica el 16 % de la variabilidad total, está altamente correlacionada de manera positiva con el catión Magnesio. Gráficamente esto se puede apreciar en la Figura 23 que muestra las direcciones de las correlaciones entre los iones y las tres componentes principales retenidas, identificando así una clasificación similar a las establecidas en la Entrada y en el Centro del Embalse.

Tabla 12: Correlación entre iones y las tres componentes principales retenidas, rotadas y no rotadas, Salida del Embalse de Yacyretá.

	Componentes no rotadas			Componentes rotadas		
	CP1	CP2	CP3	CP1	CP2	CP3
Bicarbonato	0,73	-0,39	0,15	0,83	-0,12	-0,05
Calcio	0,74	-0,31	0,11	0,81	-0,05	-0,04
Cloruro	0,17	0,86	-0,04	-0,12	0,75	0,43
Magnesio	0,08	0,51	0,8	0,03	0,02	0,95
Potasio	0,88	-0,04	0,06	0,84	0,25	0,07
Sodio	0,87	0,16	-0,06	0,75	0,46	0,08
Sulfato	-0,45	-0,49	0,55	-0,18	-0,82	0,18

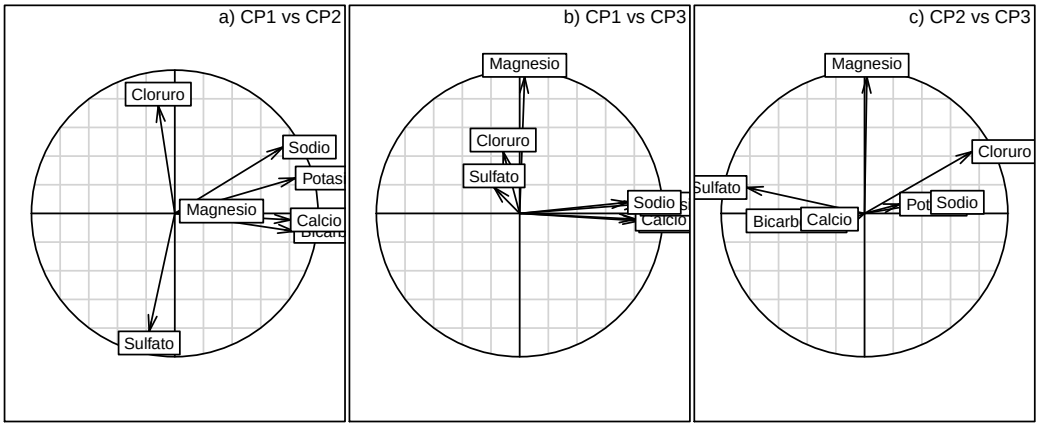

Figura 23: Circulo de correlaciones entre los siete iones y las tres componentes principales retenidas (rotadas) en la Salida del Embalse

En resumen, en cada uno de los mencionados puntos de muestreo se han observado comportamientos similares en cuanto a las correlaciones entre los siete iones y las tres componentes principales retenidas. Esto sugiere que, pasando de una estación de muestreo a otra, las distribuciones de las concentraciones medias de los diferentes iones estudiados durante los casi 10 años son relativamente homogéneas, es decir, en cada punto muestral analizado se presentaron concentraciones medias de los iones bastante similares.

4.1.4. Comportamiento de los iones según el Análisis de Conglomerados

Para el Análisis de Conglomerados, en cada uno de los tres puntos de muestreo, se utilizan métodos jerárquicos para la obtención de grupos de iones observando estructuras que se forman al crear los grupos. Se han probado varios métodos de clasificación obteniendo con ellos similares clasificaciones, por lo cual, se reproducen aquellos que fueron obtenidos por el método de "encadenamiento completo" (Complete Linkage) con el uso de correlaciones entre variables como medidas de similitud.

En la "Entrada del Embalse" (Figura 24) puede identificarse tres grupos. Estos grupos de variables coinciden exactamente con los obtenidos con el Análisis de Componentes Principales. Más precisamente, los grupos que identifican iones con similar comportamiento son:

Grupo 1: Bicarbonato, Calcio, Sodio y Potasio

Grupo 2: Cloruro y Sulfato

Grupo 3: Magnesio

La clasificación de los iones en cada uno de los grupos o clusters se puede observar en el dendrograma (Figura 24). Este gráfico sugiere el agrupamiento de las variables en tres grupos o cluster debido al mayor salto observado en el aglomerado de las mismas. Para visualizar mejor los tres grupos se separaron mediante rectángulos de diferentes colores. En el rectángulo rojo se aprecia el Grupo 1 (G1), en el azul el Grupo 2 (G2) y en el verde el Grupo 3 (G3).

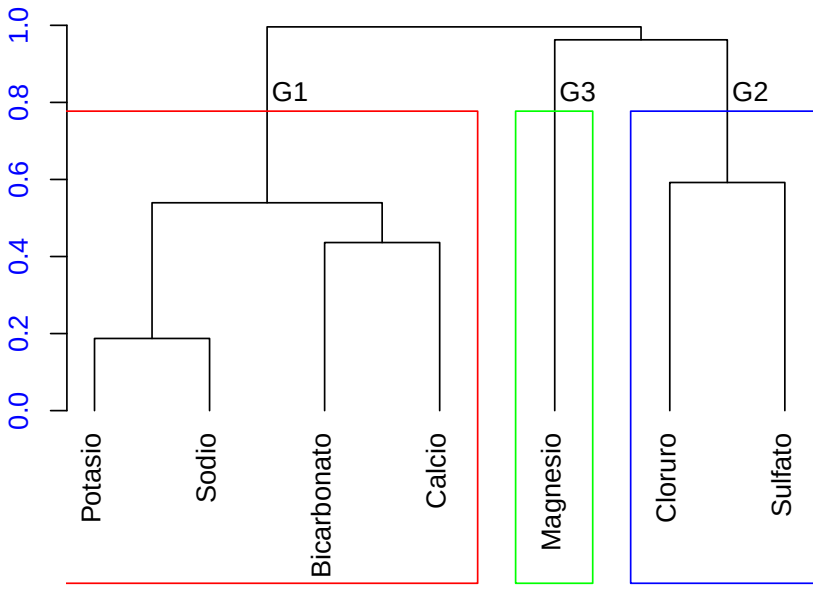

Figura 24: Dendrograma de iones en la Entrada del Embalse de Yacyretá

En el “Centro del Embalse”, se puede decir también que se forman tres grupos de variables. Sin embargo, esta vez solo uno de los grupos formados coincide con el obtenido en la “Entrada”. En efecto, el Grupo 1 del “Centro del Embalse” está conformada por los iones Bicarbonato, Calcio, Sodio y Potasio. La diferencia se encuentra en los dos grupos restantes. Mientras que en la Entrada en el Grupo 2 están los iones Cloruro y Sulfato, en el Centro se tienen los iones Cloruro y Magnesio. Es decir, Magnesio desplaza a Sulfato en el Centro del Embalse, quedando este último en el tercer grupo de manera solitaria, tal y como lo muestra la Figura 25.

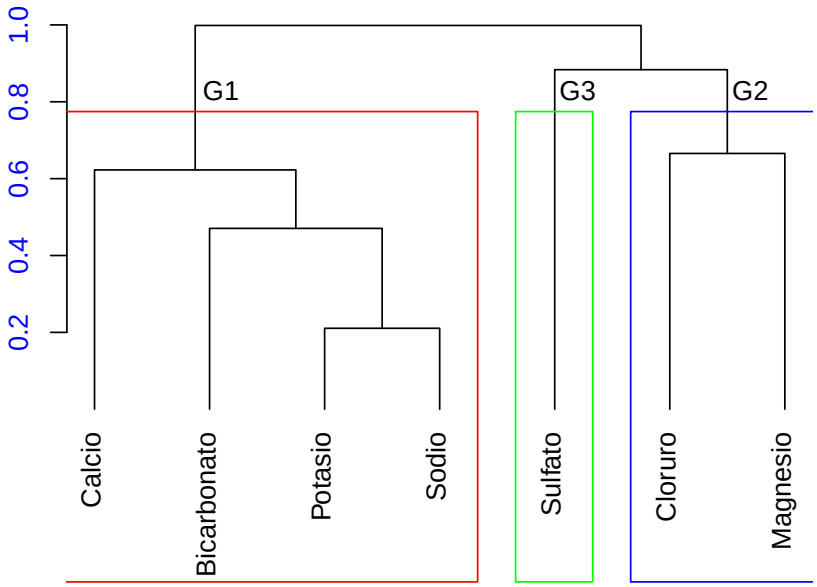

Figura 25: Dendrograma de iones en el Centro del Embalse de Yacyretá

Por último, en la Salida del sistema, se consigue una similar clasificación que en la Entrada. Se forman la misma cantidad de grupos con las mismas variables dentro de los tres grupos, aunque el orden de la aglomeración de los iones en los grupos es algo diferente a lo que se obtuvo en la Entrada del Embalse (Figura 26).

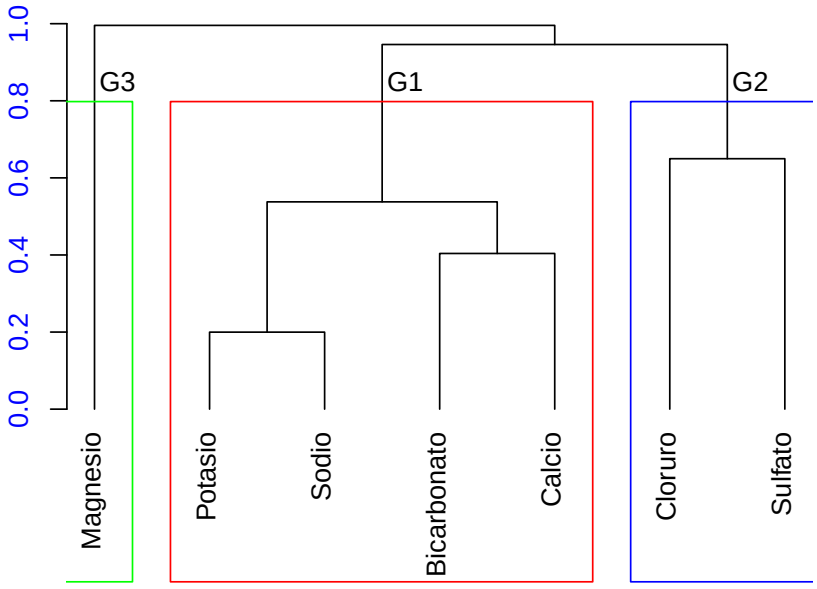

Figura 26: Dendrograma de iones en la Salida del Embalse de Yacyretá

4.2. Análisis Inferencial Multivariado de los datos

4.2.1. Comparación de iones mediante MANOVA Permutacional o no Paramétrico

En la aplicación del MANOVA no paramétrico se han utilizado 2000 permutaciones para la obtención del p-valor asociado al pseudo F. Cabe señalar que se han probado varios valores para la permutación conduciendo también estos a resultados bastante similares en cuanto al p-valor se refiere.

Tabla 13: MANOVA no-paramétrico con 2000 permutaciones para ajustar la relación entre los siete iones y las estaciones de muestreo del Embalse de Yacyretá, durante el periodo 2001-2010

	Grados de libertad	Suma de Cuadrados	Cuadrado Medio	pseudo F	p-valor
Estaciones de Muestreo	2	0,00204	0,0010188	0,46827	0,8331
Residual	341	0,74190	0,0021756		
Total	343	0,74393			

En la Tabla 13 se observan los elementos principales de un análisis de varianza, con la diferencia de que los cálculos de estos elementos son realizados de forma algo diferente al de un MANOVA clásico. En primer lugar, se puede apreciar que el factor de interés (Estaciones de Muestreo) cuenta con dos grados de libertad y el residuo con 341 grados de libertad. Estos dos valores coinciden si se aplica un MANOVA clásico. En segundo lugar, el valor de la suma de cuadrados del factor estudiado es bastante pequeño por lo que este valor indica que existe un cierto grado de similitud entre las tres "Estaciones de Muestreo" estudiadas. El p-valor evidencia que no existen diferencias estadísticamente significativas en las concentraciones medias de los siete iones en las tres Estaciones de Muestreo. Es decir, no hay evidencia suficiente para afirmar que las concentraciones medias de los siete iones varíen espacialmente, por lo menos, en los tres puntos de muestreos considerados en el Embalse de Yacyretá. Esto puede deducirse también de la Figura 27 que sugiere que en cada punto de muestreo existe bastante similitud en cuanto a las concentraciones medias de los siete iones estudiados.

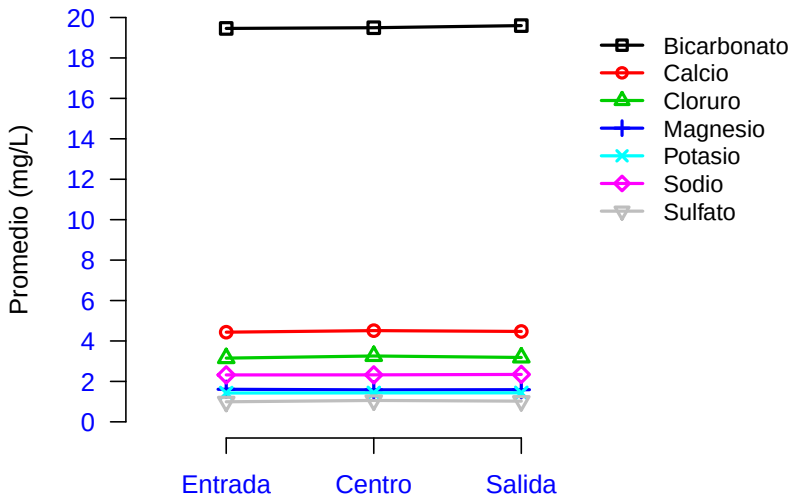

Figura 27: Perfiles de las medias de los siete iones en las tres estaciones de muestreo del Embalse de Yacyretá

En resumen, exploratoriamente, mediante el ACP y el AC, se tiene un indicio de la existencia de cierto grado de homogeneidad en el comportamiento de las concentraciones de los iones analizados, en los tres puntos de muestreo, a lo largo del período analizado. El MANOVA no paramétrico o PERMANOVA corrobora esta primera impresión con un alto grado de confiabilidad a través del p-valor. En otras palabras, con una probabilidad de error alta, la hipótesis de no diferencias significativas en las concentraciones medias de los siete iones, en los tres puntos de muestreo del Embalse de Yacyretá, y a lo largo del período 2001-2010, no es rechazada.

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

En este trabajo se ha mostrado que las técnicas de análisis de datos multivariados como el Análisis de Componentes Principales (ACP), el Análisis de Conglomerados (AC) y el Análisis Multivariado de la Varianza no paramétrico (MANOVA no paramétrico o PERMANOVA) son potentes herramientas para el estudio de las concentraciones de los iones; *Bicarbonato, Cloruro, Sulfato, Calcio, Magnesio, Sodio y Potasio* en las tres estaciones de muestreo del Embalse de Yacyretá, en el periodo 2001-2010.

Los descriptivos básicos univariados muestran similares concentraciones medias de los iones en las tres estaciones de muestreo analizados con variabilidades relativamente homogéneas, por lo menos de estación a estación. Lo mismo sugieren los gráficos descriptivos multivariados, como las estrellas y las caras de Chernoff, que son bastante parecidos en cuanto a las características presentadas de los siete iones en cada Estación de muestreo y a lo largo del período muestreado. Las diferencias del comportamiento global de cada muestra pueden notarse por períodos, esto es en primeros años se destaca preponderantemente la presencia de algún ion (sulfato o cloruro, por ejemplo) pero en los últimos años de muestreo con presencia de todos, en aparente control.

Con el ACP se ha logrado reducir la dimensión del trabajo a costa de una pequeña perdida de la información, es decir, con tres componentes principales se explica por lo menos alrededor del 75 % de la variabilidad total de las siete variables iniciales. Con esta técnica se conformaron iguales grupos de variables en cada una de las tres estaciones de muestreo: la primera componente principal agrupa las variables Bicarbonato, Calcio, Sodio y Potasio por lo que esta componente está asociada al Bicarbonato del agua y los cationes que aportan al mismo, aumento de las concentraciones de estos cationes involucran un aumento en el Bicarbonato; la segunda componente explica mejor los aniones Sulfato y Cloruro

con correlaciones opuestas con esta componente principal; mientras que la tercera componente principal determina mayoritariamente las concentraciones del cation Magnesio.

El AC también sugieren tres grupos o conglomerados de variables que coinciden con los formados por el ACP, excepto en el "Centro del Embalse" en donde difieren dos de los tres grupos de variables. Es decir, en el Centro este método agrupa el cation Magnesio con el anion Cloruro, y por otro lado, deja solo al anion Sulfato en otro grupo.

Con estas dos últimas técnicas exploratorias multivariadas se ha logrado disminuir la dimensión del problema y, además, se comprueba la existencia de tres grupos de variables bastante similares en cada una de las Estaciones de muestreo. Uno de esos grupos indudablemente está bien definida en las tres Estaciones de muestreo, el conformado por los cuatro iones; Bicarbonato, Calcio, Sodio y Potasio .

Finalmente, con la técnica inferencial del MANOVA no paramétrico o simplemente PERMANOVA se ha podido constatar, mediante una medida de confiabilidad, lo que indicaban los procedimientos descriptivos univariados y multivariados, como así también el ACP y AC, esto es, no existen diferencias respecto de la concentración de los iones en las tres Estaciones muestreadas a lo largo del Embalse Yacyretá en el periodo estudiado. Es decir, estadísticamente existe mucha homogeneidad entre las tres Estaciones de muestreo en cuanto a las concentraciones de los siete iones. Esto indica que no hay un efecto espacial significativo en las concentraciones medias de los iones estudiados (p-valor=0,8331).

5.2. Recomendaciones

El presente trabajo puede ser ampliado y mejorado incluyendo mayor cantidad de variables fisicoquímicas que, en cierta forma, contribuyan a la explicación de la calidad del agua, específicamente la salinidad, como el pH o la conductividad, entre otras.

Resulta, a su vez, sumamente importante, utilizar otras técnicas del Análisis Multivariado con la finalidad de desarrollar índices de salinidad, tanto para todo el Embalse como para los distintos puntos de muestreos, que permitan corroborar lo aquí afirmado sobre el comportamiento homogéneo de todo el sistema, a través

del tiempo. El análisis factorial y el PERMANOVA multifactorial podrían ser herramientas multivariadas útiles para lograr este fin.

Un índice general de calidad del agua puede formularse además con la mayoría de las variables fisicoquímicas medidas en las aguas del Embalse de Yacyretá con los métodos factoriales multivariados que son de amplio uso en esta área y, a partir de estos, verificar las diferencias o similitudes existentes en puntos de muestreo más desagregados y en periodos de tiempo más amplios y más actuales.

Otras técnicas para el análisis de datos multivariados, tradicionales (como el Escalonamiento Multivariado) o emergentes (como las asociadas a Minería de Datos), pueden ser utilizadas para el estudio de las concentraciones iónicas que definen,

ANEXO A

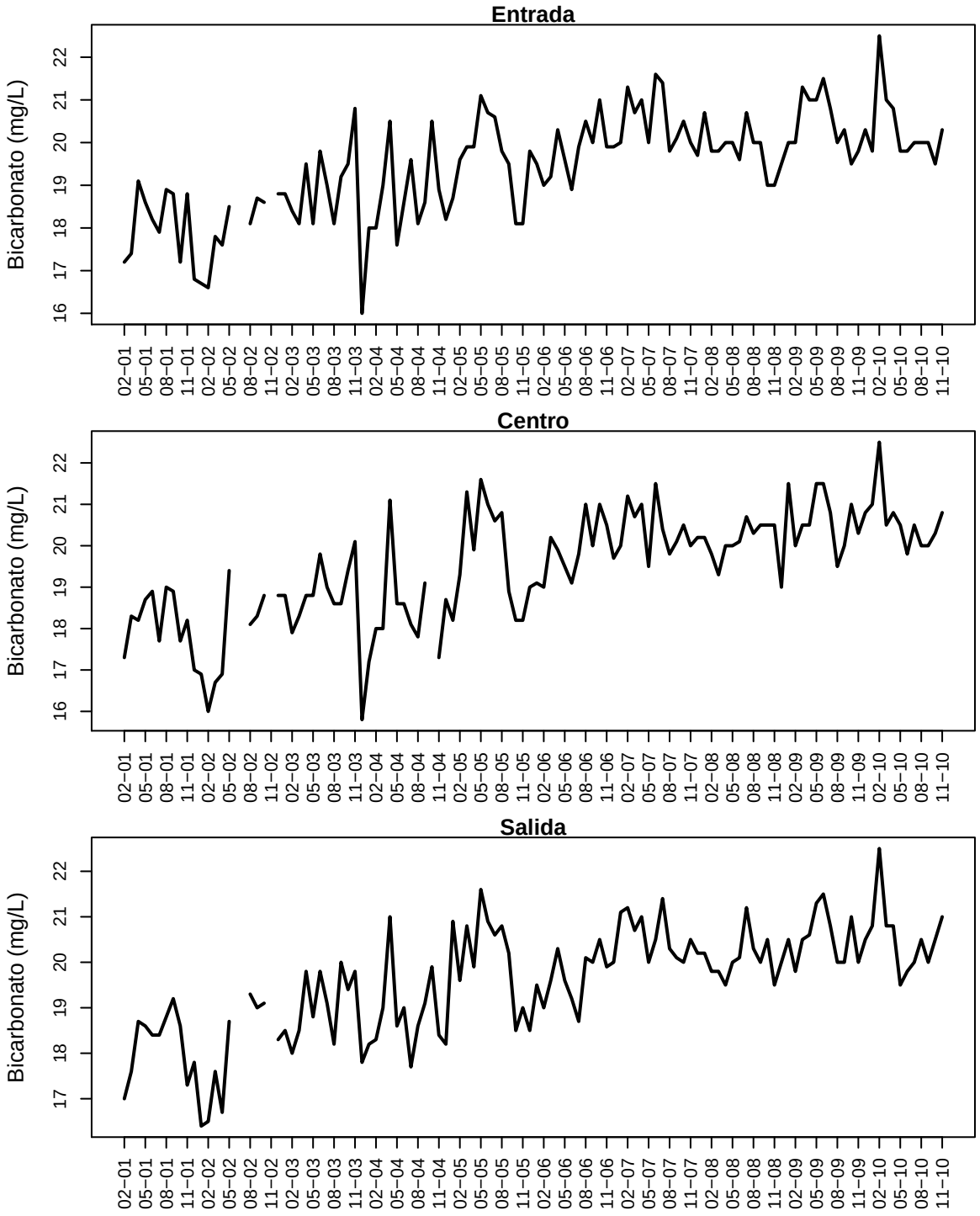

Figura 28: Concentraciones del ion Bicarbonato en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

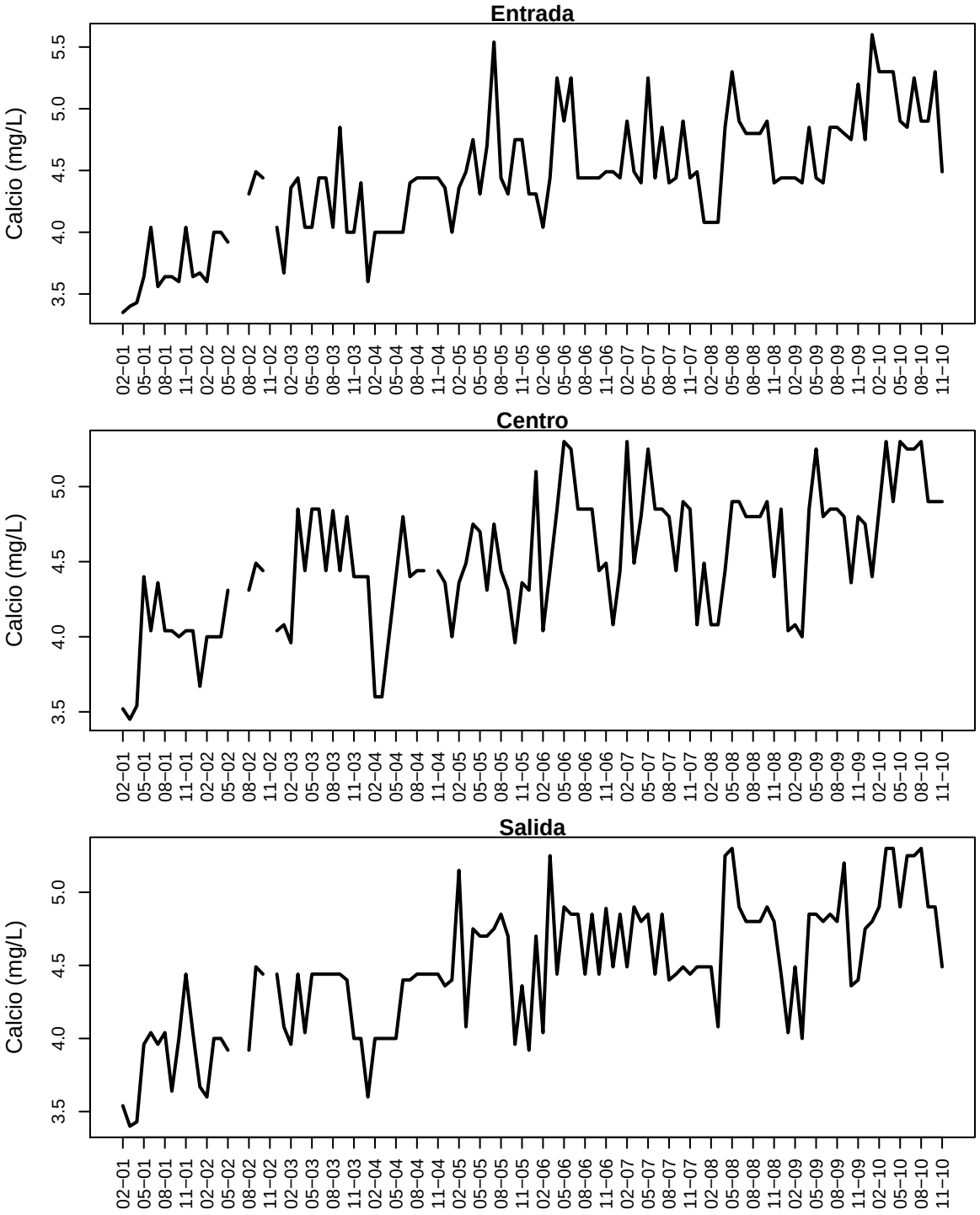

Figura 29: Concentraciones del ion Calcio en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

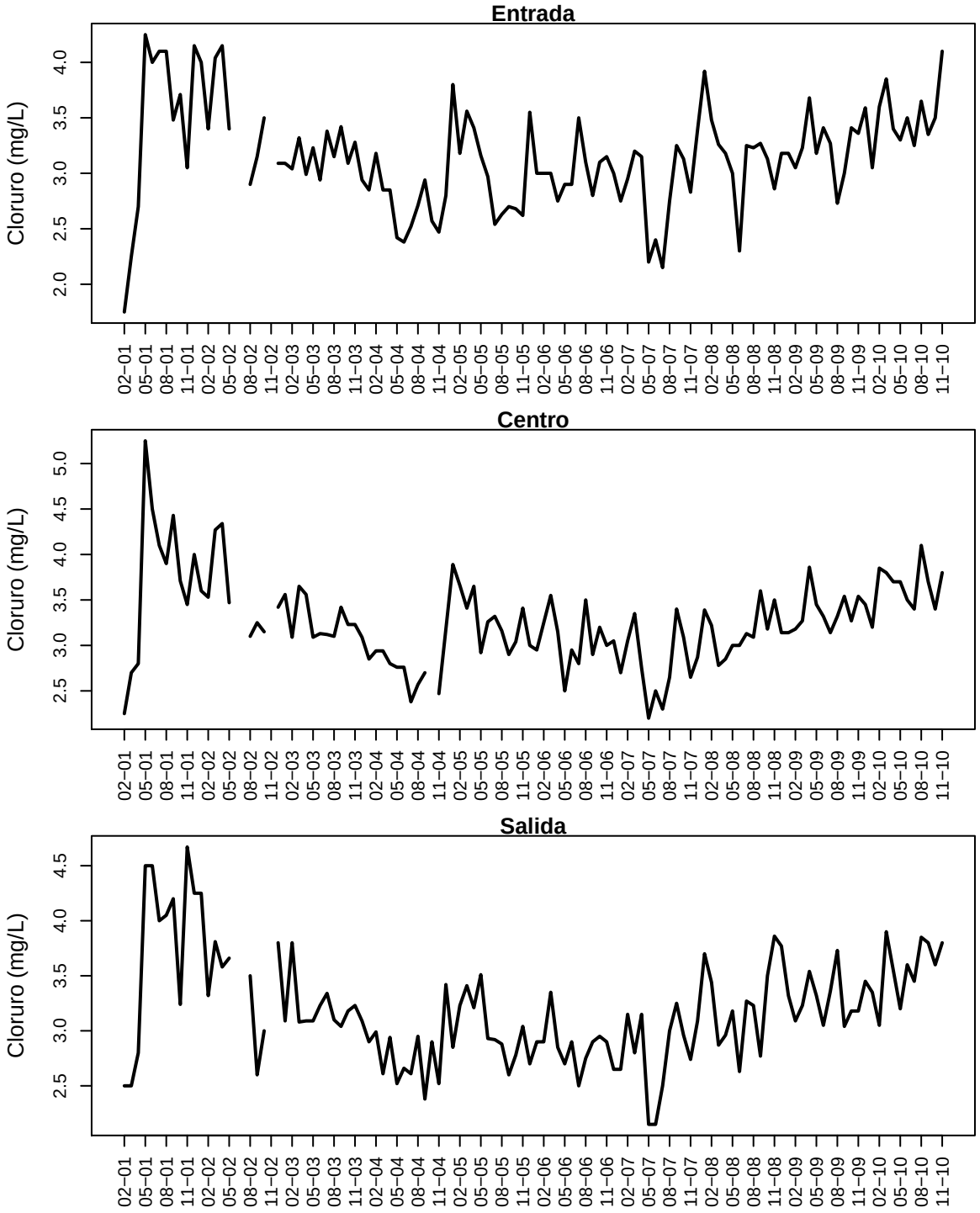

Figura 30: Concentraciones del ion Cloruro en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

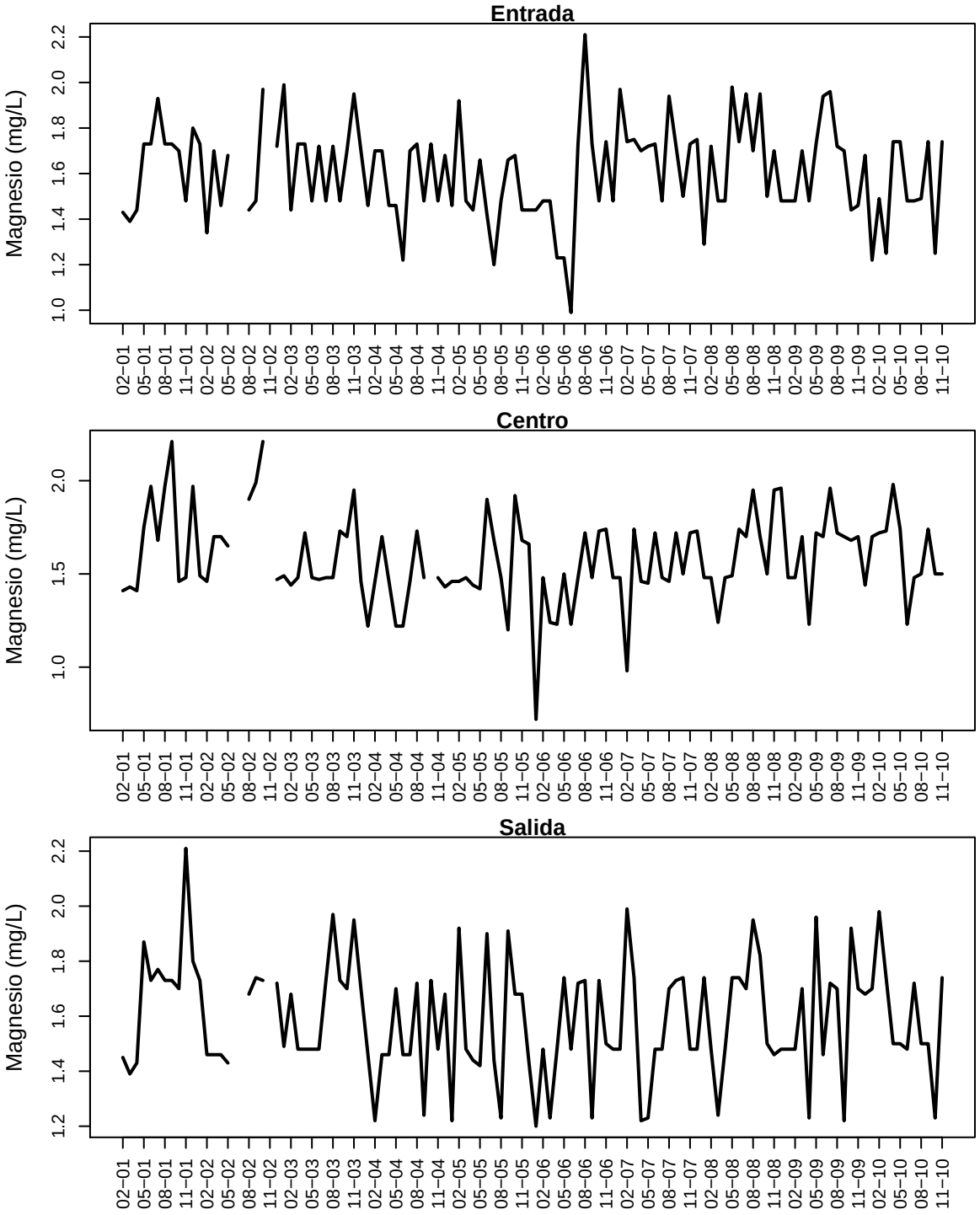

Figura 31: Concentraciones del ion Magnesio en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

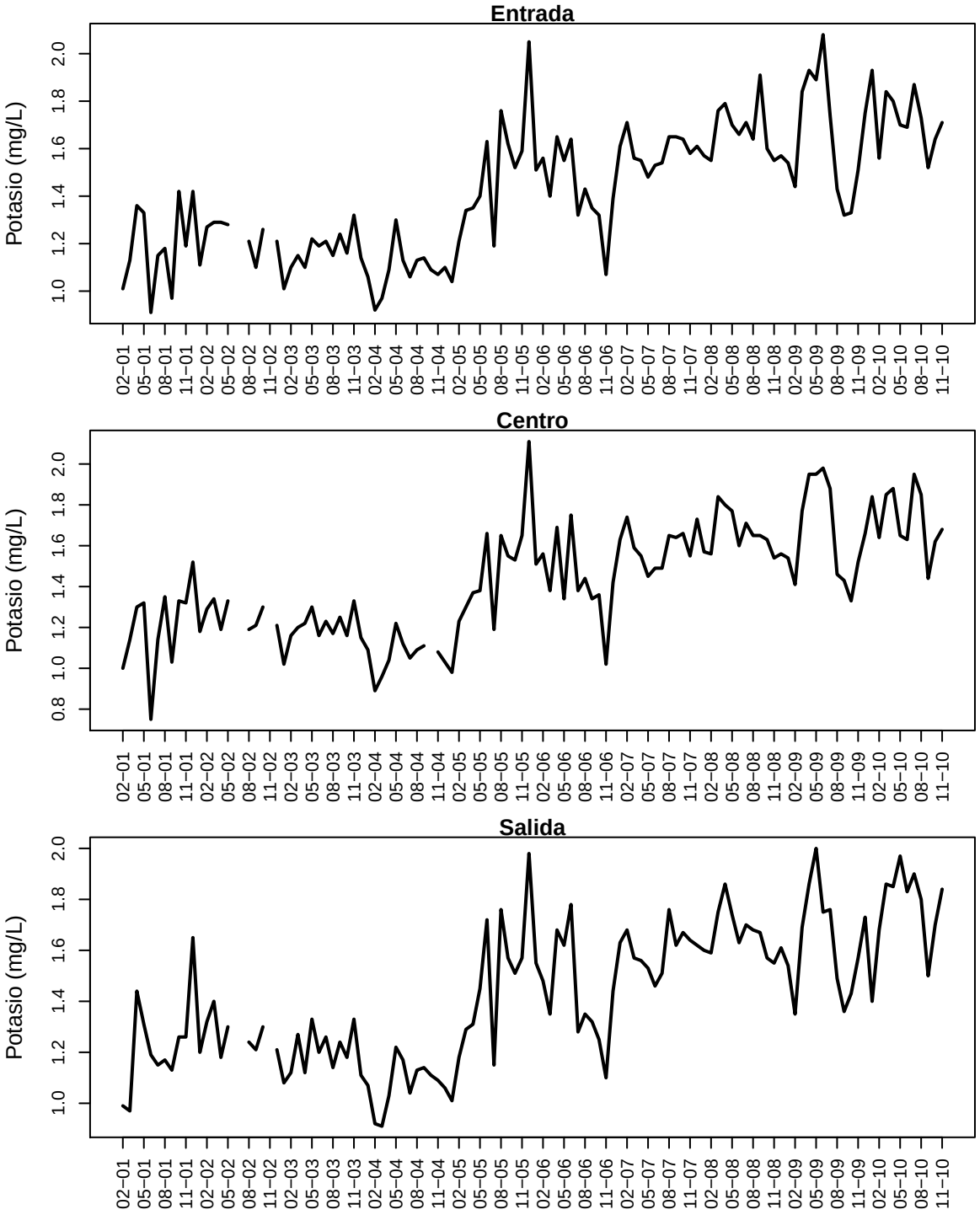

Figura 32: Concentraciones del ion Potasio en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

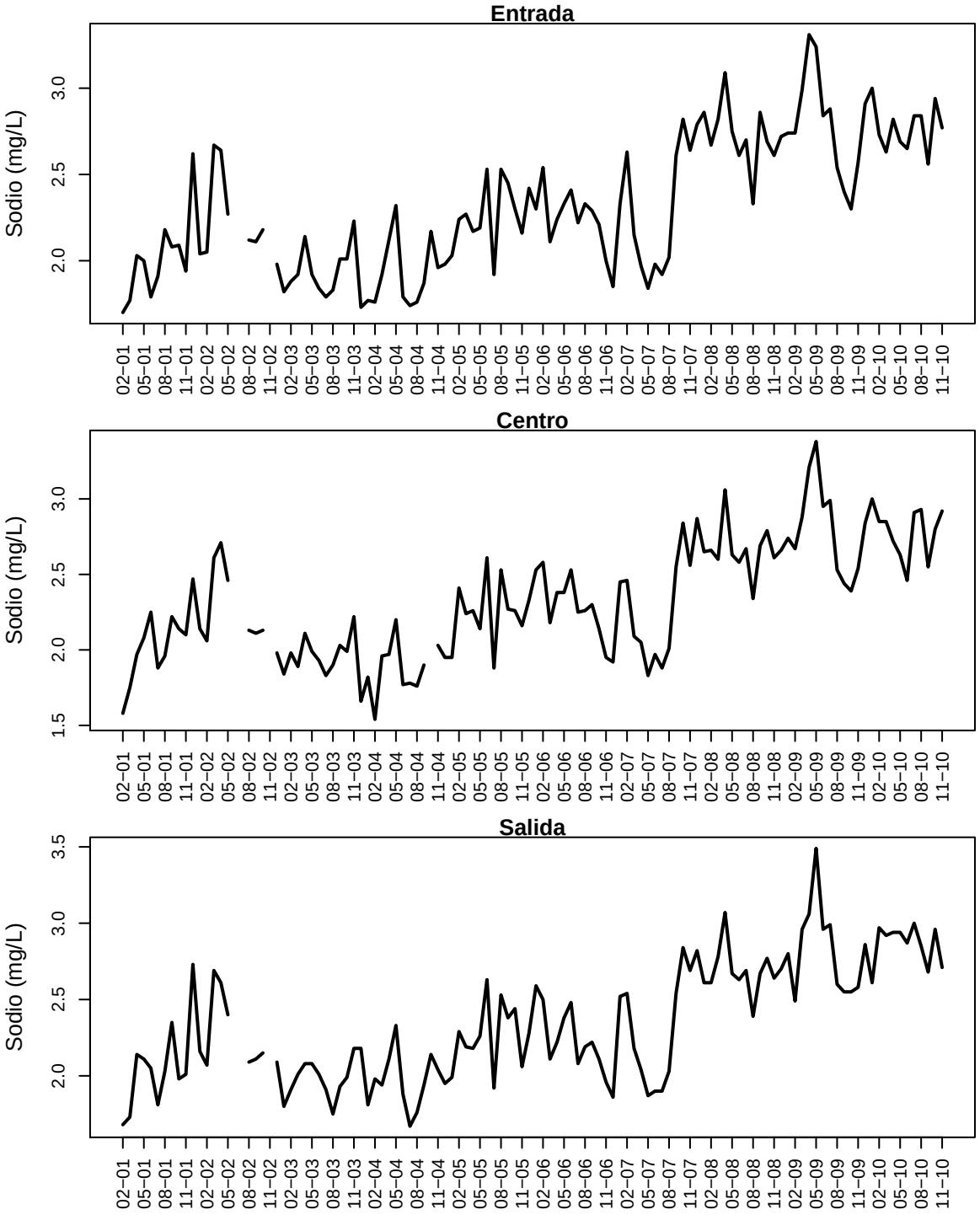

Figura 33: Concentraciones del ion Sodio en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

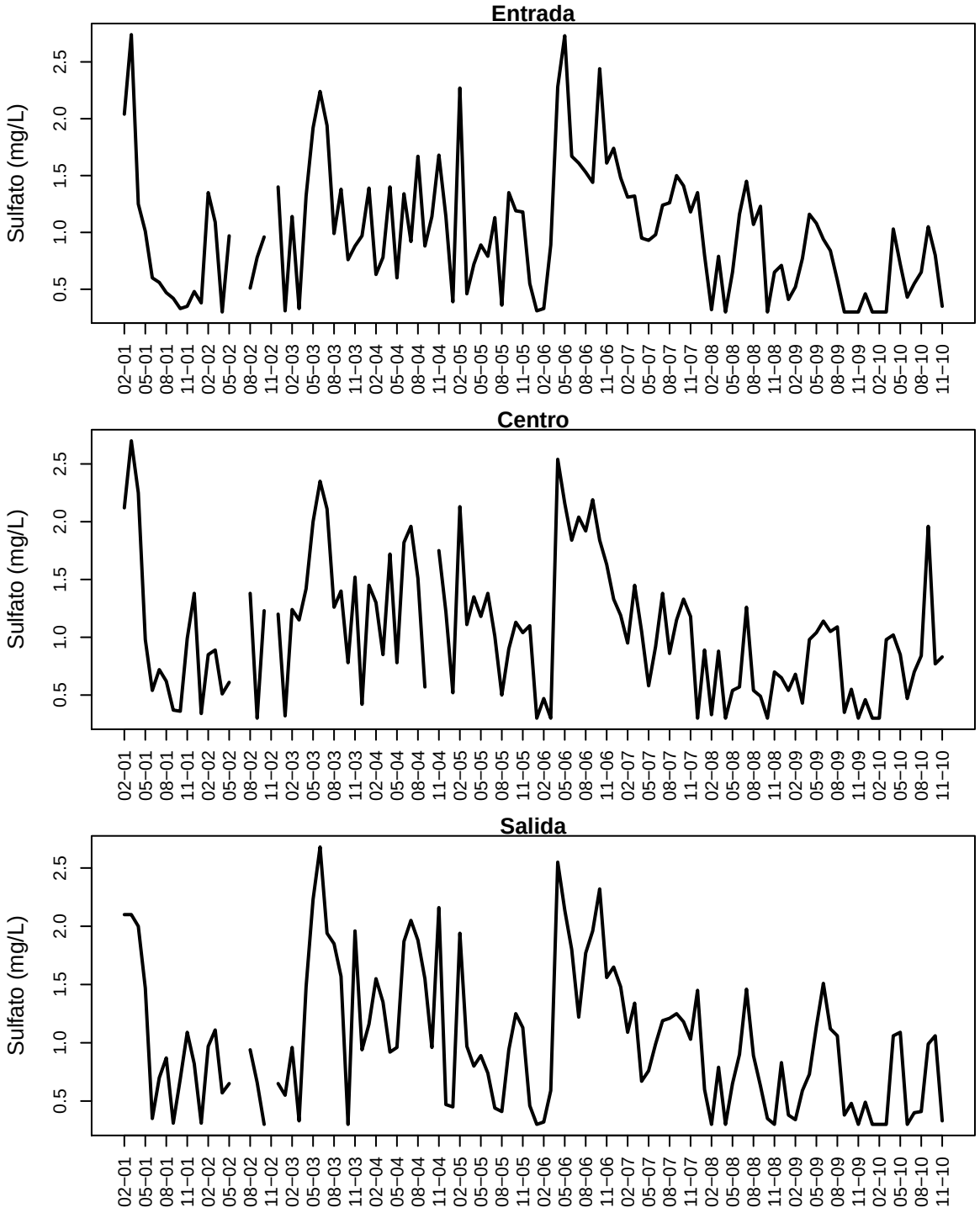

Figura 34: Concentraciones del ion Sulfato en los tres puntos de muestreo del Embalse de Yacyretá, durante el período febrero/2001 a noviembre/2010.

REFERENCIAS BIBLIOGRÁFICAS

ANDERSON, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46.

ANDERSON, M. (2014). Permutational multivariate analysis of variance (PERMANOVA). *Chichester: John Wiley and Sons, Ltd.*

ÁVILA, H.; GARCIA, S.; ROSA-ACEVEDO, J. (2015). Análisis de Componente Principales, como herramienta para interrelaciones entre variables fisicoquímicas y biológicas en un ecosistema léntico de Guerrero, México. *Revista Iberoamericana de Ciencias*, 2(3):43–53.

CASTRO, L.; CARVAJAL, Y.; ÁVILA, Á. (2012). Análisis clúster como técnica de análisis exploratorio de registros múltiples en datos meteorológicos. *Ingeniería de Recursos Naturales y del Ambiente*, (11):11–20.

CAYUELA, L. (2011). Análisis Multivariante. (en línea). Consultado 21 de junio 2017. Disponible en https://dl.orangedox.com/dyCZ78Z5w8pctsQ8O3/6-Analisis20multivariante.pdf.

CHENINI, I.; KHEMIRI, S. (2009). Evaluation of ground water quality using multiple linear regression and structural equation modeling. *International Journal of Environmental Science and Technology*, 6(3):509–519.

COLETTI, C.; TESTEZLAF, R.; RIBEIRO, T.; SOUZA, R.; PEREIRA, D. (2010). Water quality index using multivariate factorial analysis. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 14(5):517–522.

COMAS, E.; ARMENGOL, J.; SABATER, S.; SABATER, F. (1998). Variabilidad espacial y temporal de la calidad del agua en el río Urola (Guipuzkoa). *Ingeniería del agua*, 5(4):29–36.

CUADRAS, C. (2014). Nuevos Métodos de Análisis Multivariante. Barcelona, ES: CMC Editions. 304 p.

DE LA GARZA, J.; MORALES, B.; GONZÁLEZ, B. (2013). Análisis Estadístico Multivariante: Un enfoque teórico y práctico. 1ra ed. México, MX: Mc Graw Hill. 728p.

DÍAZ, A. (2013). Aplicación de Modelos de Series de Tiempo a un componente iónico (Alcalinidad Total) de la calidad del agua del Embalse de Yacyretá. Tesis (M. Sc.). San Lorenzo, PY: FACEN-UNA. 86 p.

EBY (ENTIDAD BINACIONAL YACYRETÁ). (2013). Historia de la EBY (en línea). Consultado 21 de junio 2017. Disponible en https://www.eby.gov.py/index.php/institucional/historia.

GÓMEZ, I.; PEÑUELA, G. (2016). Revisión de los métodos estadísticos multivariados usados en el análisis de calidad de aguas. *Revista Mutis*, 6(1):54–63.

HAIR, J.; ANDERSON, R.; TATHAM, R.; BLACK, W. (1999). Análisis multivariante. 5ta Edición. Madrid, ES: Prentice Hall. 832 p.

JOHNSON, R.; WICHERN, D. (1982). Applied Multivariate Statistical Analysis. 3ra Edición. Editorial Prentice Hall. 642 p.

KAUFMAN, L.; ROUSSEEUW, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

LÓPEZ, M.; PALACÍ, D. (2014). Estudio multivariante de la calidad del agua: Aplicación al río Júcar en el periodo 1990-2013. *M+ A. Revista Electrónica de Medio Ambiente*, 15(1):37–52.

OMS (ORGANIZACIÓN MUNDIAL DE LA SALUD). (2006). Guías para la calidad del agua potable. Primer apéndice a la tercera edición. Ginebra. 408 p.

PEÑA, D. (2002). Análisis de Datos Multivariantes. Mc Graw Hill/Interamericana de España, S.A.U. 539 p.

PÉREZ, C. (2004). Técnicas de Análisis Multivariante de Datos: Aplicaciones con SPSS. Madrid, ES: Pearson Educación. 672 p.

PÉREZ, C. (2015). R Lenguaje de programación y análisis de datos estadísticos. Madrid, ES: Garceta. 450 p.

R CORE TEAM. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RODRÍGUEZ, M. I.; GRANERO M.; BUSTAMANTE, M. A.; AVENA, M.; BONFANTI, E.; BUSSO, F.; GIRBAL, A. (2001). Composición iónica del embalse San Roque (Córdoba, Argentina) y su relación con el proceso de eutroficación. Seminario Internacional de Gestión Ambiental e Hidroelectricidad. INA. Complejo Hidroeléctrico Salto Grande, Entre Ríos, Argentina. 9 p.

ROJAS, H. (2010). Calidad del agua del Embalse de Yacyretá en la cota de 76 metros sobre el nivel del mar. *Reportes Científicos de la FaCEN*, 1(1):40–55.

SIERRA, C. (2011). Calidad del agua, evaluación y diagnóstico. Medellin, CO: Ediciones de la U. 458 p.

URIEL, E. (1995). Análisis de datos: Series temporales y Análisis multivariante. Madrid, ES: Editorial AC. 436 p.

VALENCIA, J. (2007). Estudio Estadístico de la Calidad de las Aguas en la Cuenca Hidrográfica del Río Ebro. Tesis Doctoral. Universidad Politécnica de Madrid. Madrid, España. 338 p. Consultado 26 de setiembre de 2017. Disponible en http://oa.upm.es/454/1/JOSE_LUIS_VALENCIA_DELFA.pdf.