UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE QUIMICA

TITULO:

"EVALUACION DE LA CALIDAD DEL AGUA DEL RIO RIMAC MEDIANTE EL ANALISIS MULTIVARIADO"

TESIS PARA OPTAR EL TITULO PROFESIONAL DE:

LICENCIADO EN QUIMICA

CESAR GREGORIO ESPIRITU LIMAY

ASESOR:

LIC. CHRISTIAN JACINTO HERNANDEZ

LIMA-PERÚ

2010

RESUMEN

El alcance de este trabajo de tesis es la aplicación de las diferentes técnicas estadísticas multivariadas para la evaluación espacial e interpretación de una gran cantidad de datos obtenidos durante el monitoreo del agua de la cuenca seca, parte de la cuenca húmeda y el principal tributario del río Rímac, el río Santa Eulalia, localizados en el departamento de Lima.

Un total de 252 muestras de agua fueron colectadas mensualmente durante el período Julio 2008 a Junio 2009 en 7 diferentes estaciones de muestreo a lo largo de la trayectoria del río y el principal tributario, analizando 20 parámetros ambientales físicos y químicos en cada estación de monitoreo. Los análisis se realizaron en los laboratorios de Química de la Facultad de Ciencias de la Universidad Nacional de Ingeniería.

El objetivo principal de esta tesis se enfocó en la evaluación de los parámetros ambientales analizados en el monitoreo ambiental y la clasificación espacial (estaciones de muestreo) del agua del rio Rímac mediante la aplicación de las técnicas multivariadas más importantes, específicamente, análisis de clúster (AC), análisis de componentes principales (ACP), análisis de factor (AF) y análisis discriminante (AD). Estas técnicas fueron analizadas con el programa estadístico Statistica versión 8.0 (StaSoft 2007).

En primer lugar se realizó el análisis inicial y pretratamiento de los datos antes de la realización de estos análisis multivariados para evaluar el comportamiento inicial de los datos y transfórmalos si es requerido en nuevos datos ambientales que cumplan los requerimientos de la normalidad multivariada.

Seguidamente se desarrollo los AC y ACP, y cuyos resultados permitieron visualizar la agrupación de las estaciones de muestreo en una sola clase, a saber, grupo I: estación Los Ángeles, Huampaní y Huachipa. Mientras que las demás estaciones no se encuentran agrupadas, a saber, Ricardo Palma, Santa Eulalia, Del Ejército, y Gambeta. Además estos análisis multivariado agruparon a los parámetros ambientales en 4 clases de acuerdo a las características comunes e interacciones que se encuentran en el río.

Por otro lado, el AF determinó las relaciones que existen entre los 20 parámetros ambientales dadas por la matriz de correlación de estos parámetros, lo cual permitió identificar y caracterizar 4 principales fuentes de contaminación que influyen en la composición fisicoquímica del agua del río Rímac, a saber, urbana, industrial, agrícola y geológica, el cual son caracterizados de acuerdo a todos los parámetros ambientales analizados. Además, este análisis evaluó las influencias de estas fuentes (puntuaciones de los factores) sobre cada una de las 7 estaciones de muestreo lo largo de todo el período de muestreo.

Finalmente el AD comprobó la clasificación realizada por el AC y ACP con una confiabilidad del 94% de acuerdo a las características fisicoquímicas de las estaciones de muestreo. Además el AD determinó los parámetros ambientales (NO_3-N, NO_2-N y PO_4 -P, Cl, DT y Fe) más discriminantes para la clasificación realizada sobre las estaciones de muestreo por el AC y ACP con una confiabilidad del 82.1 % de la clasificación.

De esta manera, de acuerdo a los resultados obtenidos, la presente tesis pretende ser una contribución inicial para una evaluación más exhaustiva del estado global del agua del río Rímac de una manera rápida y sin pérdida de información original, de manera que exista una alternativa de evaluación en el control e identificación de las fuentes de contaminación y así planear un adecuado manejo y protección del área de estudio.

INTRODUCCION

La cuenca de un río generalmente constituye áreas con una alta densidad demográfica debido a las condiciones de vida favorables tales como la disponibilidad de tierras fértiles, agua potable, riego de plantas, bebida para animales y para propósitos industriales, además de ser medios eficientes para el transporte fluvial.

En el Perú, una de las cuencas hidrográficas más importante es la cuenca del río Rímac pues cruza las ciudades más importantes y altamente pobladas del Perú, a saber, Lima y Callao. Esta cuenca desempeña un rol vital como fuente de abastecimiento de agua para el consumo humano, agrícola y energético. Sin embargo, pese a su importancia económica, y los servicios ambientales que brinda, la cuenca del río Rímac es una de las más contaminadas a nivel nacional, y enfrenta serios problemas de degradación ambiental producto de la actividad antropogénica; los cuales generan graves impactos a la naturaleza, la vida y la salud de las poblaciones.

Para establecer una norma moderna que pueda evaluar la calidad del agua, el 31 de Julio del 2008 se publicó en el diario oficial El Peruano los estándares nacionales de calidad ambiental para agua (ECA) de los diferentes parámetros ambientales (físicos, químicos y biológicos) en los diferentes ambientes de agua natural y para usos o actividades a la que se destina el agua, cambiando de esta manera la antigua ley general de aguas promulgada en 1969 y modificada en 1983.

Estos estándares nacionales son establecidos con el objetivo de conservar los ambientes acuáticos y tener un rango de calidad del agua para sus diferentes usos, tales como: producción de agua potable, recreacional, riego de vegetales, actividades marinas y bebidas para animales.

Sin embargo, para cumplir con el objetivo de los ECA es necesario realizar una evaluación constante de la calidad de cada ambiente acuático y los factores que contribuyen. Esta evaluación tiene ciertos requisitos indispensables tales como: un programa de monitoreo continuo de todos los parámetros del agua, un conocimiento de la naturaleza de estos parámetros y sus interacciones con su entorno, y por último

desarrollar técnicas de evaluación estadística adecuados que permita obtener una buena interpretación de los resultados concernientes a la realidad de cada ambiente acuático.

El éxito del programa de monitoreo sobre cada ambiente acuático conlleva a tener que analizar una gran cantidad de parámetros, tomados en diferentes fechas y provenientes de muchas estaciones de muestreo. El resultado es una matriz compleja de datos, el cual es necesario interpretar para evaluar el comportamiento fisicoquímico del agua. Por esta razón, el desarrollo de las técnicas estadísticas adecuadas es fundamental para obtener resultados significativos, especialmente cuando los grupos de datos son químicamente irregulares y extensamente evaluados.

De esta manera, la rama de la química y la estadística en conjunto, llamada quimiometria proporciona una gran extensión de técnicas estadísticas, los cuales están incrementándose en su uso para estudios ambientales relacionados no solo con las evaluaciones de los ambientes acuáticos si no también con los ambientes terrestres y la atmósfera.

De todas las técnicas estadísticas, las más importantes para una evaluación inicial sobre una gran base de datos provenientes del monitoreo ambiental son las técnicas exploratorias de datos multivariados o reconocimiento de patrones, a saber, análisis de clúster (AC), análisis de componentes principales (ACP), análisis de factores (AF), como técnicas de reconocimientos de patrones no supervisadas, y el análisis de discriminantes (AD) como técnica de reconocimiento de patrones supervisada.

Estas técnicas de la quimiometría proveen múltiples vías de análisis para visualizar y caracterizar los parámetros fisicoquímicos, las estaciones de muestreo, y es útil para evidenciar las variaciones espaciales y temporales causadas por las influencias de la naturaleza y las actividades antropogénicas en cada estación de muestreo

Utilizando estas técnicas multivariadas sobre la matriz compleja de datos del programa de monitoreo ambiental del agua del rio Rímac se tienen que los principales objetivos de la tesis son:

(1) Realizar la agrupación de las estaciones de muestreo de acuerdo a la composición fisicoquímica de sus aguas, y determinar las relaciones entre los parámetros evaluados.

(2) Designar las fuentes de contaminación provenientes de las actividades antropogénicas y la geología de acuerdo a la composición fisicoquímica del agua del rio Rímac, y las influencias sobre cada estación de muestreo.

(3) Evaluar la clasificación de las estaciones de muestreo y la reducción del número de parámetros para la evaluación de la calidad del agua.

SIGLAS Y ABREVIATURAS

EAA	Espectral Atomic Absorption
ACP	Análisis de componentes principales
AD	Análisis discriminante
AF	Análisis de factor
APHA	American Public Health Association
AWWA	American Water Works Association
CP	Componente principal
DBO_5	Demandas Bioquímica de Oxigeno al quinto día
DQO	Demanda Química de Oxigeno
EPA	Environmental Protection Agency
FD	Funciones discriminantes
gl	Grados de libertad
HDPE	High Density Poly Etilene
kNN	k Nearest Neighbor
LDD	Límite de detección
MANOVA	Multiple analysis of variance
msnm	Metros sobre el nivel del mar
NE	Noreste
SE	Sureste
NW	Noroeste
SW	Suroeste
NIST	National Institute of Stantards and Technology
SIMCA	Soft Independent Models of Class Analogy
SRM	Standard Reference Material
PM	Particular Matter
PLS	Partial Least Squares
UNEQ	Unequal Dispersed classes
UTM	Unidad Transversal de mercartor
WPCF	Water Pollution Control Federation

INDICE

INDICE DE FIGURAS

PRIMERA PARTE: MARCO TEORICO

CAPITULO 1: EL AGUA SUPERFICIAL

CAPITULO 2: QUIMIOMETRIA EN EL ANALISIS AMBIENTAL

SEGUNDA PARTE: METODOLOGIA EXPERIMENTAL

CAPITULO 3: PROCEDIMIENTO EXPERIMENTAL

TERCERA PARTE: RESULTADOS

CAPITULO 4: DISCUSION DE RESULTADOS

ANEXOS

Anexo 1

Anexo 4

INDICE DE TABLAS

PRIMERA PARTE: MARCO TEORICO

CAPITULO 1: EL AGUA SUPERFICIAL

CAPITULO 2: QUIMIOMETRIA EN EL ANALISIS AMBIENTAL

SEGUNDA PARTE: METODOLOGIA EXPERIMENTAL

CAPITULO 3: PROCEDIMIENTO EXPERIMENTAL

TERCERA PARTE: RESULTADOS

CAPITULO 4: DISCUSION DE RESULTADOS

ANEXOS

Anexo 5

Anexo 6

AGUA SUPERFICIAL

1.1 introducción

El agua superficial es toda agua natural que se encuentra en contacto con la atmósfera e incluye a los ríos, lagos, océanos, mares, estuarios, humedales, etc. Esta representa uno de los elementos básicos que sostiene la vida y los ecosistemas, un componente primario para la industria, una parte de consumo para los humanos y animales y un vector para los efluentes de las actividades domésticas e industriales (1; 2, 3).

Las características fisicoquímicas y biológicas del agua superficial dependen de los múltiples factores naturales tales como la intensidad y composición de la lluvia, las reacciones químicas entre el agua y los sedimentos, las reacciones biogeoquímicas, y las interacciones entre el agua superficial y subterránea.

Sin embargo, el desarrollo tecnológico, el crecimiento demográfico, la industrialización, las nuevas técnicas agrícolas son factores que generan la contaminación. Este contribuye a la introducción de un gran número de sustancias químicas en el agua superficial y cuyas interacciones y efectos adversos en los ecosistemas acuáticos y los seres vivos son procesos complejos que requieren de un estudio profundo (5).

El principal problema de la contaminación es el rápido aumento de las emisiones de los contaminantes en los ecosistemas acuáticos, excediendo de esa manera la capacidad de estos de asimilarlos. Para cualquier actividad general en donde se genera la contaminación hay ciertas características comunes, tales como: las fuentes de contaminación, los contaminantes, el medio de transporte, y el receptor (6).

Así, la consideración más importante para la gestión del agua superficial es la creación de abastecimientos adecuados, limpios y seguros; tanto para las diferentes actividades antropogénicas como para la sostenibilidad de los diversos ecosistemas acuáticos (1; 4).

Por esto, surge el interés de estudiar y preservar los ciclos y procesos naturales de las aguas superficiales por medio de programas de monitoreo de las aguas superficiales, los cuales son una aproximación adecuada para un mejor entendimiento de la composición fisicoquímica y de las diferentes fuentes de contaminación en el agua superficial.

1.2 Distribución

El agua de la hidrosfera se transfiere de manera continua desde los océanos a las regiones de la tierra y de nuevo a los océanos dentro de un proceso cíclico denominado ciclo del agua, el cual se mueve por la energía del sol y es considerado como un gran sistema de depuración natural de la misma (2). Una imagen desde el espacio muestra que vivimos en un mundo de agua, ya que más del 70% de la superficie terrestre está ocupada por la hidrosfera y de cuya existencia depende la vida. La Figura N°1-1 detalla que el volumen total del agua en la tierra es de 1 400 millones de km^3 de los cuales más del 97% del agua de la tierra es agua salada y que se localiza en los mares y océanos, y que cubren alrededor de las tres cuartas partes de la superficie terrestre.

Por otro lado, el volumen total de agua dulce es de 35 millones de Km^3 ó alrededor de 2.5 % del volumen total. Así, el 68.9% está en forma de hielo permanente que cubre las regiones montañosas y se encuentran en las regiones de la Antártica y el Ártico. Asimismo, un 30.8% conforman las aguas subterráneas, 0.3% se encuentra en los lagos, y solamente el 0.001% del agua de la tierra se localiza en la atmosfera (2).

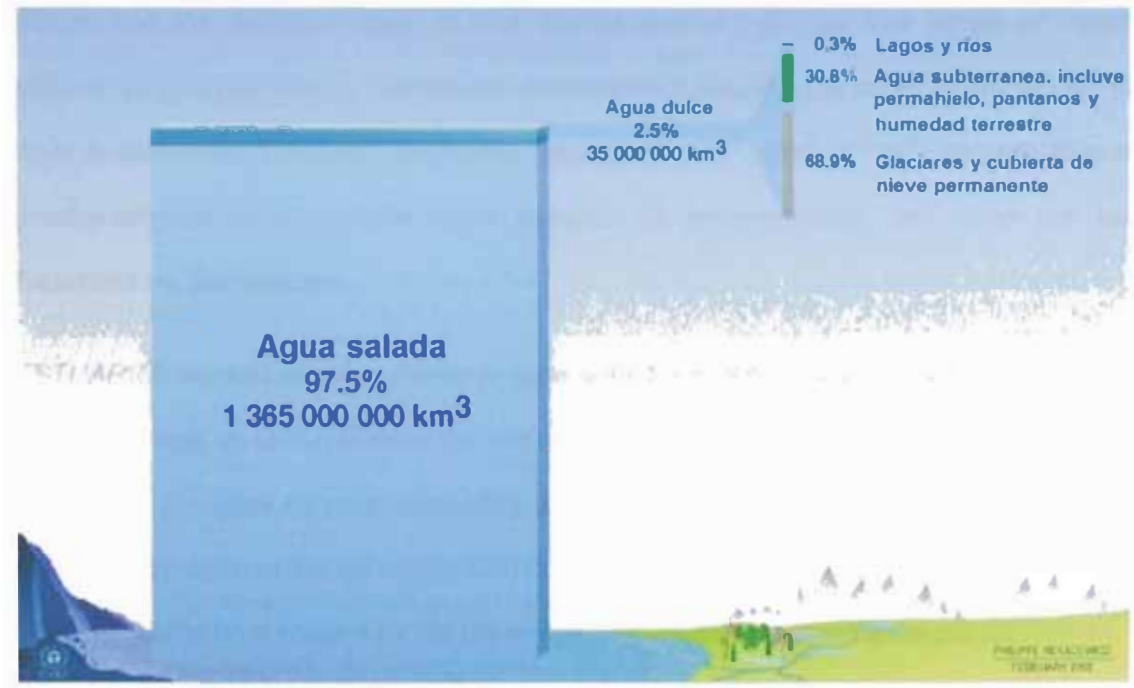

Figura N° 1-1. Distribución porcentual del agua en los diversos ambientes acuáticos de la Tierra[1].

[1] Fuente: Iger A. Shiklomanov, State Hydrological Institute (SHI, St. Petersburg) and United Nations Educational, Scientific and Cultural Organisation (UNESCO, París), 1999

1.3 Clases de agua superficial

Las muestras de aguas superficiales se clasifican de acuerdo a la composición fisicoquímica y las interacciones que tiene con el ambiente. De esta manera, tenemos la siguiente clasificación general:

RIO: ecosistema de agua dulce que comprende tanto el curso principal y los afluentes. Las aguas de río juegan un rol importante en la asimilación y transporte de las aguas residuales y municipales y las escorrentías de las tierras agrícolas, llevando el flujo de una carga significante de materia en disolución y fases de partículas de un solo sentido. Las aguas de río tienen composiciones variables y el estado ambiental en cualquier punto refleja algunas influencias importantes, incluyendo la litología de la cuenca, los aportes atmosféricos, las condiciones de clima y los aportes antropogénicos (7).

OCEANO Y MAR: hábitats de una gran cantidad de vida animal y vegetal y una gigantesca solución de iones y otras sustancias en las que existen plantas y animales. Gran parte de los constituyentes disueltos del agua de los mares y océanos son iones y sufren solo pequeñas variaciones en sus cantidades relativas (2). Los iones sodio y cloruro son los predominantes, lo que explica que el agua de mar tenga un sabor salobre. Hay muchos otros elementos, denominados trazas, que están presentes en el agua en concentraciones muy pequeñas, tales como: Ca^{2+}, Mg^{2+}, F^{-}, Sr^{2+}, Br^{-}, etc. El mar geológicamente es importante como espacio de sedimentación, así como por las fluctuaciones que ocurren.

ESTUARIO: espacio acuático donde el agua salada del océano y el agua dulce del río se mezclan y este es caracterizado por altos gradientes en la fuerza iónica y composición química (1; 6). Este es muy vulnerable a la contaminación debido a la deposición de partículas de sedimentos de origen terrestre con los contaminantes absorbidos en ellos y actuando como un sumidero de los diferentes contaminantes arrastrados por los ríos (6).

LAGO (8): cuerpo de agua dulce parcialmente encerrado y circundado de tierra que se encuentra alejada del mar, y asociada generalmente a un origen glaciar. El

comportamiento del agua del lago está sujeto a una amplia variedad de influencias y por el aporte de las aguas de los ríos y del afloramiento de aguas freáticas.

HUMEDAL: zona de tierras, generalmente planas, en la que la superficie se inunda permanente o intermitentemente. Al cubrirse regularmente esta zona de agua, el suelo se satura, quedando desprovisto de oxígeno y dando lugar a un ecosistema híbrido entre los puramente acuáticos y los terrestres. La categoría biológica de humedal comprende zonas de propiedades geológicas diversas: bañados, ciénagas, esteros, marismas, pantanos, turberas, así como las zonas de costa marítima que presentan abnegación periódica por el régimen de mareas (manglares).

1.4 Contaminación

Cuando las formas de sustancias o energía son de tal manera que los seres vivos o el ambiente acuático los pueden asimilar, transformar o eliminar continuamente, se puede considerar que existe una situación estable. Sin embargo, en la actualidad, el equilibrio que existe en el ambiente se altera debido al gran aumento en la cantidad de sustancias que entran continuamente y en muchos casos rebasando la capacidad de los ambientes acuáticos para transformar las sustancias naturales, o bien, estos carecen de la capacidad para asimilar, transformar o eliminar las sustancias sintéticas. Como consecuencia de esto, sobreviene la acumulación de sustancias y energía en los sistemas acuáticos. Esta acumulación se conoce como contaminación (5).

Un amplia definición de contaminación es "la introducción de sustancias o energía por el hombre o la naturaleza en el ambiente responsable de causar peligros a la salud humana, daño a los recursos vivientes, sistemas ecológicos y servicios productivos, o interferencia con usos legítimos del ambiente" (6; 9).

Un aumento de la contaminación degrada los diferentes medios acuáticos dentro de los ciclos naturales. Por esto, la contaminación de estos medios es un grave problema y con serias consecuencias que ha afectado a numerosas zonas a pesar de la autopurificación del propio ciclo. Uno de los principales propósitos en la actualidad es la búsqueda de

adecuados suministros de agua dulce para las crecientes necesidades manteniendo su calidad (10).

1.4.1 Naturaleza de los contaminantes

Se considera como contaminante a una sustancia presente en concentración mayor que lo natural como resultado principalmente de la actividad humana que tiene un efecto perjudicial en el medio ambiente o sobre algo de valor en ese ambiente (1). Los contaminantes tienen ciertas propiedades intrínsecas que determinan el probable efecto que tendrán después de las emisiones o descargas en el ambiente (6). Conforme a la naturaleza del agente contaminante, se suele distinguir a los contaminantes generalmente en tres tipos: químico, biológico y físico (Tabla N° 1-1).

Tabla N° 1-1. Tipos de contaminantes comunes en el agua superficial (2)

TIPO	EJEMPLO
Químico	Compuestos orgánicos iones inorgánicos Material radiactivo
Biológico	Coliformes Virus Plancton Maleza Acuática
Físico	Temperatura Color Espuma Olor Ruido

1.4.2 Tipos de contaminación

Atendiendo al modo de producción de la contaminación, se puede distinguir dos tipos:

PUNTUAL: Es producida por un foco emisor determinado afectando a una zona concreta e identificada por sus orígenes de las actividades antropogénicas, lo que permite una mejor difusión del vertido. Su detección y su control son relativamente sencillos (1).

Un ejemplo de contaminación puntual sería el vertido de aguas residuales industriales o domésticas a un ambiente acuático.

DIFUSA: Su origen no está claramente definido pues aparecen en áreas más extensas en las que coexisten múltiples focos de emisión, lo que dificulta el estudio de los contaminantes y su control individual (1). Pueden producirse posibles interacciones que agraven el problema. Principalmente corresponden a la contaminación natural, a las escorrentías agrícolas, ganaderas, urbanas y las emisiones de los vehículos (11).

1.4.3 Fuentes de contaminación

Las fuentes son particularmente importantes porque generalmente es el lugar indicado para eliminar la contaminación (1). Los diferentes parámetros de contaminación proceden principalmente de diferentes actividades antropogénicas, tales como: industrial, agrícola, urbana y en otros casos provenientes de las fuentes naturales (Figura N° 1-2).

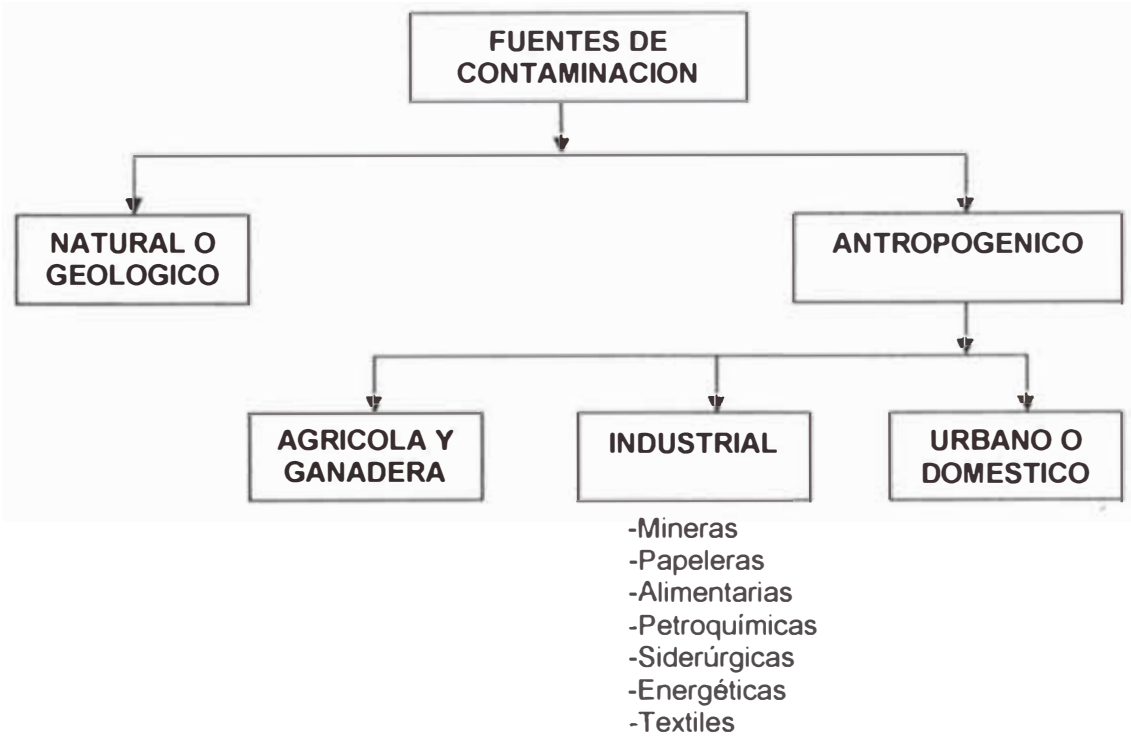

Figura N° 1-2. Fuentes de contaminación.

Los diferentes mecanismos de entrada de los parámetros de contaminación al agua superficial pueden verterse directamente, introducidas desde la atmosfera, escorrentía del suelo por las lluvias, y por el lixiviado de residuos inorgánicos y orgánicos que han

sido depositados inadecuadamente en las riberas ó en los vertederos. Dado que los parámetros del agua que pueden ocasionar contaminación pueden provenir de diferentes fuentes de origen, es importante clasificar tales fuentes (1). De esta manera, las fuentes de origen de los parámetros del agua se dividen en dos categorías principales.

1.4.3.1 Natural ó Geológico

Las aguas superficiales naturales son soluciones que contienen material suspendido y compuestos diluidos de diferente complejidad. Esto se debe al estrecho contacto que estas tienen con los compuestos químicos de la litosfera, la atmosfera y la biosfera (2). La contaminación natural o geológico de las aguas superficiales suele estar limitada en el tiempo y en el espacio, ya que está asociada a circunstancias biogeoclimaticas especiales y mucho de los compuestos químicos en la naturaleza forman parte de los ciclos de la materias, conocidos como ciclos biogeoquímicos (Anexo 1) (1; 5; 10).

En la Tabla N° 1-2 se presenta una lista de los compuestos más comunes de las aguas superficiales dulces. Algunos de tales compuestos son vitales para las plantas y los animales acuáticos. Los compuestos orgánicos naturales provienen de la elusión de los minerales de la corteza terrestres, compuestos del suelo, y de la incorporación desde los sedimentos de los procesos metabólicos de seres vivos y de los residuos de animales y vegetales (12). Por otro lado, varios de estos compuestos químicos interfieren con el uso que se destina el agua y, por tanto, se les considera contaminantes naturales.

Tabla N° 1-2. Componentes naturales de las aguas superficiales dulces (2; 9)

FUENTE DE ORIGEN	EN SUSPENSION	SOLUCION
Litosfera (minerales y rocas)	Arena Arcilla Suelos	Fe^{2+}, Mn^{2+}, Zn^{2+}, Na^{+} K^{+}, Ca^{2+}, Mg^{2+}, Cl^{-}, HCO_3^{-}, SO_4^{2-}, PO_4^{3-}
Atmosfera	Polvos y macropartículas	O_2, N_2, CO_2, PM_{10}, $PM_{2,5}$
Biosfera	Algas, animales acuáticos y otra plantas Bacterias	H_2S, CO_2, CH_4, O_2, N_2 Ácidos húmicos, fúlvicos e hidrofílicos

Todos estos compuestos sufren una serie de procesos químicos y biológicos que forman parte de la capacidad autodepuradora del agua y en su mayoría son eliminados. La presencia de casi todas estas impurezas escapa al control humano, y es imposible eliminarlas (2).

1.4.3.2 Antropogénico

Esta fuente representa la mayor contribución a la mayor variedad de parámetros en el agua. Cuantitativamente, los parámetros registrados son menores que las fuentes naturales pero sus efectos se multiplican porque sus efluentes se localizan en áreas reducidas, que a su vez son las que tienen mayor cantidad de población, y además porque sus emisiones son más intensas.

A diferencia de la contaminación natural, la contaminación de origen antropogénico puede ocurrir también por la entrada al ambiente de sustancias sintéticas (xenobióticas) y, por lo tanto, tiene una gran variedad de causas. Por lo mismo, ocasiona diversos problemas y efectos adversos tanto a corto como a largo plazo (5). Existen 3 principales fuentes de contaminación de origen antropogénico:

a) Urbano o doméstico

La contaminación de origen urbano o doméstico es el resultado del uso del agua en viviendas, actividades comerciales y de servicios, lo que genera aguas residuales, que son devueltas al agua superficial con contenidos de residuos fecales (con alta carga biológica), desechos de alimentos (grasas, restos orgánicos, etc.), y en la actualidad con un importante incremento de productos químicos (lejías, detergentes, cosméticos, etc.).

b) Agrícola y ganadera

La contaminación de origen agrícola deriva, principalmente, del uso de fertilizantes, abonos, herbicidas, pesticidas, fungicidas, residuos agrícolas, etc., los cuales son arrastrados por el agua de riego antes de entrar a los ambientes de agua superficial (fenómeno llamado escorrentía), llevando consigo sales compuestas de nitrógeno, fósforo, azufre y trazas de elementos organoclorados. La agricultura incrementa la

cantidad de la escorrentía removiendo la vegetación, la compactación del suelo, el decrecimiento de la población animal del suelo (insectos, gusanos, etc.) y disminuyendo las concentraciones de la materia orgánica (13).

En explotaciones ganaderas, la contaminación procede de los restos orgánicos que caen al suelo y de vertidos con aguas cargadas de materia orgánica. Esto sucede debido a que los desechos de animales contienen nitratos y fosfatos que se pueden aprovechar cuando se usan correctamente (fertilizantes) y pueden ser dañinos cuando terminan en las fuentes de agua (2).

c) Industrial

La contaminación de origen industrial es la que produce un mayor impacto por la gran variedad de materiales y fuentes de energía que pueden aportar al agua: materia orgánica, metales pesados, incremento de pH y temperatura, radioactividad, aceites, grasas, petroquímicos, detergentes, residuos sólidos, etc. Entre las industrias más contaminantes se encuentran las petroquímicas, las agroalimentarias, las energéticas (térmicas, nucleares, hídricas, etc.), papeleras, siderúrgicas, alimenticias, textiles y mineras.

1.5 Programa de monitoreo

El monitoreo ambiental es la observación y estudio del ambiente acuático, terrestre o atmosférico, el cual involucra el registro de muchos parámetros para evaluar el estatus de cada ambiente y sus interrelaciones con las diferentes influencias antropogénicas o naturales (14). Para el caso de la contaminación actual y el daño potencial o real a los ambientes acuáticos, se requiere el desarrollo de una red de monitoreo, el cual es el elemento crítico en la evaluación, restauración, y protección de la calidad del agua. Esta red de monitoreo se ha vuelto cada vez mas importante debido al aumento de las poblaciones humanas que añade altas concentraciones y en grandes cantidades, componentes contaminantes a los diferentes ambientes (14; 15).

De esta manera, estas redes o sistemas de monitoreo permiten el control de las sustancias acuáticas de las aguas domésticas, las aguas superficiales, y el control de efluentes. Tales sistemas están muy relacionadas a las mediciones analíticas "clásicas" tales como la colección de muestras y los análisis de laboratorio.

Pero también existen otras mediciones sofisticadas que están basadas en el uso de técnicas automáticas, esto es, sensores para obtener sistemas adecuados de señales de prevención y facilitar la gestión de los recursos y procesos hídricos para tomar decisiones ambientales adecuadas.

1.6 Parámetros de análisis

Los programas de monitoreo supervisan la calidad del agua y requieren de la recolección de una gran cantidad de parámetros fisicoquímicos y biológicos, considerados como parámetros de análisis, tomados en diferentes tiempos y sitios. Estos datos obtenidos forman la piedra angular de los procesos de toma de decisiones con respecto al estatus de la calidad del agua (3).

Tabla N° 1-3. Parámetros mas analizados en los programas de monitoreo (8)

PARAMETROS FISICOS	PARAMETROS INORGANICOS	PARAMETROS ORGANICOS	PARAMETROS BIOLOGICOS
Color	Alcalinidad	Tolueno	Coliformes fecales
Caudal	Cloruros	Xileno	Coliformes totales
Conductividad	Durezas	Etilbenceno	Salmonella
Color	DBO5	Estireno	Escherichia Coli
Olor	DQO	Detergentes	Vibrion Cholerae
pH	Sulfatos	Fungicidas	Zooplancton
Sólidos Totales	Nitratos	Herbicidas	Fitoplancton
Sólidos Disueltos	Nitritos	Fertilizantes	Huevos de Helmintos
Sólidos Suspendidos	Fosfatos	Pesticidas	Enterococos
Turbidez	Metales	Aceites y grasas	
Temperatura	No metales	Ácidos húmicos	
Oxigeno disuelto	Cianuros	Ácidos fúlvicos	

Así en la Tabla N° 1-3 muestra los parámetros más importantes en el marco de los programas de monitoreo o para propósitos de actividades de investigación.

Adicionalmente, existe una lista de más de 75 parámetros de análisis específicos en los estándares para el agua potable según la Agencia para la Protección Ambiental de los Estados Unidos (EPA) (4).

1.7 Métodos de análisis

En la actualidad existe un periodo de evolución de la investigación científica de los ambientes acuáticos caracterizada por el desarrollo de nuevas y mejores técnicas analíticas e instrumentales, pero al mismo tiempo surgen una serie de problemas a solucionar (16).

Para llevar a cabo tales investigaciones es necesario que se conozca la naturaleza y cantidad de los contaminantes químicos y de otras especies en el agua. Por consiguiente, el desarrollo de las nuevas técnicas de análisis de agua es fundamental para la obtención de información precisa para la solución de los diferentes problemas de los ambientes acuáticos. Sin embargo, existen diferentes técnicas de análisis de agua para el mismo parámetro ambiental, por lo cual se tiene que elegir el más adecuado dependiendo de las características de cada muestra de agua, los instrumentos y reactivos de laboratorio.

Uno de los métodos de análisis más usados a nivel mundial pertenece a la Asociación Americana de Salud Pública (APHA). Esta asociación creó un comité para estudiar los diferentes métodos analíticos disponibles y publicó las recomendaciones del comité como "Métodos Estándares de Análisis de Agua" en 1905 (4). Desde esa época, con la colaboración de la American Water Works Association y la Water Environment Federation, el campo que abarcan los "Métodos Estándares" se ha extendido hacia diferentes tipos y análisis de agua y que actualmente se encuentra en la vigésimo primera edición (2005) (17).

La importancia del uso de los "métodos estándares" permite asegurar la calidad de los contaminantes ambientales, las cuales tienen una influencia directa en la formulación de la política ambiental. Puesto que, los datos de todos los contaminantes producidos en el

marco de los programas de monitoreo forman el centro del proceso de toma de decisiones con respecto al estatus de la calidad del agua. Estos datos, junto con la información asociada, esto es, con la ubicación de la muestra y las condiciones de muestreo, deben tener un alto y definido nivel de confiabilidad. Esto permite que la interpretación y las regulaciones realizadas tengan una base sólida para que en posteriores procesos puedan ser tomadas en cuenta.

1.8 Regulación de la calidad del agua

EL objetivo de la regulación de la calidad del agua es alcanzar el desarrollo ambiental sostenible que implica la puesta en marcha de un conjunto de vías de progreso económico, social y político que atienda las necesidades del presente sin comprometer la capacidad de las generaciones futuras (10).

Existen diferentes instituciones internacionales que establecen el control de la calidad de las aguas para las diferentes actividades antropogénicas y de preservación del agua. Estas reglamentaciones establecen parámetros legales básicos tanto para sus niveles de calidad como su obtención. De esta manera los países del mundo toman como referencias estas reglamentaciones internacionales para regular la calidad del agua. En el Perú, las leyes y regulaciones ambientales están relacionadas a los límites máximos permisibles de los efluentes industriales, y los estándares nacionales de calidad ambiental del agua (ECA).

Esta nueva regulación de los estándares nacionales de calidad ambiental de agua fue emitida en Julio del 2008 (Anexo 2), el cual está basado en la calidad del agua para el uso recreativo, riego y preservación de los diversos sistemas acuáticos.

1.9 La estadística en el análisis del agua (16; 18; 19)

La estadística ambiental, en este estudio referida al análisis del agua, se ha desarrollado rápidamente en los últimos 10 ó 15 años en respuesta a un incremento de la preocupación de las personas, organizaciones y gobiernos para proteger los diferentes ecosistemas acuáticos.

Este tipo de análisis estadístico difiere de otras aplicaciones estadísticas (i.e. estadística industrial, estadística médica, etc.) debido al énfasis para la aplicación de una gran variedad de métodos y modelos necesarios para la evaluación inicial, monitoreo y el respectivo control de la contaminación del agua para su respectivo manejo sostenible. Este análisis estadístico está también relacionado a los datos ambientales que influyen en la calidad del agua, tales como: el manejo de la fauna y flora, el cambio climático, el efecto invernadero y las actividades antropogénicas (pesca, agricultura y ganadería).

De esta manera, surgen muchos problemas tanto básico como complejos del estudio del agua, y por el cual existen muchas técnicas estadísticas, desde realizar simples gráficos de datos hasta construir modelos iterativos y estimación de parámetros. Por esto, es necesario conocer la naturaleza y las principales características estadísticas de los datos ambientales que tiene los ecosistemas acuáticos, tales como:

Parámetros en trazas: los análisis de datos ambientales en el agua están conectados con los problemas de análisis de trazas y ultratrazas. Estos análisis requieren de un alto grado de exactitud y precisión debido a que la medida de los parámetros se encuentra en concentraciones de ppb y ppt, los cuales pueden provocar daños humanos y ecológicos.

Existencia de muchos parámetros: la característica de cualquier ambiente acuático es multiparamétrica, esto es, la existencia de muchos parámetros (físicos, químicos y biológicos) en un sólo ambiente. Estos parámetros pueden provenir de la industria, la agricultura, la ganadería y la población que emiten mezclas de productos que contienen muchos compuestos. Para considerar simultáneamente a todos, y sus potenciales reacciones e interacciones, se requiere la aplicación de técnicas estadísticas multivariadas supervisadas y no supervisadas.

Múltiples fuentes de contaminación en el ambiente: el mecanismo de transporte de la mayoría de los parámetros contaminantes o no en el ambiente acuático es en general desconocido. Este mecanismo esta a menudo asociado con cambios en el carácter químico y la concentración de las especies químicas individuales en el ambiente.

Comportamiento variable: Los diferentes ambientes acuáticos nunca están estáticos, puesto que existen interacciones de los compuestos fisicoquímicos del agua, suelo, la atmosfera y los seres vivos (plantas, animales y seres humanos). Estas interrelaciones dependen de múltiples factores naturales y antropogénicos, produciendo reacciones químicas, físicas y bioquímicas.

Valores atípicos o extremos: los valores que no se encuentran en el rango común o en la tendencia de los valores analizados en ciertos ecosistemas acuáticos son bastantes comunes. Estos valores pueden ocurrir debido a una nueva fuente de contaminación como los efluentes o descargas de contaminantes provenientes de la actividad antropogénica. Por otro lado también se debe a los errores de factor humano tales como: producidos en el muestreo, en los análisis o por errores de apunte, por lo que rechazar o modificar estos valores conlleva a errores de interpretación de resultados.

Distribución de los datos ambientales: el problema de la distribución es debido a las diferentes técnicas estadísticas que tratan con modelos lineales, y cuyo procedimiento se realiza con datos que tienen una distribución característica e independiente; sin embargo, en el caso de los datos ambientales, estas asunciones no son válidas. Por tanto, se tiene que estudiar inicialmente la distribución de los datos y modificarlo antes de realizar los métodos estadísticos a desarrollarse.

La estadística ambiental está relacionada también a muchos otros problemas que involucra el agua, tales como: económico, político, médico, científico y tecnológico. La comprensión y la solución de tales problemas están relacionadas con el análisis cuantitativo del agua, en particular en la adquisición y análisis de datos.

QUIMIOMETRIA EN EL ANALISIS AMBIENTAL

2.1 Introducción

Por mucho tiempo el rango extenso de los métodos matemáticos y estadísticos ha proporcionado una excelente oportunidad para la descripción cuantitativa de los efectos y resultados experimentales de las ciencias naturales. Estos métodos son aplicados, en particular, en aquellas disciplinas científicas que investigan los efectos provenientes de muchas influencias (16).

Inicialmente, la adquisición de los datos era un paso limitante en los procesos analíticos clásicos. Esta situación cambio notoriamente en la década de los 50s debido a la introducción de una gran cantidad de nuevos métodos instrumentales en la química analítica, los cuales permitieron que el campo de la química analítica se vuelva cada vez más compleja debido al aumento de los datos químicos en décadas recientes (16).

No obstante, junto con el aumento de los datos químicos vino el desarrollo de la ciencia informática, una herramienta muy poderosa para la solución de problemas matemáticos complejos y que permite almacenar y tratar a altas velocidades un elevado número de datos analíticos. Esta herramienta hizo más fácil para los químicos, especialmente para los químicos analíticos, utilizar computadoras con métodos matemáticos y estadísticos avanzados en sus propios campos de trabajo (16).

Así, la ciencia informática permitió una reducción, representación clara (en términos de visualización), y extracción de la información relevante sobre la gran cantidad de datos. Además posibilitó la descripción detallada y cuantitativa de los mecanismos y de las relaciones entre los distintos componentes del entorno (16; 10).

Por lo tanto, esos dos desarrollos condujeron a la formación de una nueva subdisciplina química, llamada quimiometría, la cual ha sido definida como: "Una rama de la química que usa la matemática, la estadística descriptiva e inferencial para diseñar o seleccionar procedimientos experimentales óptimos, extraer a partir de los análisis de los datos experimentales la máxima información química relevante y extender el conocimiento de los sistemas químicos" (16; 20; 21; 22).

La aplicación de la quimiometría en el análisis ambiental puede ser muy útil para solucionar ciertos problemas ambientales, permitiendo la planificación y optimización de los procesos analíticos en las diversas etapas de un análisis ambiental, diseñando las operaciones de monitoreo y los experimentos analíticos.

De esta manera, la quimiometría permite la comprensión de un gran conjunto de parámetros adquiridos en el monitoreo ambiental, tales como: la eliminación de información redundante, la visualización de relaciones entre parámetros y estaciones de muestreo, la identificación y caracterización de fuentes de contaminación. Esto permite una mejor interpretación de los resultados para los trabajos de línea base en la evaluación del impacto ambiental y planes de manejo ambiental (10; 16).

2.2 Clasificación de los métodos quimiométricos

El rango de estudio de la quimiometría va de la estadística simple a los análisis de datos sofisticados, por lo que esta gran variedad de métodos requiere de una clasificación que permite una orientación adecuada. La Figura N° 2-1 provee una visión general de los métodos quimiométricos empleado en los análisis e interpretaciones de datos.

Sin embargo, actualmente el principal interés de estudio está enfocado en los análisis multivariado de los datos. En general, para la correcta aplicación de los métodos quimiométricos se tiene que tener en cuenta que *la interpretación de los resultados de estos métodos deben estar relacionadas con la realidad para obtener buenos resultados.*

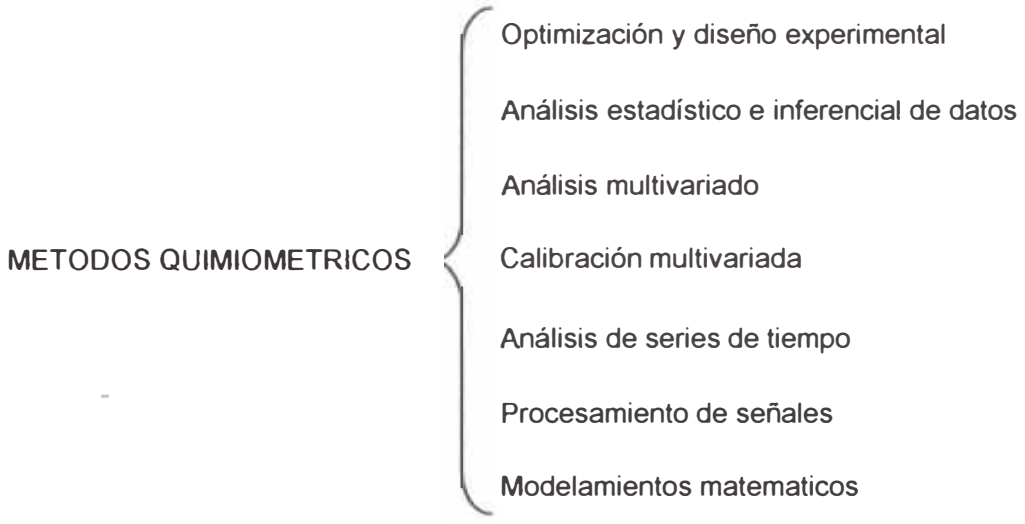

Figura N° 2-1. Métodos quimiométricos para el análisis e interpretación de datos (23).

2.3 Análisis y pretratamiento de datos

2.3.1 Generalidades

El análisis y pretratamiento de datos es un conjunto de técnicas estadísticas descriptivas e inferenciales que se lleva a cabo antes de los análisis multivariado, y se usan sobre los datos de una matriz original X(n, m) de *n* observaciones y *m* variables (ecuación 1). El uso de estos análisis puede influir de manera positiva o negativa en los resultados al final del proceso de análisis de acuerdo al procedimiento a ser utilizado (24).

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & x_{ij} & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad (1)$$

El análisis y pretratamiento de datos son la parte más importante en cualquier análisis quimiométrico, puesto que analiza el comportamiento inicial de las variables, esto es, detecta anomalías o errores en las distribuciones univariante de los datos, y remueve o reduce fuentes irrelevantes de variación (aleatoria o sistemática) (21).

2.3.2 Relleno de espacios vacíos (21)

La condición más importante para llevar a cabo el análisis y pretratamiento de datos es que la matriz de datos sea completa, debido a que una matriz incompleta de datos produce resultados inconclusos o erróneos.

Estos espacios vacios se produce por diversos motivos, tales como:

a) Algunos datos no se han podido obtener a causa de un error accidental durante la experimentación y no es posible repetir las medidas en las mismas condiciones.
b) Algunos datos se hallan por debajo del límite de detección (LDD) y, por tanto, su valor no se puede conocer sin cambiar el método.

Puesto que no es posible operar con matrices de datos incompletos, es necesario eliminar o rellenar los espacios vacios de algún modo. Sin embargo, eliminar los vacios

es una mala alternativa y procediendo de esa manera se desecha la información útil aportada por los datos disponibles en filas o columnas eliminadas. La forma correcta es aplicar los criterios conocidos como "relleno medio" y "relleno al azar" que atribuyen valores supuestos a los datos que no están disponibles.

Así en la aplicación del relleno medio, el vacio se sustituye por la media de la variable, o si es posible, por la media de la categoría a la que pertenece el objeto. En la técnica del relleno al azar, el vacío se sustituye por un valor al azar, si bien, acotado dentro de los límites de la variable, o si es posible, dentro de la categoría a la que pertenece el objeto.

2.3.3 Normalidad multivariada

La normalidad multivariada es la asunción de que cada uno del conjunto de datos de las variables de estudio y todas sus combinaciones lineales tengan una distribución normal o llamada "distribución de Gauss" de sus grupos de datos. Puesto que la importancia del estudio de la normalidad se debe a que muchos datos coleccionados en los estudios del medio ambiente tienen una distribución exponencial aproximada o no tienen ninguna distribución.

Sin embargo, proceder con los procedimientos estadísticos multivariados y las pruebas estadísticas de los resultados sin el uso de los cálculos matematicos adecuados llevaría a conclusiones erróneas, ya que la mayoría de estos procedimientos estadísticos se basan en la hipótesis de la distribución normal multivariado o distribución gaussiana de los datos de las variables de estudio (25; 26) La falta de normalidad multivariada es violada si cualquiera de las variables individuales no están normalmente distribuidas o no tienen homocedasticidad

Por esto, la necesidad de que los grupos de datos de las variables tengan una tendencia hacia la distribución normal; ya que resulta más accesible realizar estos procedimientos matematicos; siendo los resultados de fácil interpretación. Así, una manera corta para examinar la normalidad multivariada es examinando cada variable individual dentro de la base de datos (25), en el cual la falta de normalidad se comprueba examinando las características de las variables individuales, aun en muestras con una gran cantidad de

datos, que tienden a disminuir los efectos de la falta de normalidad debido al Teorema del Límite Central (20; 25; 26). Existen varios métodos para evaluar la normalidad en un conjunto de datos y se pueden dividirse en dos grupos: los métodos descriptivos y los métodos inferenciales.

2.3.3.1 Métodos descriptivos

Los métodos descriptivos están basados en gráficos, en el cual se observa la distribución de los datos, tales como: Histogramas, gráficos de Cajas y Bigotes, gráficos Q-Q y P-P. Siendo, el histograma el más importante, pues permite visualizar los datos observados distribuidos en rangos y frecuencias con una distribución normal hipotética, calculando los estimadores descriptivos estadísticos principales de la normalidad, tales como: la media, la desviación estándar, la curtosis y el sesgo (15; 16; 28).

Mientras que los gráficos de cajas y bigotes es una herramienta estadística que puede ser formado a partir de los estadísticos descriptivos paramétricos y no paramétricos, tales como: la mediana, los percentiles, el máximo y mínimo. Estos gráficos son de gran utilizada para comparar y visualizar la distribución de los datos iniciales (16; 20; 29).

2.3.3.2 Métodos Inferenciales

La segunda forma para comprobar la normalidad de una distribución se efectúa a través de la estadística inferencial por medio de la test de hipótesis. Sin embargo, no existe un test óptimo para probar la hipótesis de normalidad y la razón es que la potencia relativa del test depende del tamaño muestral y de la distribución natural que generan los datos.

Existe una extensa lista de test o pruebas estadísticas diseñadas para verificar la distribución normal de un conjunto de datos. Entre las pruebas más utilizadas podemos mencionar a las de Shapiro-Wilk y Kolmogorov-Smirnov.

a) Prueba de Kolmogorov-Smirnov: Esta prueba calcula la distancia entre la función empírica de la muestra y la distribución normal teórica. La hipótesis nula (H_o) que se pone a prueba es que los datos proceden de una población con una distribución normal frente a una alternativa (H_1) de que no es así. Si la distancia calculada es mayor que la

encontrada en las tablas y el p-valor menor que el fijado a un cierto nivel de confianza (mayormente un p-valor de 0.05 o 95% de nivel de confianza) entonces se rechaza la hipótesis nula.

b) Prueba de Shapiro-Wilks: Se basa en calcular el coeficiente de correlación, r, entre dos grupos de datos y cuanto más cerca este a 1, mayor será el grado de normalidad de la distribución. El contraste evalúa la distribución del estadístico r^2 bajo la hipótesis nula (H_o) de que los datos proceden de una población con distribución normal y proporciona un p-valor que rechaza normalidad a un cierto nivel de confianza (mayormente a un p-valor de 0.05 o 95% de nivel de confianza dicha), cuando el ajuste es bajo, es decir, cuando el estadístico toma valores bajos.

Desde un punto de vista de la cantidad de datos de la variable, el contraste de Shapiro Wilks es, en términos generales, más conveniente en muestras pequeñas (numero de observaciones, n, menor que 30), mientras que el contraste de Kolmogorov- Smirnov, es adecuado para muestras grandes.

2.3.4 Transformaciones de datos

La transformación de datos permite que los procedimientos en los análisis multivariado y los modelamientos posteriores, las cuales se basan en ciertas suposiciones matemáticas, sean desarrollados adecuadamente. Esta transformación se realiza separadamente para los grupos de datos de cada variable de la matriz de datos X (ecuación 2), debido que no es seguro si una misma transformación será útil para las diferentes variables.

$$X^T = \begin{pmatrix} x_{11}^T & x_{12}^T & \cdots & x_{1m}^T \\ x_{21}^T & x_{22}^T & \cdots & x_{2m}^T \\ \vdots & \vdots & x_{ij}^T & \vdots \\ x_{n1}^T & x_{n2}^T & \cdots & x_{nm}^T \end{pmatrix} \quad (2)$$

Las transformaciones pueden ser necesarias para que los análisis posteriores sean desarrollados rápidamente, esto es, los datos transformados "adopten las asunciones" de los análisis y modelos de comportamiento generalmente lineal, razón por la que es

usado muy frecuentemente en trabajos científicos. Sin embargo, aunque la transformación de datos son recomendados como un remedio para evitar los atípicos y para las faltas de normalidad, linealidad, y homocedasticidad, estos no son universalmente recomendados. Seis transformaciones se usan comúnmente con los datos ecológicos y ambientales para la transformación de de los datos (Tabla N° 2-1).

Tabla N° 2-1. Tipos de transformaciones de datos más comunes (25)

TRANSFORMACION	FUNCION
Logarítmica	Log(x) ó Ln(x)
Raíz cuadrada	$\sqrt{x}$
Arcoseno	$sen^{-1}(x)$
Reciproca	1/x
Potencia	x^p
Box- Cox	f(x)

La *transformación logarítmica* (de mayor uso el logaritmo natural o logaritmo a la base e) es útil para distribuciones de datos con sesgo a la derecha, la *transformación de la raíz cuadrada* se usa mayormente en datos de recuento, la *transformación arcoseno* está relacionada principalmente a los datos asignados a proporciones (porcentajes) y la *transformación reciproca* se usa mayormente en datos de índices de registro.

Una transformación más flexible es *la transformación de potencia* que usa una potencia *p* para transformar los valores de *x* a x^p. Los valores de *p* tienden a ser optimizados para cada variable; cualquier número real es bastante razonable para p, excepto para p = 0 donde una transformación logarítmica es realizado (25).

Una forma particular de las transformaciones que ayudan a transformar los datos originales de una manera que la nueva variable tenga datos con una distribución tan cercana a la normalidad como sea posible es conocida como *transformación de Box Cox* (15; 18; 19). Esta transformación es una versión ligeramente modificada de la transformación de potencia, y definida para valores x positivos (ecuación 3) (22).

$$x_{Box\text{-}Cox} = \begin{cases} (x^{\lambda}-1)/\lambda & (\text{para } \lambda \neq 0) \\ \log(x) & (\text{para } \lambda = 0) \end{cases} \qquad (3)$$

Encontra el λ requiere de un procedimiento iterativo, el cual se basa en tratar de encontrar los diferentes valores de λ hasta que la funcion del logaritmo de la funcion de verosimilitud L sea maximizado (ecuación 4). Pero incluso el λ óptimo no garantiza que los datos transformados por Box-Cox tengan una distribucion normal.

$$L = -\frac{n}{2}\log(s_T^2) + (\lambda - 1)\sum_{i=1}\log(\hat{x}_i) \qquad (4)$$

Donde n es el numero de datos a transformar, $\hat{x}_i$ es el iésimo dato transformado y s_T^2 es la varianza de la variable transformada. En casos de valores x negativos, se le agrega una constante λ_2, tal que $x + \lambda_2 > 0$, resultando la ecuacion 5.

$$x_{Box\text{-}Cox} = \begin{cases} ((x+\lambda_2)^{\lambda_1} - 1)/\lambda_1 & (\text{para } \lambda_1 \neq 0) \\ \log(x+\lambda_2) & (\text{para } \lambda_1 = 0) \end{cases} \qquad (5)$$

2.3.5 Autoescalado o Transformación Z (16; 21)

El autoescalado se refiere a las manipulaciones de los datos de las variables de una matriz de datos transformados (ecuación 2) con el propósito de que todas las variables contengan datos con media cero (centrado) y una varianza de uno (escalado), de este modo todas las variables tienen igual peso estadistico.

El autoescalado es normalmente aplicado despues de la transformación de los datos y es logrado sustraendo la media de la variable j, $\bar{x}_j$, de sus datos originales x_{ij} y dividido por la desviación estándar s_j de la respectiva variable (ecuación 6). Donde z_{ij} son los datos autoescalados de la variable j.

$$Z_{ij} = \frac{\bar{x}_{ij}^T - \bar{x}_j^T}{s_j} \qquad (6)$$

El autoescalado desplaza el centroide de los puntos de los datos al origen y cambia la escala de los ejes, por consiguiente las distancias relativas entre los puntos de los datos cambian. Esta etapa es la parte más importante del análisis exploratorio de datos en la quimiometria para evitar cualquier efecto de la escala de unidades de las variables en las mediciones de distancia y correlaciones entre ellas. Una desventaja es la obtención de

variables con valores pequeños. Por otro lado, si los datos incluyen valores extremos o atípicos, se aconseja usar versiones robusta de centrado y escalado. La posibilidad más simple es reemplazar las medias aritméticas de las columnas por las medianas de columnas, y la desviación estándar de las columnas por la desviación absoluta de la mediana (MAD).

2.3.6 Matriz de Varianza y Correlación

La relación entre las variables puede expresarse de forma numérica mediante la matriz de varianza-covarianza y de correlación. A partir de las varianzas (ecuación 7) y covarianzas de todas las m variables transformadas (ecuación 8), es calculada la matriz de varianza-covarianza o llamada matriz de varianza (ecuación 9) (20).

$$s_{ij}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 \qquad \text{para } j = 1,\ldots,m \tag{7}$$

$$Cov(j,k)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k) \qquad \text{para } j,k = 1,\ldots,,m\ ;\ j \neq K \tag{8}$$

$$C=\begin{pmatrix} s_{11}^2 & cov(1,2) & \cdots & cov(1,m) \\ cov(2,1) & s_{22}^2 & \cdots & cov(2,m) \\ \vdots & \vdots & s_{ij}^2 & \vdots \\ cov(m,1) & cov(m,2) & \cdots & s_{mm}^2 \end{pmatrix} \tag{9}$$

Como resultado se obtiene la matriz simétrica de covarianza usada en casos donde las escalas de los datos de las variables son iguales o comparables. Sin embargo, si las escalas son diferentes, las variables son escaladas. De esa manera, si los datos de las variables son autoescalados se obtiene la matriz de correlación (ecuación 10) (27).

$$R=\begin{pmatrix} 1 & r_{21} & \cdots & r_{m1}) \\ r_{21} & s_{22}^2 & \cdots & r_{m2} \\ \vdots & \vdots & s_{ij}^2 & \vdots \\ r_{m1} & r_{m2} & \cdots & s_{mm}^2 \end{pmatrix} \tag{10}$$

De esta manera, los cálculos de la matriz de correlación y covarianza son prerrequisitos para tener una observación inicial de la idoneidad de la aplicación de los métodos factoriales (ACP y AF), el cual se lleva a cabo mediante las pruebas de esfericidad de Bartlett y Kaiser-Meyer- Olkin (KMO) (7; 16).

2.4 Análisis Multivariado

2.4.1 Generalidades

A pesar del amplio campo de aplicación de la quimiometría, la parte más importante y utilizada de esta es el análisis multivariado para la obtención de resultados relevantes a partir múltiple datos químicos. En general, el análisis multivariado puede definir como: "*el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariados en el sentido de que hay varias variables medidas para cada individuo u objeto estudiado*" (22)

El análisis multivariado es una poderosa herramienta para analizar y estructurar grupos de datos que han sido obtenidos a partir de sistemas complejos, y para realizar modelos matemáticos empíricos que son capaces de predecir los valores de propiedades importantes que no son directamente medibles. Su razón de ser radica en un mejor entendimiento del fenómeno objeto de estudio obteniendo información que los métodos estadísticos simples son incapaces de conseguir. La importancia de todos los analisis mutlivariado en los estudio de los sistemas acuaticos es caracterizar y evaluar la calidad del agua de acuerdo a las variables obtenidas, y evidenciar las variaciones temporales y espaciales causados por las influencias antropogénicas y naturales (7, 30, 31).

2.4.2 Clasificación

Las principales técnicas multivariadas empleadas en aplicaciones ambientales se clasifican de acuerdo al tipo de información que son capaces de extraer (Tabla N° 2-2).

Así, las técnicas de reconocimientos de patrones supervisado es principalmente exploratorio y su finalidad es detectar la presencia de grupos de objetos o variables, mientras las técnicas de reconocimiento de patrones no supervisado abordan problemas de clasificación, en donde se dividen los objetos o variables en grupos y permiten clasificar un nuevo objeto o variables en los grupos existentes. Por otro lado, los métodos factoriales encuentran relaciones entre los objetos y variables cualitativamente mientras que los análisis de correlaciones y regresiones lo hacen de manera cuantitativa.

Tabla N° 2-2. Clasificación de las técnicas multivariadas (10; 16)

TECNICA	OBJETIVO
Técnicas de reconocimiento de patrones no supervisado	
Análisis de clúster, Mapeo no lineal, análisis de componentes principales	Encontrar relaciones y semejanzas (grupos o clases) en los datos
Técnicas de reconocimiento de patrones supervisado	
Análisis multivariado de la varianza (MANOVA), análisis discriminante, K-vecinos próximos (kNN), maquina de aprendizaje lineal, clasificación de Bayes, Modelo suave independiente de analogía de clases (SIMCA), clasificación de clases de diferente dispersión (UNEQ)	Demarcación cuantitativa de las clases a priori, relación entre propiedades de clases y variables
Métodos factoriales	
Análisis de factores, análisis de componentes principales, análisis de correlación canónica	Encontrar factores o relaciones entre las variables y/o objetos cualitativamente
Correlaciones y regresiones	
Regresión lineal múltiple, regresión de componentes principales, regresión en mínimos cuadrados (PLS)	Descripción cuantitativa de las relaciones entre las variables

2.5 Análisis de clúster (AC)

2.5.1 Generalidades

El análisis de clúster (AC) abarca una familia de métodos cuyo propósito básico es agrupar objetos basados en las características comunes que poseen (7). Este análisis es útil para resolver problemas de clasificación, de manera que encuentra y hace visible clústeres naturales dentro de una matriz de datos, el cual contiene un gran número de objetos caracterizados por muchas variables (16).

Los análisis normalmente se refieren a los objetos (en el espacio de la variable), pero también es usado para las variables (en el espacio de los objetos) o para ambos casos, variables y objetos, que en conjunto se nombra como individuos (22). Un ejemplo de la utilidad de agrupación es aplicada a 21 estándares de aminoácidos (Figura N° 2-2) mediante el análisis de clúster. Estos estándares se agrupan de acuerdo a las características de las estructuras de sus compuestos químicos (Figura N° 2-3).

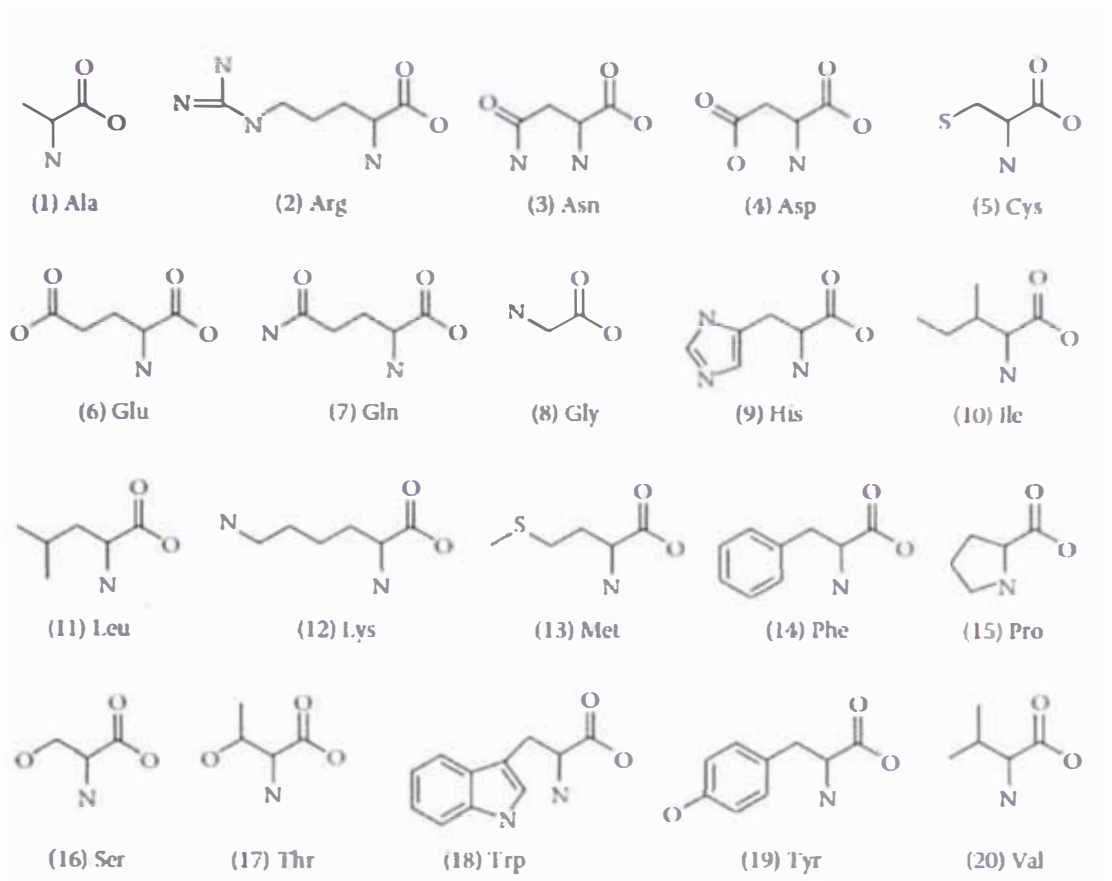

Figura N° 2-2. Estructuras y codificación química de 21 estándares de aminoácidos (22).

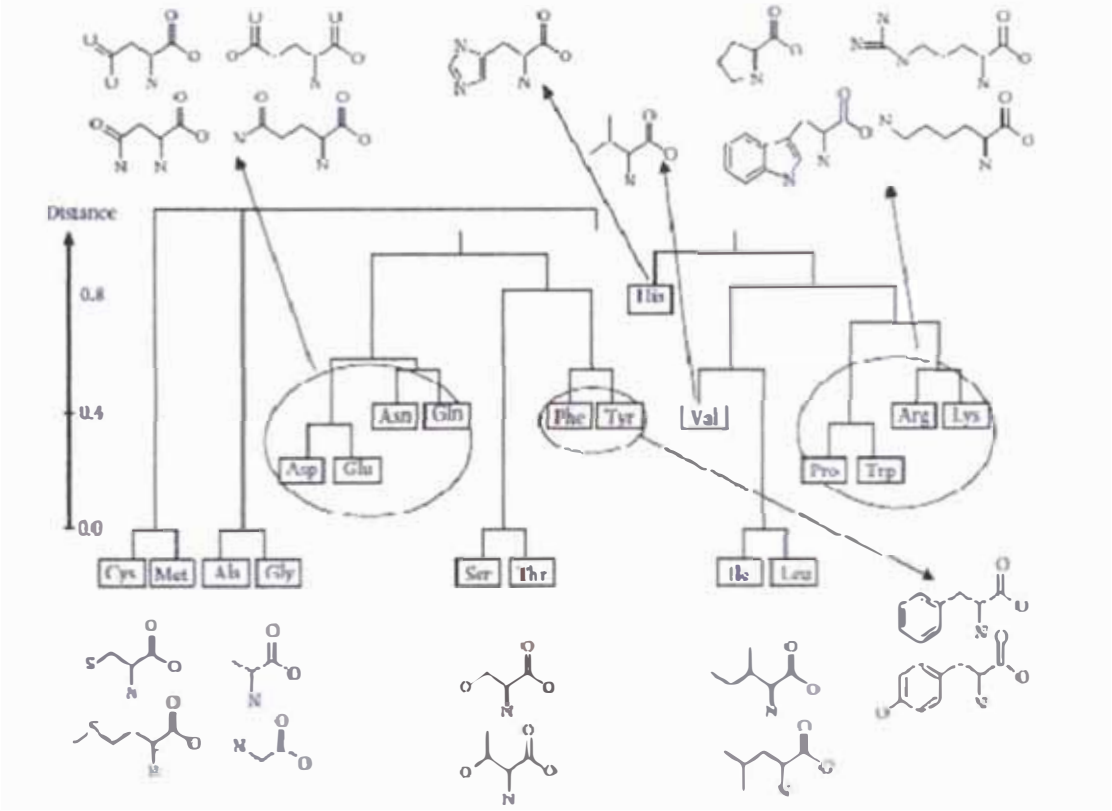

Figura N° 2-3. Agrupación de los 21 estándares de aminoácidos por el análisis de clúster (22)

El análisis de clúster se lleva a cabo mientras no exista información disponible sobre cualquier pertenencia de los individuos a una clase de grupo. Por consiguiente, el análisis de clúster es parte de las técnicas conocidas como técnica de reconocimiento de patrones no supervisado (23).

El análisis de clúster pueda ser dividido en dos etapas fundamentales:

1. La elección de la medida de semejanza: Una medida de semejanza (proximidad) está definida a medir la "cercanía" de cada par de individuos. Mientras más pequeña sea la medida de semejanza entre los individuos, más homogéneas serán.

2. La elección de un algoritmo de agrupamiento: Sobre la base inicial de las medidas de semejanza, el algoritmo es llevado a cabo para asignar individuos a un clúster, y evaluar el correcto agrupamiento de los clústeres debido a que el número "correcto" de clústeres son desconocidos.

El resultado de cualquier método de análisis de clúster es encontrar un agrupamiento óptimo en el cual los individuos dentro de cada clúster tienen un grado de asociación fuerte y los individuos en los distintos clústeres tienen un grado de asociación débil.

Normalmente no se puede esperar una única solución para el análisis de clúster, pues los resultados dependen de la medida de semejanza, el algoritmo de agrupamiento, los parámetros elegidos; y también de las características iniciales de la matriz de datos.

Por lo tanto, el éxito final del análisis de clúster está establecido si cada clúster describe, en términos de los datos coleccionados, la clase al cual los individuos pueden ser asignados.

El análisis de clúster es aplicado en muchos campos tales como las ciencias naturales, medicina, sociología, siquiatría, arqueología, economía, marketing, etc. (26). En cada caso un grupo de datos heterogéneos son analizados con el objetivo de identificar subgrupos homogéneos.

2.5.2 Análisis inicial

El punto de partida de un análisis de clúster es seleccionar y organizar un número de objetos razonable, *n*, y sus *m* variables correspondientes en una matriz de datos autoescalados Z *(n, m)* (ecuación 11) de la matriz de datos transformados (ecuación 2). Esta matriz de datos también puede ser escrita como la ecuación 12. Donde $\bar{x}_i$ es una fila (vector objeto en la fila *i*) y y_j es una columna (vector variable autoescalado en la columna *j*).

$$Z=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & z_{ij} & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{pmatrix} \tag{11}$$

$$Z_x=\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_i \\ \vdots \\ \bar{x}_n \end{pmatrix} \text{ ó } Z_y = \begin{pmatrix} \bar{y}_1 & \bar{y}_2 & \bar{y}_j & \cdots & y_m \end{pmatrix} \tag{12}$$

2.5.3 Medida de semejanza

La idea fundamental en el análisis de clúster para encontrar semejanzas en una matriz de datos entre los individuos es calcular sus medidas de semejanza (Tabla N° 2-3). Así, dos individuos son asignados a la misma categoría o clase y tener las mismas propiedades si sus medidas de semejanza son pequeñas.

La medida de semejanza más usada para el AC es la distancia euclidiana, el cual se deriva de la geometría analítica (ley de Pitágoras) (16; 21). La medida de semejanza City-Block o Manhattan describe una distancia absoluta y puede ser fácilmente entendible.

La medida de semejanza de Minkowski es una generalización de las dos anteriores medidas, y es usada mayormente en el reconocimiento de outliers cuando el factor p sea un número infinito (16).

Mientras, la distancia de Mahalanobis es una medida de distancia introducida por Mahalanobis en 1936, y considera la distribución de los objetos en ciertos grupos en el

espacio de la variable. Esta medida de semejanza calcula la distancia de cada objeto a los centroides de cada grupo, y se diferencia de la distancia euclidiana en que tiene en cuenta la dispersión y la correlación de las variables aleatorias (caracterizado por la matriz de covarianza). El autoescalado de los datos no es necesario en el cálculo de esta distancia (20; 23).

Tabla N° 2-3. Medidas de semejanza más comunes entre los individuos de una matriz de datos

MEDIDAS DE SEMEJANZA	ECUACION
Distancia Euclidiana	$d_{euclidiana}(\bar{x}_i,\bar{x}_j) = d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$, i,j = 1,...,n
Distancia City-Block (Manhattan)	$d_{City\text{-}Block}(\bar{x}_i,\bar{x}_j) = d_{ij} = \sum_{K=1}^{m}\lvert x_{ik}-x_{jk}\rvert$, i,j = 1,...,n
Distancia de Minkowski	$d_{Minkowski}(\bar{x}_i,\bar{x}_j) = d_{ij} = \left(\sum_{k=1}^{m}(x_{ik}-x_{jk})^p\right)^{1/p}$, i,j = 1,...,n
Distancia de correlación de Pearson	$d_{Pearson}(\bar{y}_i,\bar{y}_j) = d_{ij} = 1-\lvert r(\bar{y}_i,\bar{y}_j)\rvert$, i,j = 1,...,m
Distancia de Mahalanobis	$d_{Mahalanobis}(\bar{x}_i,\bar{x}_A) = d_{iA} = [(\bar{x}_i-\bar{x}_A)^T C_A^{-1}(\bar{x}_i-\bar{x}_A)]^{1/2}$, i =1,...,n

C_A^{-1}: Inversa de la matriz de covarianza de los vectores objetos y vector centroide del grupo A ($\bar{x}_A$).

Por otro lado, las *m* variables descritas por la *n* objetos pueden también ser sujetas a análisis de clúster. Una elección para la enlace de semejanza entre cada par de variables es la distancia basada en el coeficiente de correlación de Pearson r (16; 22). Al final del proceso, la distancia entre los individuos es descrita y organizada en matrices de semejanza simétricas y de diagonal de ceros (ecuaciones 13 y 14).

$$D_{objetos} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix} \quad (13)$$

$$D_{variables}=\begin{pmatrix} 0 & d_{12} & \cdots & d_{1m} \\ d_{21} & 0 & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & 0 \end{pmatrix} \quad (14)$$

2.5.4 Clasificación de las técnicas de agrupación

Después de seleccionar la medida de semejanza entre los individuos y obtener la matriz de semejanza, tanto para objeto como variables, se tiene que elegir el algoritmo de agrupamiento apropiado. Para esto, la tarea de identificar clústeres de datos supone que hay una estructura de varios grupos inherente en la matriz de datos.

Sin embargo, las técnicas en general no asumen que un individuo pertenezca a solo un clúster, si no que puede ser parte de dos o más grupos. Así, los algoritmos de agrupación que efectúan una distribución de los individuos en grupos separados no siempre darán la solución deseada. Por esta razón muchas técnicas surgen de acuerdo al tamaño y forma de los clústeres (Figura N° 2-4).

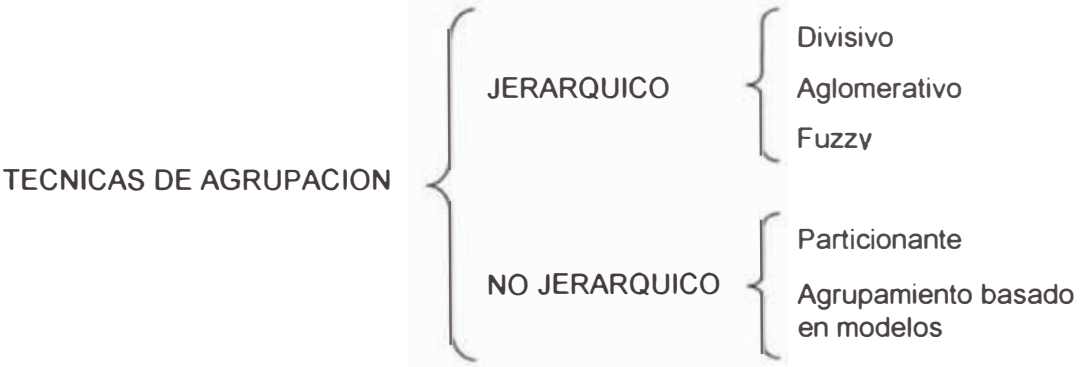

Figura N° 2-4. Técnicas de agrupación del análisis de clúster (22).

2.5.5 Técnicas de agrupación jerárquica

Las técnicas de agrupación jerárquica son las más usadas en los análisis multivariado. Estos algoritmos producen un grupo diverso de clústeres, distribuyendo jerárquicamente en cada etapa los individuos en diferentes clústeres. Existen dos tipos de técnicas (22):

Técnica jerárquica aglomerativo: En la etapa inicial, cada uno de los individuos forma un solo clúster. Luego en cada etapa, cada individuo o clúster es unido en otros clústeres más heterogéneos, hasta que finalmente todas las muestras forman un solo clúster.

Técnica jerárquica divisiva: En la etapa inicial, todos los individuos forman un solo clúster. Luego en cada etapa cada individuo o clúster es separado en clústeres más homogéneos, hasta que finalmente las muestras forman cada uno un solo clúster.

2.5.6 Técnicas de agrupación jerárquica aglomerativo

La técnica jerárquica aglomerativo es más usada en la práctica debido a que la unión de clústeres es matemáticamente menos demandante que la división de clústeres, puesto que, no solamente en cada etapa, el número de clústeres a ser divididos tienen que ser calculados, sino que también los individuos que formarán los nuevos clústeres tienen que ser identificados (22).

2.5.6.1 Tipos de enlace

La información básica para la unión de los clústeres es la medida de enlace entre los clústeres. Se denota como $d_i^{(a)}$ y $d_i^{(b)}$ para i = 1,..., *n* a todas las distancias asignadas a los clúster *a* y *b*, con un número de individuos de clúster n_a y n_b, respectivamente. Entonces la medida de enlace entre los dos clústeres *a* y *b* puede ser determinado por diferentes tipos de enlace (Tabla N° 2-4) (22), tales como:

El enlace promedio, el cual calcula la distancia de enlace promedio (ponderada o no ponderada) de cada par de distancia de los clústeres *a* y *b*. El enlace máximo, el cual define la distancia de enlace entre los dos clústeres *a* y *b* como la máxima distancia de cada par de medidas de semeja pertenecientes a cada clúster. El enlace mínimo define la distancia de enlace entre los dos clústeres *a* y *b* como la mínima distancia entre cada par de distancias pertenecientes a cada clúster.

El método de centroide es casi similar al enlace promedio y calcula el centro de gravedad de cada par de distancias entre los clústeres *a* y *b*. Esto, sin embargo, no conduce estrictamente a que la distancia se incremente dentro de los procedimientos de agrupación (22). El procedimiento de determinar la distancia entre los objetos de un grupo y otro para los 3 primeros tipos de medida de enlace se puede apreciar gráficamente en la Figura N° 2-5.

Debido a que el análisis de clúster es una técnica de reconocimiento de patrones, es necesario obtener resultados que puedan ser fácilmente interpretables. En muchas aplicaciones los resultados que son más fáciles de interpretar son obtenidos por el método de Ward (16).

Así, el método de Ward está relacionado con el factor $n^{\bullet}$, el cual incorpora el número de individuos en el clúster resultante, la suma de estos números, y el numero de muestras en cada clúster (16). Este método es el más utilizado por los trabajos de investigación actualmente.

El objetivo del análisis de clúster es unificar clústeres que tenga la mayor semejanza posible, de manera que la variación dentro de estos grupos no se incremente drásticamente en cada paso de enlace dado, resultando grupos de individuos tan homogéneos como sea posible.

Tabla N° 2-4. Tipos de enlaces de agrupamiento de los clústeres

TIPO DE ENLACE	MEDIDA DE ENLACE	RESULTADO
Enlace promedio (average linkage)	$\text{Promedio}_i\left\{d_i^{(a)}, d_i^{(b)}\right\}$	Dendograma con clústeres moderados
Enlace mínimo (single linkage)	$\text{Mínimo}_i\left\{d_i^{(a)}, d_i^{(b)}\right\}$	Dendograma con pocos clústeres grandes
Enlace máximo (Complete linkage)	$\text{Máximo}_i\left\{d_i^{(a)}, d_i^{(b)}\right\}$	Dendograma con muchos clústeres pequeños
Método de centroide (Centroid linkage)	$\text{Centroide}_i\left\{d_i^{(a)}, d_i^{(b)}\right\}$	Dendograma con tamaño de clústers dependiente del número de muestras
Método de Ward (Ward's method)	$\text{Centroide}_i\left\{d_i^{(a)}, d_i^{(b)}\right\} \times n^{\bullet}$	Dendograma con la mejor estructura de clústeres

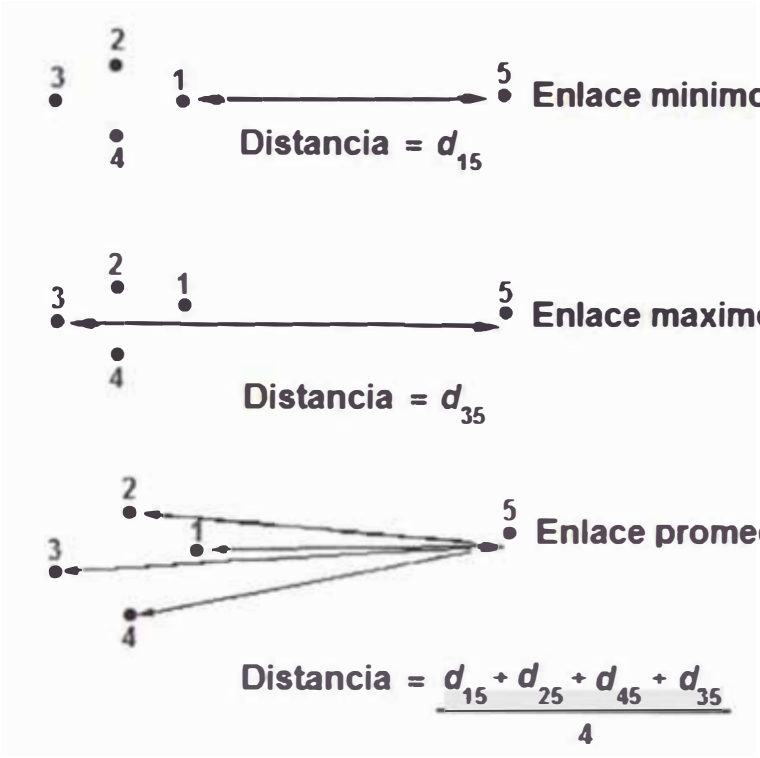

Figura N° 2-5. Tipos de enlace de unión para los individuos en la formación de un clúster.

2.5.6.2 Dendograma

El proceso de unificación de los clústeres de los objetos y variables puede ser gráficamente representado por un dendograma (Figura N° 2-6), donde el eje X representa los individuos, y el eje Y representa la distancia de agrupamiento expresado como (Dlink / Dmax)*100. Esta distancia representa la relación estandarizada entre la distancia de unión para un individuo particular divido por la distancia máxima (7; 28; 32).

De esta manera, el dendograma visualiza a todos los individuos, los pasos de agrupamiento de los clústeres y la distancia entre ellos. Sin embargo, la agrupación de los clústeres es irreversible, en el sentido de que los individuos o clústeres que son unidos en un clúster mayor no pueden ser separados luego, y de esa manera cualquier error de agrupamiento inicial no podrá ser corregido en etapas posteriores (26).

Por consiguiente, el dendograma puede ser usado para identificar manualmente un número de clústeres inherentes en la matriz de datos X^T, cortando el dendograma a un cierto valor del eje Y, esto es, a una cierta (Dlink / Dmax)*100. Por esto, el dendograma es muy útil para discutir muchos resultados posibles de los procesos de agrupamiento.

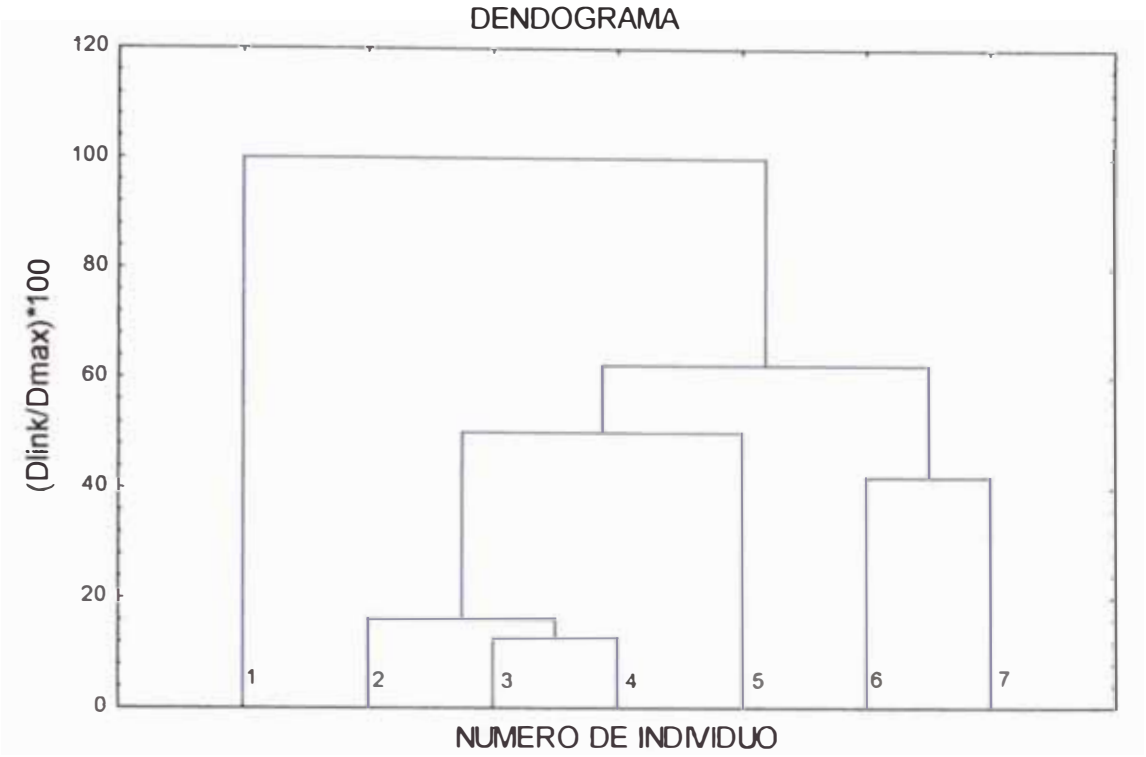

Figura N° 2-6. Dendograma del análisis de clúster jerárquico.

2.6 Análisis de componentes principales (ACP)

2.6.1 Generalidades

El análisis de componentes principales (ACP) es una técnica estadística que fue propuesta por Karl Pearson en 1901 como parte del análisis de factores y es considerado como la técnica de cambio más significante en la visión del químico en los análisis de datos (23). El ACP se basa en las correlaciones entre las variables para desarrollar a partir de estas correlaciones un grupo pequeño de variables no observables llamadas componentes principales. Esto se logra a través del cálculo de los eigenvalores y eigenvectores a partir de la matriz de correlación (32; 33, 34), por lo que el ACP forma parte de las llamadas técnicas factoriales, además de ser conocido como una técnica de reconocimiento de patrones no supervisado (21).

El objetivo principal del ACP es reducir las dimensiones de una base de datos multidimensional (muchas variables y muchos objetos observados) en un espacio reducido de pocas dimensiones. Este espacio reducido tiene como ejes principales a los componentes principales (CPs), el cual son combinaciones lineales de las variables

originales (35, 36). En la Figura N° 2-7 se visualiza los dos primeros CPs para uno grupo de datos conformado por dos variables. Estos son ejes ortogonales y en direcciones que maximizan la varianza posible o representatividad de cada CP (7; 33; 37)

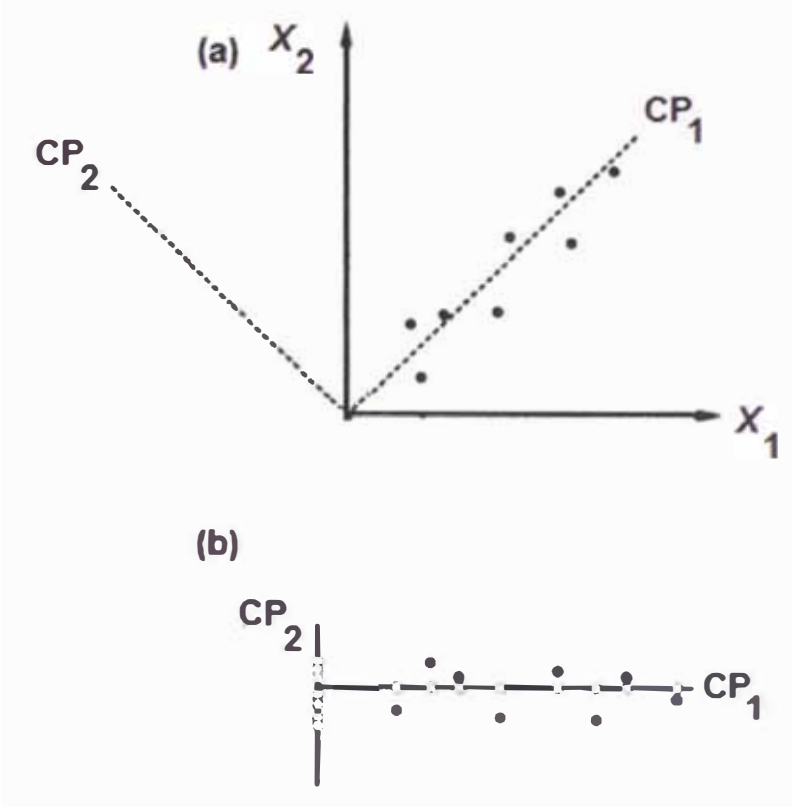

Figura N° 2-7. (a) Diagrama que ilustra las dos componentes principales CP1 y CP2 para dos variables, X_1 y X_2. (b) Puntos referidos a los ejes de las componentes principales donde los puntos en negrita indican los datos y los blancos la proyección sobre los ejes (46).

Teniendo la primera componente principal la máxima varianza posible, la segunda componente principal la máxima varianza posible restante y así sucesivamente en orden creciente. Todas las componentes representan y explican las relaciones y agrupamientos de las variables y objetos sin perder mucha de la información original de las relaciones entre estos, esto es extraer la mayor varianza posible de los datos. La reducción de la dimensión por el PCA es usado principalmente para (20):

- Visualización de datos multivariado mediante gráficos de dispersión observables.
- La transformación de variables muy correlacionadas en grupos pequeños de PCs no correlacionados que pueden ser usados por otros métodos.
- La separación de la información relevante (descrita mediante unas pocos CPs) del ruido.

El ACP es adecuado para grupos de datos con variables correlacionadas tal como sucede a menudo con los datos de origen químico. Las variables constantes o variables altamente correlacionadas no causan problemas para el PCA; sin embargo, los valores extremos o atípicos (outliers) pueden tener una fuerte influencia en los resultados, por lo cual es importante el análisis y pretratamiento de los datos antes de realizar el ACP (22).

2.6.2 Análisis inicial

El punto central del ACP es reducir y proyectar la matriz de datos autoescalados, *Z(n, m)*, de la matriz de datos transformados $X^T(n, m)$, con *n* objetos y *m* variables en un espacio de menor dimensión que tienen como ejes a los componentes principales, *c*, mediante la descomposición de la matriz *Z(n, m)* en el producto de las matrices de puntuaciones *S* y la matriz transpuesta de cargas *L* (ecuación 15) (16; 33).

$$\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & z_{ij} & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1c} \\ s_{21} & s_{22} & \cdots & s_{2c} \\ \vdots & \vdots & s_{ij} & \\ s_{n1} & s_{n2} & \cdots & s_{nc} \end{pmatrix} x \begin{pmatrix} l_{11} & l_{21} & \cdots & l_{1c} \\ l_{21} & l_{22} & \cdots & l_{2c} \\ \vdots & \vdots & l_{ij} & \vdots \\ l_{m1} & l_{2c} & \cdots & l_{mc} \end{pmatrix}^T \quad (15)$$

$$Z = S \times L^T \quad (16)$$

De esta manera, existen dos espacios de menor dimensión c, tanto para los objetos como para las variables. Graficando cada dos o tres columnas de la matriz S se obtiene los agrupamientos de los objetos en gráficos de bidimensionales o tridimensionales. De la misma manera, graficando las columnas de la matriz L los gráficos muestra las correlaciones de las variables (38).

2.6.3 Técnicas de análisis

Los algoritmos para el cálculo de las matrices de puntuación y cargas más usados para el ACP son los siguientes:

1. Mínimos cuadrados parciales iterativo no lineales (NIPALS).
2. Descomposición de valores singulares (DVS).

2.6.4 Descomposición de valores singulares (DVS)

La descomposición de valores singulares (DVS) está basado en la descomposición de la matriz *Z(n, m)* en 3 matrices (ecuación 17) (20).

$$Z = S \times \sqrt{\Lambda} \times V \quad (17)$$

Donde $\sqrt{\Lambda}$ es la matriz diagonal que contiene a las raíces cuadráticas de los autovalores de los componentes principales, V es la matriz de los autovectores correspondientes a los autovalores y S es la matriz de puntuaciones de los componentes principales.

La solución de la ecuación 17 se basa en el cálculo de la matriz de los autovectores a partir de la matriz de correlación de las variables, *R*, de la matriz *Z* (16; 39). Esto se realiza mediante la descomposición de esta matriz de correlación, *R*, en una matriz de autovectores *V* y una matriz diagonal de autovalores Λ (ecuación 18).

$$R \times V = V \times \Lambda \quad (18)$$

Sin embargo, la ecuación 18 permite ser desarrollado a partir de la solución no trivial de su determinante, denominada ecuación característica (ecuación 19).

$$|R - \Lambda I| = 0 \quad (19)$$

A partir del producto vectorial entre la matriz de autovectores *V* y la matriz diagonal de las raíces cuadráticas de los autovalores, $\sqrt{\Lambda}$, se calcula la matriz de las cargas o matriz de coordinación de las cargas de los componentes principales correspondiente a las variables, *L* (ecuación 20) (40).

$$L = V \times \sqrt{\Lambda} \quad (20)$$

Calculándose luego la matriz de puntuaciones de los componentes principales correspondientes a los objetos, *S*, a partir de la matriz de los datos autoescalados, *Z*, y la matriz de cargas *L* (ecuación 21) (39).

$$S = Z.L \quad (21)$$

Sin embargo, los datos de esta matriz no están ponderados con el peso de cada componente principal, por lo que finalmente se calcula la matriz de coordinación de las puntuaciones de los componentes principales para los objetos (22; 39) (ecuación 22).

$$S_c = S \times \sqrt{\Lambda} \tag{22}$$

Los valores de la matriz de coordinación de las cargas (L) varían en el rango de +1 y -1. Así, los componentes principales con las cargas de las variables más cercanas a los valores extremos del rango tienen una mejor explicación o mayor peso sobre estas variables. Mientras los valores más cercanos a cero son menos explicados.

Por otro lado, las variables con los valores de sus coordinaciones de las cargas, los cuales están representados como vectores en las graficas de dispersión de las coordinaciones de las cargas, más cercanos, esto es, con los vectores que tengan los ángulos más agudos entre ellos tienen a ser más semejanzas. Mientras los que tienen ángulos cercanos a 180° tienen una relación inversa.

2.6.5 Estimación del número de componentes principales (20)

El uso de todos los componentes principales después de la composición de la matriz de datos usualmente no es adecuado. Para decidir el número de componentes principales, existen algunos criterios estadísticos como: porcentaje de la varianza explicada, criterio de Káiser o del autovalor > 1), la prueba de sedimentación y la validación cruzada.

2.6.5.1 Porcentaje de la varianza explicada

La calidad de la representación de los datos puede ser evaluada por el porcentaje de la varianza explicado por los CPs (ecuación 23), que está en función de la relación de los autovalores de cada CP de la ecuación 18 y la suma de todos los autovalores (21).

$$\%\text{Varianza del } CP_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \times 100\% \tag{23}$$

Donde CP_i es el componente principal i-esimo, y n es el número de componentes principales. Por lo que un alto porcentaje de varianza de los datos se traduce en una alta

representatividad y explicación de la información de las relaciones entre los datos en un espacio de CPs.

De esta manera, el primer componente principal (CP1) tienen el mayor porcentaje de varianza posible, esto es, representa y explica la mayor cantidad de información posible, el segundo componente principal (CP2) tiene la mayor cantidad de varianza restante no explicada por el CP1, y así sucesivamente hasta cubrir la varianza total de la matriz de datos, esto es, la representación y explicación de tota la información de las relaciones de los datos.

Usualmente un porcentaje adecuado de la varianza es fijado por los primeros 3 o 4 primeros componentes principales (22; 41). Esto permite la observación de las relaciones entre los datos en gráficos visibles bidimensionales y tridimensionales.

2.6.5.2 Criterio de Kaiser o autovalor >1

Se basa en el hecho de que que indica que sobre una base de datos estandarizados (varianza igual 1 y un valor medio igual 0), el promedio de los autovalores de cada componente principal es uno.

Esto es debido a que la suma de los eigenvalores es exactamente igual al número de variables para los datos autoescalados de la matriz Z. Por lo cual cada CP debe explicar al menos una parámetro ambiental y así los componentes con valores mayores a 1 son considerados importantes u óptimos (15; 42).

2.6.5.3 Prueba de sedimentación

Este gráfico representa el número de componentes principales junto con los eigenvalores. Tal gráfico se asemeja al perfil de una montaña, pues después de una pendiente empinada aparece una región más plana llamada zona de sedimentación, (Figura N° 2-8).

El número de componentes principales a tomar en cuenta son aquellas que contienen la parte de la grafica con la pendiente más vertical, y los cuales normalmente concuerda con los componentes que tienen los autovalores mayores que uno.

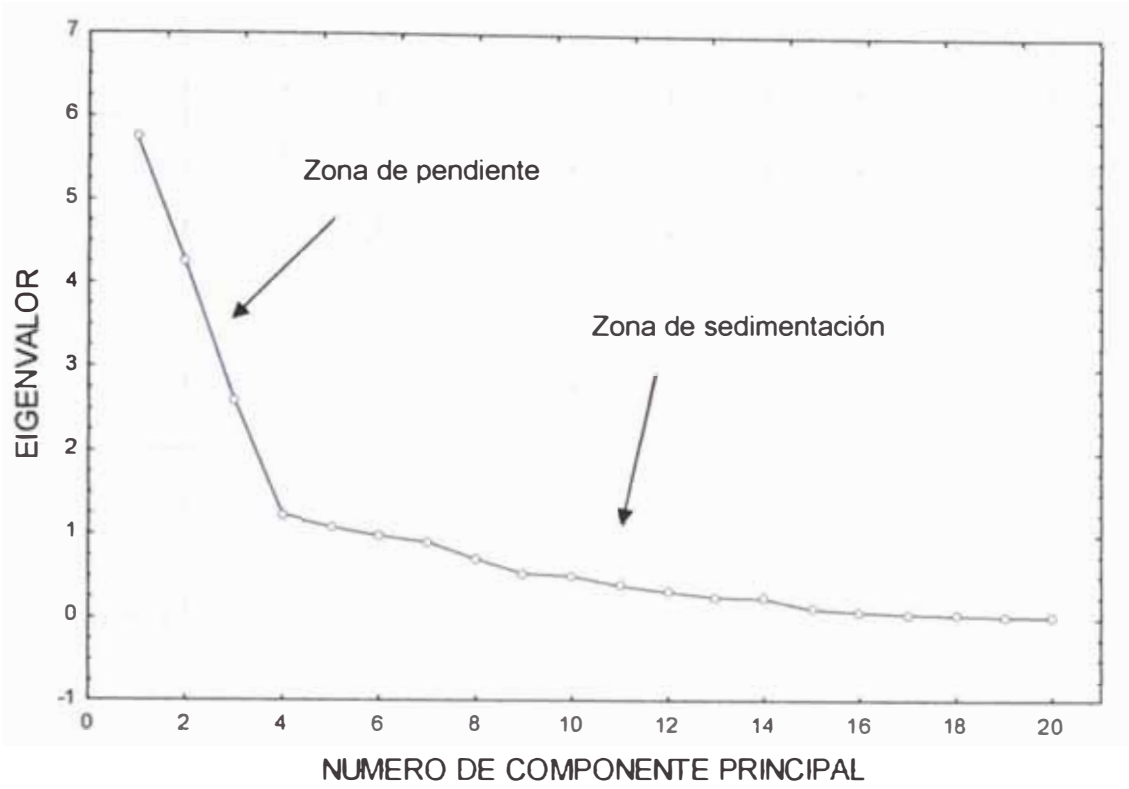

Figura N° 2-8. Gráfico de sedimentación del análisis de componentes principales.

2.7 Análisis de factor (AF)

2.7.1 Generalidades

El análisis de factor (AF) es un método que tiene características comunes con el ACP, ambos forma parte de los análisis factoriales, pues tienen como origen la simplificación de las matrices complejas de datos tomando en cuenta las relaciones inherentes que existen entre las variables, y los cuales son visualizados en su matriz de correlación (43).

El objetivo principal del análisis de factor (AF) es determinar las estructuras de las variables (método de clasificación de datos) y las influencias de cada una de estas estructuras sobre los objetos ((44). Esto se visualiza en un espacio de dimensiones reducidas que contiene como ejes a un número pequeño de características comunes latentes (no observadas) llamados factores, que usualmente explica la misma cantidad de información de las relaciones entre los objetos y las variables que la matriz de datos originales (22). Sin embargo, el AF se diferencia del ACP (método de reducción de datos), no usa toda la información o varianza de los datos de la matriz de correlación

inicial, pues el AF divide la varianza total de los datos en tres partes: varianza de los factores comunes (comunalidades), varianza de los factores específicos y los residuales o errores (16; 26). De esta manera, el AF solo utiliza la varianza de los factores comunes, esto es, la información o varianza que cada variable tiene en común con las otras variables (40).

Finalmente, los procesos de rotación (ortogonal y oblicua) sobre los valores de los factores facilita la interpretación de estos valores, visualizando una estructura más simple de los factores tanto para las variables como los objetos.

2.7.2 Análisis inicial

El punto central del análisis de factor es reducir la matriz de datos autoescalados de la matriz transformada X^T, *Z (n, m)* de *n* objetos y *m* variables en el producto de la matriz *F (n, c)* y la transpuesta de la matriz *L (m, c)* más una matriz *E (n, m)* (ecuación 24) (16; 22).

$$\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & z_{ij} & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1c} \\ f_{21} & f_{22} & \cdots & f_{2c} \\ \vdots & \vdots & f_{ij} & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nc} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & \cdots & l_{1c} \\ l_{21} & l_{22} & \cdots & l_{2c} \\ \vdots & \vdots & l_{ij} & \vdots \\ l_{m1} & l_{2c} & \cdots & l_{mc} \end{pmatrix}^T + \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & e_{ij} & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nm} \end{pmatrix} \quad (24)$$

$$Z = F \times L^T + E \quad (25)$$

Donde *F* y *L* son las matrices de puntuaciones y de cargas de los *c* factores respectivamente, *E* es la matriz de los factores específicos o los factores no explicados m-*c* (residuales o errores). Estos factores específicos no están relacionados a los factores comunes *c y* normalmente representan a la varianza única de cada variable, los interferentes aleatorios, los errores u otras fuentes de variación (20). De acuerdo al método de SVD, la ecuación 25 puede ser expresada como la ecuación 26.

$$Z - E = F \times \sqrt{\Lambda} \times V \quad (26)$$

Donde $\sqrt{\Lambda}$ es la matriz diagonal que contiene a las raíces cuadráticas de los autovalores de los componentes principales, V es la matriz de los autovectores correspondientes a los autovalores y F es la matriz de puntuaciones de los componentes principales.

2.7.3 Técnicas de análisis

El análisis de factor puede basarse tanto en la técnica de extracción de los factores mediante el ACP (técnica de los componentes principales) como en las técnicas de análisis que no extraigan toda la varianza de las variables (técnica de los factores principales); solo la proporción de la varianza debido a los factores comunes (40). De esta manera, las principales técnicas de extracción son dadas en la Figura N° 2-9.

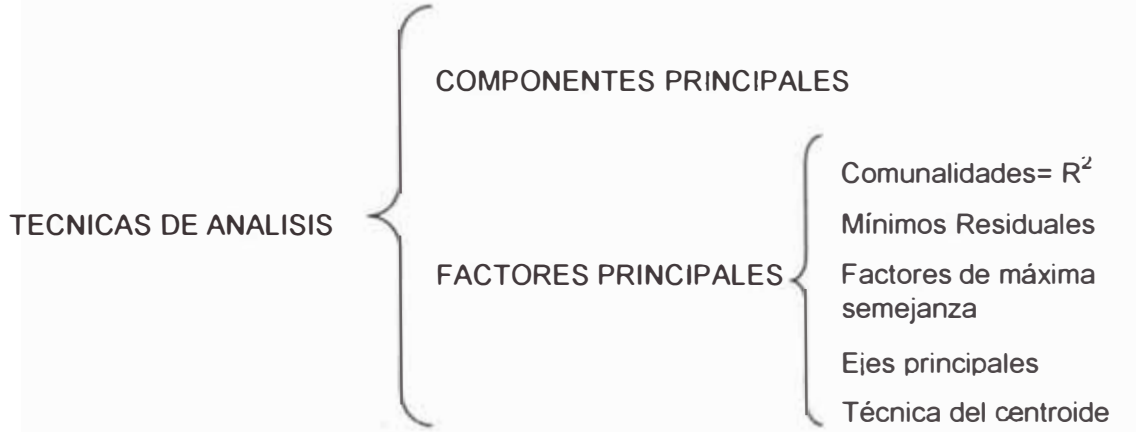

Figura N° 2-9. Clases de técnicas del Análisis de Factor.

2.7.4 Análisis de las comunalidades= R^2

El punto inicial de esta técnica de extracción de los factores principales es el cálculo de la varianza de los factores comunes a partir de la matriz de correlación reducida *R'*. Esta matriz proviene de la diferencia entre la matriz de correlación de las m variables de la matriz *Z(n, m)*, *R*, y la matriz de la varianza de los factores específicos, *K* (ecuación 27):

$$R' = \begin{pmatrix} 1 & r_{21} & \cdots & r_{m1} \\ r_{21} & 1 & \cdots & r_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{pmatrix} - \begin{pmatrix} \sigma_1^2 & \sigma_{21} & \cdots & \sigma_{m1} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} h_1^2 & r_{21}-\sigma_{21} & \cdots & r_{m1}-\sigma_{m1} \\ r_{21}-\sigma_{21} & h_2^2 & \cdots & r_{m2}-\sigma_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}-\sigma_{m1} & r_{m2}-\sigma_{m2} & \cdots & h_m^2 \end{pmatrix} \quad (27)$$

$$R' = R-K \quad (28)$$

Donde σ_j^2 y σ_{ij} son las varianzas y covarianzas de los factores específicos de la matriz K respectivamente, y corresponden a la fracción de la varianza y covarianza de los datos que no pueden ser explicadas por los factores comunes, y $h_j^2 = 1 - \sigma_j^2$ para j = 1,2,..., m,

son llamadas comunalidades o las varianzas explicadas por los factores comunes. El primero paso es el cálculo de cada columna de la matriz de cargas (L), el cual reproduce la matriz de correlación reducida R′(ecuación 28) mediante la multiplicación de estas matrices de cargas y la diferencia de las matriz de varianza de los factores específicos (ecuación 29) (20).

$$R' = L \times L^T - K \tag{29}$$

El siguiente paso es calcular columna por columna la matriz de autovalores y sus respectivos autovectores a partir de la matriz de correlación reducida mediante la ecuación 30.

$$R' \times V = V \times \Lambda \tag{30}$$

La ecuación 30 permite ser desarrollado a partir de la solución no trivial de su determinante, denominada ecuación característica (ecuación 31).

$$|R' - \Lambda I| = 0 \tag{31}$$

De esta manera, se determina el número de autovalores > 1 y por consiguiente se considera solamente a los autovectores y factores (columnas de la matriz L) que sean representados por estos autovalores.

Luego, a partir del producto vectorial entre la matriz de cargas, *L*, y la matriz de correlación total, R, se calcula la matriz de coordinación de las puntuaciones de de los factores, *B (*ecuación 32).

$$L = R \times B \tag{32}$$

Finalmente, se calcula la matriz de puntuaciones de los factores F, a partir de la matriz de los datos autoescalados *Z*, y la matriz de coeficientes de las puntuaciones *B* (ecuación 33).

$$F = Z \times B \tag{33}$$

Estas puntuaciones son importantes, pues miden la influencia de los factores comunes sobre las *m* variables.

2.7.5 Métodos de rotación

Después de la extracción de los factores, una matriz de cargas óptimas de los factores es obtenida a través de la rotación de los ejes de los factores; esto es, transformar los factores comunes abstractos en factores interpretables, el cual ayuda que la interpretación de las cargas de los factores sea más fácil.

Los métodos de rotación logran que las variables puedan ser divididas en grupos separados, en concordancia con las características comunes que tienen las variables, de acuerdo a ciertos criterios de varianza tomados de acuerdo al tipo de rotación (16). Existen dos principales tipo de rotación de ejes: la rotación ortogonal y oblicua.

2.7.5.1 Rotación ortogonal: Para una rotación ortogonal el sistema de coordenadas de los factores es rotado en un cierto ángulo (Figura N° 2-10). El objetivo es que las nuevas coordenadas corte a las clases de los objetos en una manera óptima. Así, la rotación ortogonal de la matriz de cargas L a una matriz de cargas rotadas ortogonales $L_{rotación}$ se obtiene por medio de la multiplicación de la matriz L con una matriz de transformación rotación T (20; 40) (ecuación 34).

$$L_{rotación} = L \times T \qquad (34)$$

Las rotaciones ortogonales rescatan la estructura de los factores independientes y una vez desarrollada, los ejes de la matriz de cargas de los factores rotados, $L_{Rotación}$, mantienen su ortogonalidad. Ejemplos típicos de rotaciones ortogonales son:

Rotación varimax: Maximiza la varianza de las cargas cuadráticas de todas las variables de cada factor. En otras palabras, la rotación permite incrementar la participación de las variables con la más alta contribución (maximiza los valores de las cargas más altas de las variables) y la vez reduce las variables con una menor contribución (minimiza los valores de las cargas más bajas de las variables) de cada factor (39; 40; 45).

Rotación quartimax: Maximiza la varianza de las cargas cuadráticas de cada variable a lo largo de los factores. Esto es, la rotación permite maximizar las cargas más altas y al mismo tiempo minimiza las cargas más bajas dentro de cada variable (39; 40).

Rotación ecuamax: Maximiza tanto las varianzas de las cargas cuadráticas de cada variable dentro y a lo largo de cada factor. Esta rotación puede ser considerada como una "mezcla única" entre la rotación varimax y quartimax (39; 40).

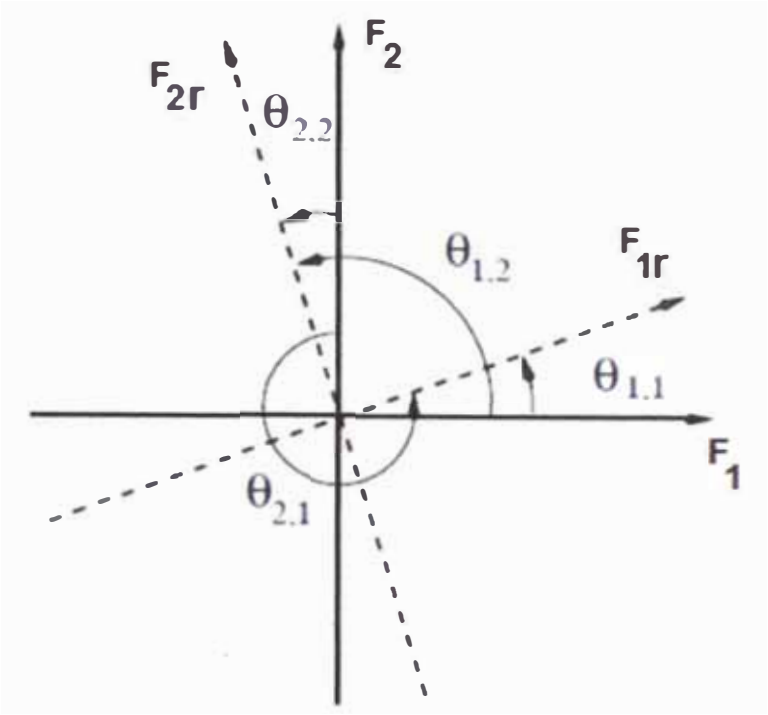

Figura N° 2-10. Rotación de los ejes de los CPs para el Análisis de Factor.

2.7.5.2 Rotación oblicua: En esta rotación los ejes de los factores no son ortogonales, por lo que estos factores pueden no estar correlacionados. Esto permite que esta rotación sea más útil conceptualmente que las rotaciones ortogonales pero la interpretaron, descripción y presentación de los resultados no es sencillo ni práctico (16; 40).

2.8 Análisis discriminante (AD)

2.8.1 Generalidades

El análisis discriminante (AD) es un grupo de métodos y herramientas estadísticas descriptivas e inferenciales que se desarrolla con un conocimiento a priori de la pertenencia de los objetos a una clase particular. El AD, de una manera parecida al análisis factorial (ACP y AF), crea nuevas variables llamadas funciones discriminantes (FD), los cuales son combinaciones lineales de las variables originales. Estas funciones

maximizan la separación de los objetos en las clases conocidas de los objetos en un espacio reducido, que tiene como ejes principales a las FD. En la Figura N° 2-11 se observa la clasificación de un grupo de individuos mediante dos funciones discriminantes (FD_1 y FD_2) y la proyecciones de estos(26).

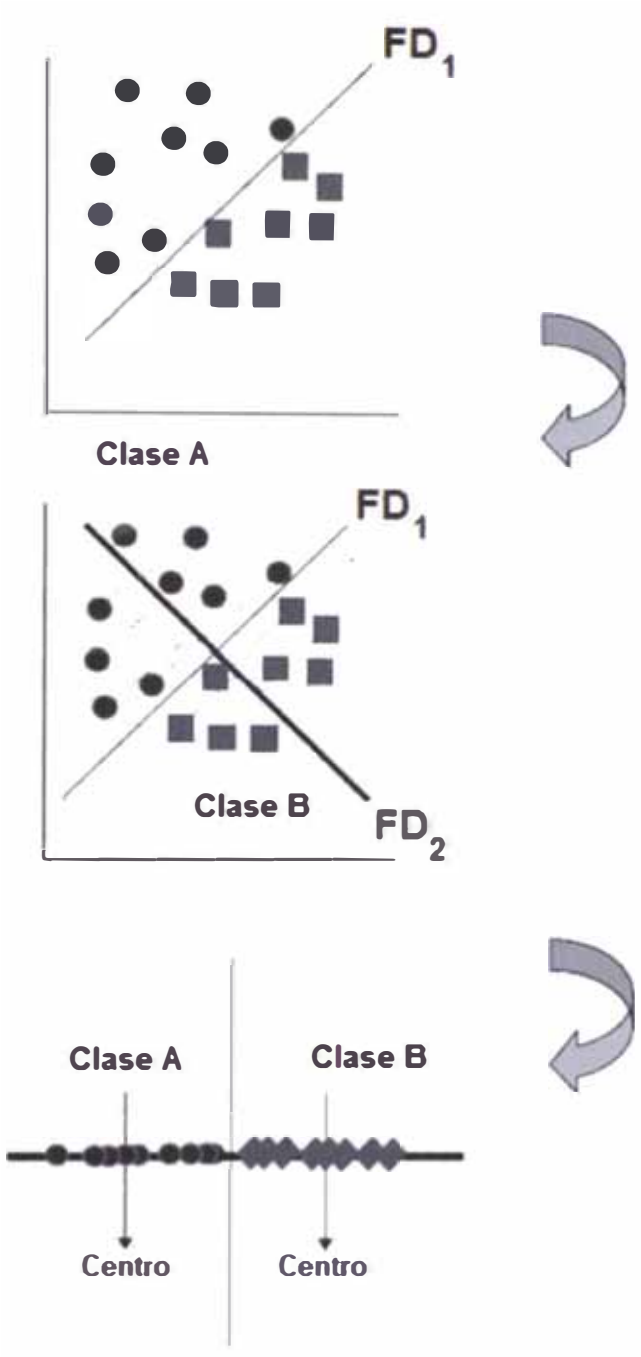

Figura N° 2-11. Clasificación en dos clases de un grupo de individuos mediante dos funciones discriminantes (FD_1 y FD_2) de los objetos y sus proyecciones sobre la FD_2 (23).

Sin embargo, para obtener el mejor procedimiento y resultado para un número de clases mayores que dos, el análisis discriminante supone que los datos deben seguir una

distribución normal multivariado, esto es, la existencia de diferentes clases de centroides de los grupos, y la semejanza de las matrices de varianzas entre las diferentes clases (24; 40; 44). Esto conlleva a que los datos iniciales deben ser transformados a variables que tengan una distribución normal aproximada de sus datos. Por otro lado, a diferencia de los otros métodos multivariado, la estandarización de los datos no tienen ningún efecto sobre el resultado del análisis discriminante (46).

Las aplicaciones del AD es determinar la correcta asignación de los objetos a las diferentes clases existentes mediante la visualización de la matriz de clasificación, determina las variables más predictivas o discriminantes de la clasificación conocida y calcular las funciones de clasificación que permiten clasificar a los nuevos objetos incluidos en cierta clase o grupo de variables (39).

2.8.2 Análisis inicial

En la matriz inicial de datos transformados se asume que los n objetos, medidos por m variables independientes, se clasifican en k diferentes clases (g_k), donde cada una de las clases consiste de $n_1, n_2, \ldots, n_k$ objetos (ecuación 35).

$$\boldsymbol{X} = \begin{bmatrix} \begin{Bmatrix} x_{11}^T & x_{12}^T & \cdots & x_{1m}^T \\ \vdots & \vdots & \cdots & \vdots \\ x_{i1}^T & x_{i2}^T & \cdots & x_{im}^T \end{Bmatrix} g_1 \\ \vdots \\ \begin{Bmatrix} x_{k1}^T & x_{k2}^T & \cdots & x_{km}^T \\ \vdots & \vdots & \cdots & \vdots \\ x_{l1}^T & x_{l2}^T & \cdots & x_{lm}^T \end{Bmatrix} g_2 \\ \vdots \\ \begin{Bmatrix} x_{j1}^T & x_{j2}^T & \cdots & x_{jm}^T \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1}^T & x_{n2}^T & \cdots & x_{nm}^T \end{Bmatrix} g_k \end{bmatrix} \tag{35}$$

El análisis discriminante para una cantidad de clases mayores a 2 se basa en el análisis de varianza multivariada (MANOVA) de una sola vía (25). Este análisis determina las matrices de varianzas o suma de los cuadrados entre y dentro de las clases (ecuaciones 36 y 37 respectivamente) para el análisis de las clases de los objetos (16).

$$B = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{36}$$

$$W = \sum_{i=1}^{K} \sum_{n=1}^{n_k} (x_{in}-x_i)(x_{in}-x_i)^T \quad (37)$$

Donde *k* es el número de clases, n_k es el número de objetos en la clase *k*, *x* es el vector de las medias totales, x_i es el vector de las medias la clase *i*, y x_{in} es el vector objeto de la clase *i* en la posición *n-ésimo*.

La matriz B describe la dispersión o varianza entre las medias de las clases sobre la media total y la matriz W describe la dispersión o varianza de cada objeto dentro de cada clase con su respectiva media. Estas dos matrices son el punto inicial para el análisis discriminante.

2.8.3 Funciones discriminantes

El procedimiento matemático para encontrar las funciones discriminantes se basa en las funciones discriminantes lineales de Fisher, el cual se enfoca en calcular los autovectores y sus correspondientes autovalores de la matriz BxW^{-1} (ecuación 38), tal como se conoce del procedimiento de las técnicas factoriales (ACP y AF) (16; 20).

$$\begin{aligned} (BxW^{-1}).e_1 &= \lambda_1.e_1 \\ (BxW^{-1}).e_2 &= \lambda_2.e_2 \\ &\vdots \\ (BxW^{-1}).e_q &= \lambda_q.e_q \end{aligned} \quad (38)$$

La ecuación 34 permite ser desarrollado a partir de la solución no trivial de su determinante, denominada también ecuación característica (ecuación 39).

$$|BxW^{-1}-\Lambda I|=0 \quad (39)$$

La solución resulta en *q* autovalores (λ_i) y los *q* autovectores (e_q) no correlacionados (ortogonales) entre sí correspondientes a sus respectivos autovalores, los cuales expresan la parte de la varianza extraída de la matriz $B.W^{-1}$.

De esta manera, el primer autovalor, λ_1, representa la parte máxima de la varianza total extraída por el primer autovector. El segundo autovalor, λ_2, representa la parte máxima de la varianza total restante extraída por el segundo autovector, y así sucesivamente en

orden descendente. La suma de todos los autovalores representa la varianza total de la matriz $B.W^{-1}$ (16; 46).

Finalmente, las funciones discriminantes (ecuación 40) son calculadas a partir de la suma de los productos de los coeficientes estandarizados de los autovectores y las variables transformada de los objetos (16; 40).

$$F_i = e_{i1}Y_1 + e_{i2}Y_2 + \ldots + e_{im}Y_m \qquad (40)$$

Donde *i* es la i-esimo función discriminante y *m* es el número de variables independientes.

La primera función discriminante refleja la mejor separación posible entre los grupos, la segunda, ortogonal a la primera, refleja la mejor separación entre los grupos sobre la base de asociaciones no usada en la primera función discriminante. Este procedimiento continúa hasta que todas las separaciones entre los grupos sean evaluadas (16; 40). El número máximo de autovalores con respecto a la clasificación de los objetos en un cierto número de clases *k* es calculado por la ecuación 41.

$$q = \min(K-1, m) \qquad (41)$$

Donde: *k* es el número de clases de los objetos existentes, *m* el número de las variables independientes.

Por otro lado, las funciones discriminantes de manera similar a las ecuaciones de regresión, calcula las puntuaciones de cada objeto original (coordenadas de los objetos en el nuevo espacio de las funciones discriminantes) a partir de la ecuación 40 (26).

2.8.4 Prueba F multivariada

Una comparación entre las medias multivariadas o centroides de las clases es realizada mediante la prueba F multivariada. Esta prueba permite determinar el poder discriminatorio de las variables y la significancias de las funciones discriminantes canónicas. La prueba tiene como hipótesis nula:

(H_o): *las medias multivariantes (centroides) de los grupos son iguales*. Para esto, la prueba F es calculada a partir del producto de un factor dependiente de los grados de libertad de B y W, y una función que caracteriza la relación de las dos matrices B y W (16) (ecuación 42).

$$F=\frac{Df_2}{Df_1}.Spur(B.W^{-1}) \quad (42)$$

Donde Df_2 son los grados de libertad de la matriz B, Df_2 son los grados de libertad de matriz H y Spur es la función que caracteriza la relación de las matrices B y W.

Existen muchos estadísticos multivariados para caracterizar el Spur de la prueba F multivariada, y el cual permite contrastar la hipótesis nula (H_o). Los estadísticos más comunes son: el criterio de la traza de Pillai-Bartlett, el criterio de la traza de Hotelling-Lawley, la lambda de Wilks, y la raíz más grande de Roy ((25; 40 . Para muestras grandes los tres primeros estadísticos conllevan al mismo p-valor (25).

Sin embargo, el más utilizado es la lambda de Wilks, el cual compara las matrices W y B en forma similar al F estadístico univariado y puede ser expresado en términos de los eigenvalores de la matriz BxW^{-1} (16) (ecuación 43).

$$\lambda_{WILKS}=\frac{|W|}{|B+W|}=\prod_{i=1}^{n_F}\frac{1}{1+\lambda_i} \quad (43)$$

Donde n_F es el número de funciones discriminantes y λ son los eigenvalores de las funciones discriminantes.

El estadístico λ es reducido a un escalar mediante el uso de las determinantes. Así la información en las matrices W y B sobre la separación de los vectores de las medias de cada clase, x_1, x_2,..., x_k es cambiada a un escalar, el cual puede determinar si la separación de los vectores de las medias es significante (26).

Debido al uso de las determinantes, la lambda de Wilks es llamado el criterio de la determinante, y toma valores entre 0 a 1 de forma que, cuanto más cerca este de 0, mayor es el poder discriminante de las variables consideradas y los vectores de las

medias de las clases son mas separadas comparadas a la variación dentro de las clases y cuanto más cerca de 1, menor es dicho poder (20; 26). Finalmente la prueba F multivariada se expresa según la ecuación 44 (26; 40).

$$F_{(Df_1,Df_2)} = \left(\frac{Df_2}{Df_1}\right) \times \left(\frac{1-\lambda^{1/t}}{\lambda^{1/t}}\right) \quad (44)$$

Donde:

$Df_1 = m \times (k-1)$, $Df_2 = w \times t - \frac{1}{2}(DF_1-2)$

$$t = \sqrt{\frac{DF_1^2-4}{m^2+(k-1)^2-5}}, \quad w = n-1-\frac{1}{2}(m+k)$$

2.8.5 Prueba de significancia del χ2

Cuando se evalúa la significancia de las funciones discriminantes canónicas, F_i, se puede utilizar la prueba de significancia, χ^2 (chi cuadrado), cuya hipótesis H_o: *las medias de las funciones discriminantes canónicas en cada clase son iguales*. De esta manera, la distribución del χ^2 está relacionada a la λ de Wilks mediante la ecuación 45 (26)

$$\chi_i^2 = -\left[n-1-\left(\frac{m+k}{2}\right)\right] \times \ln(\lambda_i),\ i = 1, \ldots, K-1 \quad (45)$$

Donde n es el número de objetos, m es el número de variables y k es el número de clases.

Cada valor de la aproximación de la distribución χ_i^2, con (m-i+1) (k-i) grados de libertad, está relacionado con cada función discriminante cuadrática (F_i), rechazando la hipótesis nula (H_o) si el p-valor asociada al valor del χ_i^2 es menor que 0.05 (nivel de confianza del 95%). Por lo que las funciones discriminantes significantes permiten visualizar la mejor separación entre las clases (26).

2.8.6 Reducción de variables

Las funciones discriminantes canónicas igual que los componentes principales y factores del ACP y AF respectivamente, son combinaciones lineales de todas las variables de la

matriz de datos inicial. Por esta razón, no hay una reducción real de las variables desde el punto de vista práctico (experimental) (16).

Sin embargo, el análisis discriminante desarrolla técnicas de reducción de las variables, eliminando de esta manera las variables que llevan información redundante, i.e., variables altamente correlacionadas entre ellos. Este proceso de eliminación encuentra el grupo de variables más discriminantes u óptimas con un poder de discriminación estadísticamente suficiente y un teniendo un riesgo de error aceptable.

Para llevar a cabo este proceso de eliminación, se calcula el valor del estadístico F parcial de cada variable, el cual muestra la significancia de cada una de las variables después de ajustarse a otras variables, esto es, la separación o discriminación en relación a las otras variables de las clases de los objetos (26).

Este valor F parcial se calcula a partir del llamado lambda parcial de Wilks, el cual permite evaluar los procesos de selección de la variables para el cálculo de un óptimo grupo de variables.

Esta Lambda parcial de Wilks es definida como el incremento multiplicativo en la lambda de Wilks que resulta de agregar o eliminar la variable respectiva (ecuación 46) (16; 39).

$$\lambda_{\text{parcial de la variable}} = \frac{\lambda_{\text{antes de agregar o eliminar la variable}}}{\lambda_{\text{después de agregar o eliminar la variable}}} \qquad (46)$$

Finalmente, el correspondiente valor del F estadístico parcial de cada variable se calcula a partir de la ecuación 47 (16; 40).

$$F_{(k-1,n-k-m)} = \left(\frac{n-K-m}{K-1}\right) x \left(\frac{1-\lambda_{\text{parcial de la variable}}}{\lambda_{\text{parcial de la variable}}}\right) \qquad (47)$$

Donde n es el número de objetos, K es el número de grupos y m es el número de variables independientes.

El valor F parcial para una variable indica su significancia estadística en la discriminación entre grupos, esto es, es una medida de hasta qué punto una variable tiene una

contribución única a la predicción de la pertenencia de la clase. Existen dos tipos de valor F parcial:

F de entrada: Expresa la disminución que se produce en la λ de Wilks si se incluyen una variable dada entre las que no están dentro de la función discriminante.

F de salida: Expresa el incremento que se produce en la λ de Wilks si se elimina de la función discriminante una variable dada.

2.8.7 Modos de reducción de variables

2.8.7.1 Modo estándar

El método estándar o directo calcula los respectivos valores de F de salida de cada variable. Los valores F de salida más altos y con un p-valor < 0.05 (significancia del 95%) corresponden a las variables más discriminantes o que mejor separan las clases de los objetos.

2.8.7.2 Modo stepwise (paso a paso)

Existen dos modos:

Modo de forward stepwise (paso a paso hacia delante): comienza con la variable más discriminante, esto es, la que maximiza la separación entre las clases y por consiguiente tenga el valor F estadístico de entrada más alta. Luego, en cada etapa, se incluye las variables con el siguiente valor F de entrada más alta, hasta aquel valor en el cual la disminución de la λ de Wilks sea inapreciable y la variable no entre al modelo (26; 39).

Modo de backward stepwise (paso a paso hacia atrás): comienza con todas las variables y calcula el F estadístico parcial de cada variable que resulta de eliminar la variable correspondiente en presencia de las otras variables (F de salida). Luego, en cada etapa, se elimina la variable menos discriminante, esto es, la que en menor grado maximiza la separación entre las clases y por consiguiente tenga el valor del F de salida más pequeño. Esto continúa hasta que el incremento de la λ de Wilks no sea significativo y la variable se elimine del análisis ((26; 39).

2.8.7 Clasificación bayesiana

Uno de los otros objetivos del análisis discriminante es evaluar correctamente la clasificación de cada objeto en las diferentes clases. Esta evaluación es visualizada y resumida en una matriz de clasificación, llamada también matriz de confusión (46).

La matriz de clasificación se basa en minimizar las distancias de Mahalanobis entre los objetos dentro de los grupos y su respectiva media, asumiendo que hay un igual porcentaje de pertenencia de los objetos a cada clase (23; 25). De esta manera, a distancias de Mahalanobis más pequeñas con respecto a la media, la pertenencia de un objeto a una clase es más probable (20; 23).

Sin embargo, si los objetos de todas las clases obedecen a una distribución normal multivariada, las matrices de clasificación pueden ser modificadas a partir de la función de aproximación bayesiana o teorema de Bayes (20; 22). Este teorema mejora las asignaciones de los objetos a los grupos obteniendo de esta manera una mejor clasificación.

El principio de esta nueva clasificación es que cada clase tiene una probabilidad predefinida (priori). Así, la asignación de una objeto, *x*, caracterizada por *m* características a una clase *j* de todas las clases *k* está basada en maximizar la probabilidad posterior P(j|x) mediante el teorema de Bayes (ecuación 48) (20).

$$P(j|x) = \frac{p(x|j)P(j)}{p(x)} \qquad (48)$$

La probabilidad posterior es calculada a partir de la función de densidad de probabilidad para la clase considerada p(x|j), la probabilidad de cada clase *P(j)* y la función de la densidad de probabilidad sobre todas las clases p(x). Por lo tanto, un objeto es clasificado o asignado a una clase j, cuando tenga la mayor probabilidad posterior.

PROCEDIMIENTO

EXPERIMENTAL

3.1 Reactivos y estándares

Todas las soluciones de los reactivos, de grado analítico, fueron preparadas en agua destilada/desionizada. Mientras que los estándares para los análisis de Alcalinidad, Dureza Total y Dureza Cálcica, fueron preparadas a partir de las sales secadas en la estufa a 105 °C correspondientes y lo estándares para los análisis de nitratos (NO_3-N), nitritos (NO_2-N), sulfatos (SO_4), fosfatos (PO_4-P), cloruros (Cl), y los metales Fe, Pb, Zn, Cr, Al, Cd fueron preparados a partir de los soluciones patrón trazables a SRM de NIST (MERCK). Estos materiales nos permiten asegurar la trazabilidad de las muestras.

3.2 Equipos de análisis

Los siguientes equipos fueron utilizados en los análisis químicos:

- Espectrofotómetro de Absorción Atómica Perkin Elmer modelo AA200.
- Espectrofotómetro Visible Shimadzu UV-1201V.
- pH metro Orión modelo 420A.
- Multiparámetro HANNA, modelo HI98129.
- Agitador Corning Stirrer/ hot plate modelo - Pc 620.

3.3 Estaciones de muestreo

En este estudio, las estaciones de muestreo fueron elegidos de modo que sean sitios distribuidos geográficamente en toda la cuenca del río Rímac y representativos de las características de las aguas que se encuentran tanto en el río Santa Eulalia y el alto Rímac antes de la unión de estos. De esta manera, se tomaron 7 estaciones de muestreo para la colección del agua de río (Tabla Nº 3-1), las cuales son sitios representativos de los diferentes tipos de áreas, tomando en cuenta algunos factores de influencia relacionados a las actividades antropogénicas y el factor geográfico. (Tabla Nº 3-2).

En el Anexo 3, se muestran la galería de fotos de las variaciones del cauce del río Rímac, tanto en época seca (Mayo – Noviembre) como lluviosa (Diciembre – Abril), mientras el Anexo 4 muestra los mapas de cada una de áreas de influencia en las

estaciones de muestreo, así la Santa Eulalia y Ricardo Palma están localizadas en la provincia de Huarochirí, las estaciones Los Ángeles, Huachipa y Huampaní están a las afueras de la ciudad de Lima, la estación Del Ejercito se encuentra entre los distritos de Rímac y cercado de Lima y la estación Gambeta, más cercana al mar, en el distrito del Callao.

Tabla Nº 3-1. Estaciones de muestreo del río Rímac

ESTACION DE MUESTREO	COORDENADAS UTM	ALTURA (msnm)
Ricardo Palma	N8681435.66; E318965.61; 18L	954
Santa Eulalia	N8681580.95; E318454.71; 18L	941
Los Angeles	N8676414.35; E309913.10; 18L	703
Huampaní	N8675671.95; E306868.55; 18L	641
Huachipa	N8671116.27; E293090.84; 18L	380
Del Ejercito	N8668310.40; E277563.91; 18L	133
Gambeta	N8668496.28; E268490.46; 18L	17

Tabla N° 3-2. Características físicas y urbanas de las estaciones de muestreo del río Rímac

ESTACION DE MUESTREO	CARACTERISTICAS FISICAS
Ricardo Palma	Ubicada a 10 metros aguas abajo del Puente Ricardo Palma – Distrito de Lurigancho, con riberas arenosas y de regular vegetación, de aguas con presencia de efluentes y residuos sólidos domésticos. El ancho del cauce varia de 3 metros en época seca a 20 metros en época lluviosa.
Santa Eulalia	Ubicada a 5 metros aguas abajo del Puente Santa Eulalia – Distrito de Lurigancho, con riberas de piedras medianas, de aguas con presencia de efluentes y residuos sólidos domésticos. El Ancho del cauce varía de 10 metros en época seca a 15 metros en época lluviosa.
Los Ángeles	Ubicada a 50 metros aguas abajo del Puente Los Angeles – Distrito de Chaclacayo, con riberas arcillosas de abundante vegetación, y con presencia de residuos domésticos. El ancho del cauce varia de 5 metros en época seca a 20 metros en época lluviosa.
Huampaní	Ubicada a 5 metros aguas abajo del puente Huampaní – Distrito de Chaclacayo, con riberas arenosas y presencia de piedras grandes, de poca vegetación, con poca presencia de residuos domésticos y nula industrial cercanas. El ancho del cauce varia de 10 metros en época seca a 30 metros en época lluviosa.
Huachipa	Ubicada a 10 metros aguas abajo del Puente Huachipa – Distrito de Ate-Vitarte, con riberas piedrosas y de extensa vegetación, y con presencia de residuos domésticos e industriales. El ancho del cauce varía de brazos de 5 a 10 metros en época seca a 50 metros en época lluviosa.
Del Ejercito	Ubicada a 20 metros aguas arriba del puente Del Ejército – Distrito de Cercado de Lima, con riberas limosas y piedras medianas, de nula vegetación y de aguas negras con altas cantidades de residuos domésticos e industriales. El ancho del cauce varia de 5 metros en época seca a 50 metros en época lluviosa.
Gambeta	Ubicada a 20 m aguas arriba del puente Gambeta – Distrito del Callao, con riberas arenosas y piedrosas, de nula vegetación y de aguas negras con altas cantidades de residuos domésticos e industriales. El ancho del cauce varia de 2 metros en época seca a 55 metros en época lluviosa.

3.4 Metodología de monitoreo

Un total de 252 muestras del agua del río Rímac fueron colectadas durante el período de 1 año a partir de Julio 2008 a Junio 2009, el cual abarcó estaciones del año de gran variabilidad climática.

El programa de monitoreo consistió en colectar 3 muestras de agua en cada una de las 7 estaciones de monitoreo para los diferentes análisis químicos. Este programa se llevó a cabo en un día de la semana (días entre la tercera y cuarta semana del mes), comenzando en horas de la mañana (hora promedio de 9:00am) y terminando finalizada la tarde (hora promedio de 6:00 pm).

El procedimiento siguiente de monitoreo se estableció de acuerdo al criterio de las normas estándares de APHA-AWWA-WEF (17).

La toma de las muestras se realizó a la mitad de la profundidad del rio y a la mitad de la corriente, en meses de caudales bajos (Mayo-Noviembre), sin embargo, por seguridad las muestras en meses de caudales altos (Diciembre-Abril) se tomaron cerca a la orilla. Estas muestras colectadas se guardaron en botellas nuevas de polietileno de alta densidad (HPDE) de 1 litro previamente lavadas tres veces por la misma muestra.

En el caso del almacenamiento de las muestras para los análisis de los cationes metálicos, estas se preservaron acidificando a pH < 2 las muestras, por medio de la adición de 20 gotas de acido nítrico concentrado en uno de los 3 envase de 1 litro. Luego, las botellas fueron almacenadas y transportadas al laboratorio en coolers a bajas temperaturas (alrededor de 4 °C) sin alcanzar la solidificación de la muestra.

En el laboratorio, las muestras fueron guardadas a 4 °C hasta el respectivo análisis de cada parámetro ambiental de acuerdo al tiempo de almacenamiento recomendado por los Métodos Estándares APHA-AWWA-WEF (Anexo 5), con las respectivas cadenas de custodia para cada mes de muestreo. El Anexo 6 indica la cadena de custodia del monitoreo en Enero del 2009.

3.5 Diagramas de Procedimiento

La Figura Nº 3-1 presenta los diagramas del análisis ambiental y quimiométrico del agua del río Rímac, los cuales representan dos trabajos diferentes pero unidos para la evaluación multivariado de la calidad de los resultados del agua. De esta manera, La primera parte involucra todo el proceso de monitoreo y análisis de las muestras y la segunda parte involucra el análisis quimiométrico de los resultados del análisis del agua.

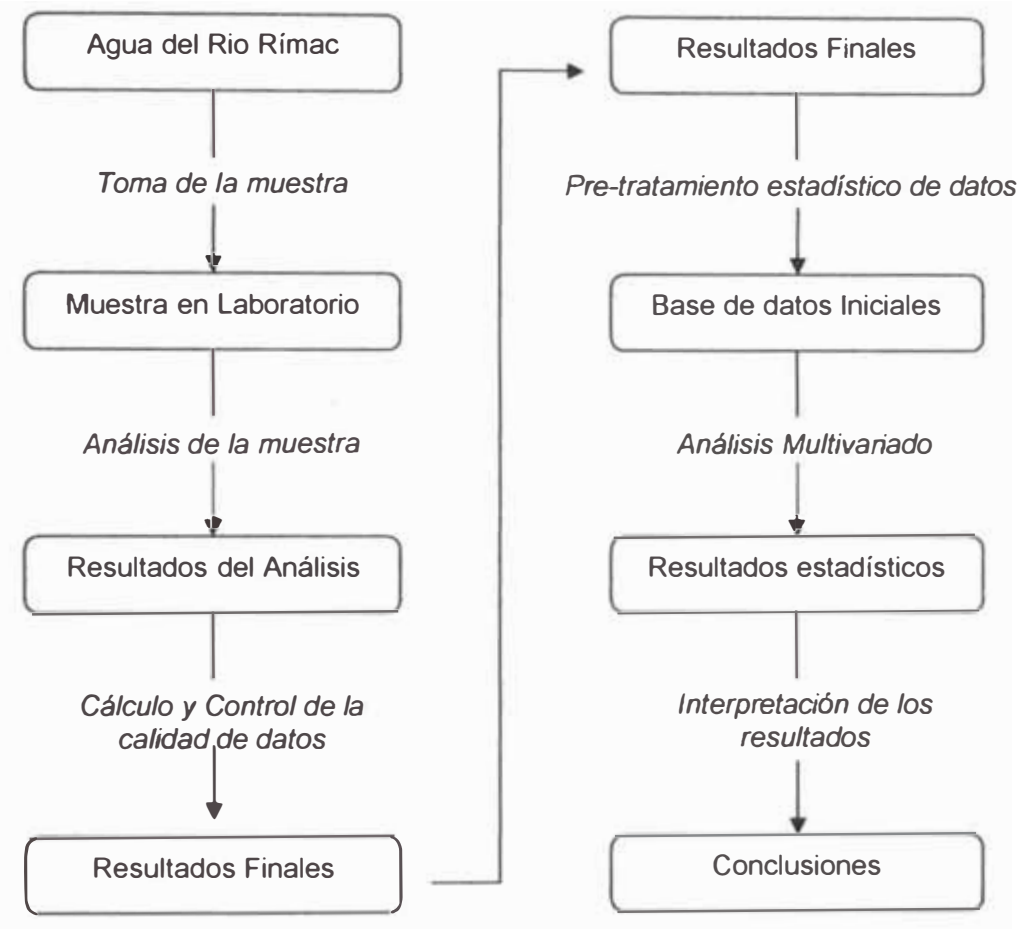

Figura Nº 3-1. Diagramas de procedimiento experimental del análisis ambiental y quimiométrico.

3.6 Métodos de análisis y parámetros

La mayoría de los parámetros analizados fueron elegidos en base a una lista de parámetros físicos y químicos prioritarios mantenidos en los programas anuales de monitoreo de Sedapal y Digesa en las aguas del río Rímac (49; 50). De esta manera, la lista de parámetros (Tabla Nº 3-3), excepto los parámetros in Situ T, pH, CE, se analizó en el laboratorio de Análisis Instrumental y Ambiental de la Facultad de Ciencias de la

Universidad Nacional de Ingeniería de acuerdo a los métodos analíticos basados en los métodos normalizados del APHA-AWWA-WEF (Anexo 8) (17).

Tabla N° 3-3. Parámetros analizados del agua del río Rímac

PARAMETROS	ABREVIACION	UNIDADES	METODO ANALITICO
Temperatura	T	°C	Multiparámetro Hanna (*in situ*)
pH	pH	Unid. de pH	Multiparámetro Hanna (*in situ*)
Conductividad eléctrica	CE	uS/cm	Multiparámetro Hanna (*in situ*)
Sólidos suspendidos totales	SST	mg / L	Secado a 103 – 105 °C
Sólidos totales	ST	mg / L	Secado a 103 – 105 °C
Nitrógeno del nitrato	NO_3-N	mg / L	Reducción por cadmio
Nitrógeno del nitrito	NO_2-N	mg / L	Colorimétrico
Sulfato	SO_4	mg / L	Turbidimétrico
Fosforo del fosfato	PO_4-P	mg / L	Acido ascórbico
Cloruro	Cl	mg / L	Argentométrico
Alcalinidad total	Alc	mg / L	Titulación
Dureza total	DT	mg / L	Titulación con EDTA
Dureza Cálcica	DCa	mg / L	Titulación con EDTA
Hierro total	Fe	mg / L	Espectrometría de absorción atómica
Plomo total	Pb	mg / L	Espectrometría de absorción atómica
Zinc total	Zn	mg / L	Espectrometría de absorción atómica
Cobre total	Cu	mg / L	Espectrometría de absorción atómica
Cromo total	Cr	mg / L	Espectrometría de absorción atómica
Aluminio total	Al	mg / L	Espectrometría de absorción atómica
Cadmio Total	Cd	mg / L	Espectrometría de absorción atómica

RESULTADOS Y

DISCUSIONES

4.1 Introducción

4.1.1 Area de estudio

4.2 Análisis y pretratamiento de datos

4.2.1 Análisis inicial de datos

4.2.2 Pretratamiento de datos

4.3 Análisis multivariado

4.3.1 Análisis de clúster (AC)

4.3.2 Análisis de componentes principales(ACP)

4.3.3 Análisis de factor (AF)

4.3.4 Análisis discriminantes (AD)

4.1 Introducción

4.1.1 Area de estudio (47; 48)

El río Rímac, uno de los ríos más importantes del Perú, de una longitud de 140 km, inicia su recorrido en la vertiente occidental de la cordillera de los Andes a una altitud de aproximadamente 5,508 m.s.n.m en el Nevado Paca, recorriendo las provincias de Lima y Huarochirí con dirección de noreste (NE) a suroeste (SE). Los principales tributarios del río Rímac son los ríos Blanco, Aruri, Santa Eulalia y las quebradas Huaycoloro y Seca.

La cuenca del río Rímac (Figura N° 4-1) limita al NE con la cuenca del río Mantaro, al SE con la cuenca del río Lurín, por el Noroeste (NW) con la cuenca del río Chillón y por el Suroeste (SW) con el Océano Pacífico. Esta cuenca tiene una extensión aproximada de 3,312 km^2, separada en la cuenca alta con una extensión de 2,237.2 km^2, desde las estribaciones occidentales de la cordillera de los Andes hasta Chosica, donde caen precipitaciones significativas, y en la cuenca baja con una extensión de 895.2 km^2, considerada a partir de Chosica, incluyendo la cuenca de la quebrada de Jicamarca (unión de las quebradas Huaycoloro y Seca), hasta la desembocadura del río en el Océano Pacífico, donde sólo esporádicamente ocurren precipitaciones.

La característica geomorfológica de la cuenca está determinada por la presencia de un valle juvenil, con una sección transversal estrecha y de relieve muy agreste, donde las marcadas variaciones de pendiente se relacionan con los cambios en las condiciones geológicas y tectónicas. Estas condiciones generan una morfología muy dinámica que se va modificando rápidamente, sobre todo a lo largo del curso principal y en el cauce de los torrentes activos que afluyen en la zona media y baja de la cuenca (ríos tributarios tales como Santa Eulalia, San Mateo y quebradas que se activan en los meses de verano en la costa de Lima, tal como la quebrada de Huaycoloro).

El Caudal del río Rímac proviene del escurrimiento natural originado por las precipitaciones sobre la sierra central del Perú, el deshielo de los nevados y los caudales liberados de las lagunas.

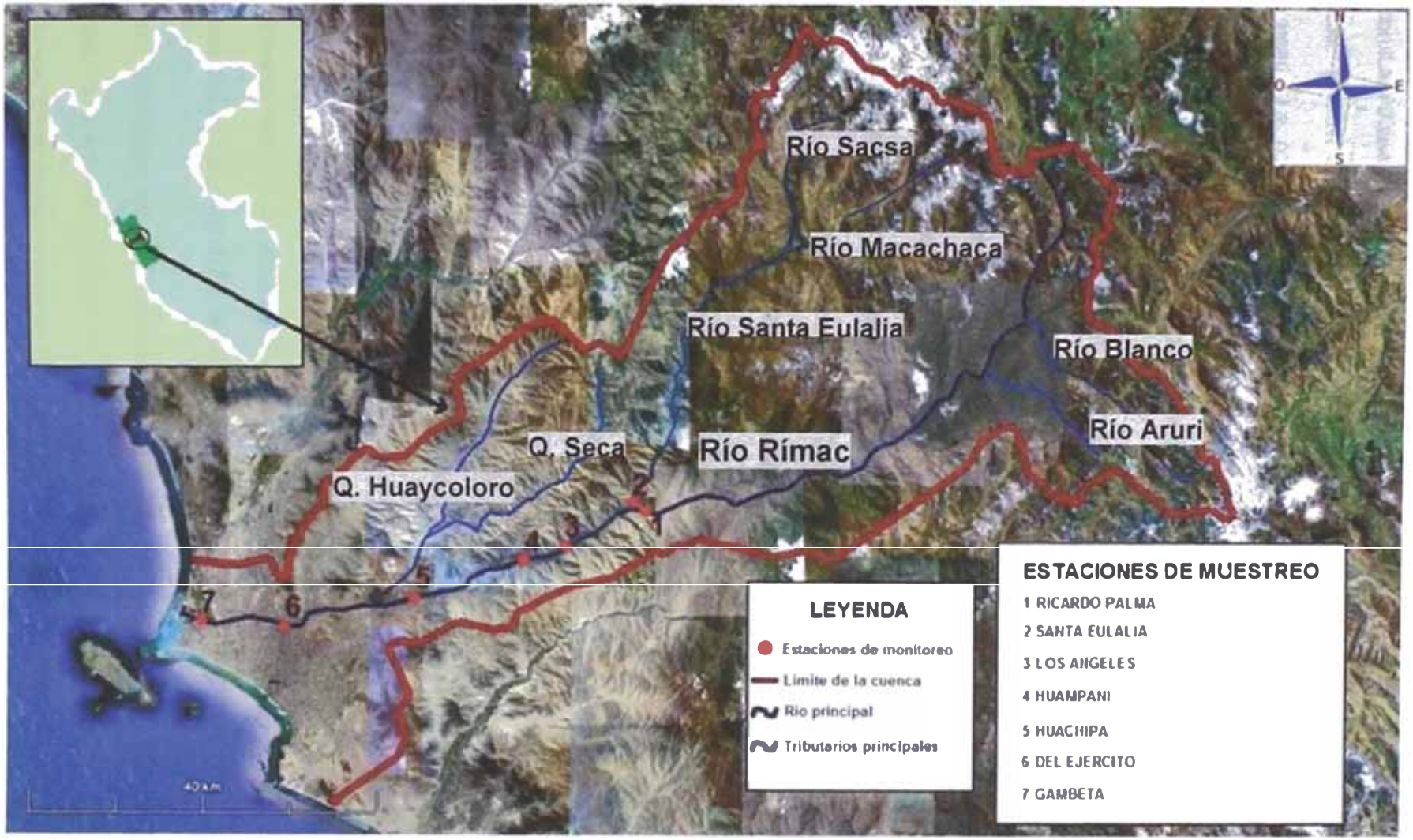

Figura N° 4-1 . Mapa de la cuenca del rio Rímac y sus principales afluentes. *Fuente : Google Earth.*

La importancia del río Rímac se debe principalmente a que es una fuente para el consumo diario de agua potable y de producción de energía para la población de Lima. Estas fuentes de consumo son dadas por la planta de tratamiento de agua para Lima, llamada La Atarjea (manejada por la empresa estatal SEDAPAL), y las centrales hidroeléctricas de Huampaní, Matucana, Huinco, Barbablanca, y Juan Carosio (también conocida como Moyopampa) respectivamente.

Por otro lado, el río Rímac soporta un amplio rango de actividades antropogénicas, entre las cuales se destacan:

Las actividades mineras, las cuales es particularmente intensa en las zonas más altas de la cuenca húmeda, tanto en la parte principal del río Rímac como en la sub cuenca del río Blanco, siendo los yacimientos polimetálicos más importantes los de Casapalca, Tamboraque, Millontingo, Pacococha, Colqui, Venturosa, Caridad, Lichicocha y Cocachacra, ubicadas en la provincia de Huarochirí.

Las actividades industriales, las cuales son particularmente intensas en las llamadas zonas industriales de los distritos de la provincia del Callao y el Cercado de Lima, y que junto con las actividades domésticas o urbanas descargan las aguas residuales en varias zonas del río tales como la margen izquierda del río Rímac (Cercado de Lima). Adicionalmente, otro problema que tienen estas actividades es el drenaje que produce los contaminantes sólidos ubicados en diferentes botaderos ubicados en las riberas de los ríos.

Las actividades agrícolas, las cuales son intensas en las zonas del alto Rímac pertenecientes a la provincia de Huarochirí y en la quebrada de Jicamarca y en menor intensidad en las zonas de las afuera de Lima y el distrito del Callao.

Otras actividades son las recreacionales establecidas en las márgenes del río Rímac, mayormente en sitios ubicados en las afuera de la ciudad de Lima, en los distritos de Chosica, Santa Eulalia y Chaclacayo, donde se ubican restaurantes turísticos, centros vacacionales y de recreación, que sirven como puntos de escape para los limeños en los periodos de invierno.

4.2 Análisis y pretratamiento de datos

Todos los cálculos realizados y los resultados dispuestos en tablas y gráficos en cada análisis estadístico se realizó con el programa estadístico Statistica 8.0 (Statsoft. 2007).

4.2.1 Análisis inicial de datos

En la Tabla N° 4-1 se reportó todos los resultados de las concentraciones de las 252 muestras coleccionadas durante el periodo de Julio 2008 y Junio 2009. Puesto que algunos valores de los parámetros son menores al límite de detección (LDD) de sus respectivos análisis, estos se registraron iguales a la mitad del LDD para poder obtener la matriz con datos numéricos (41; 51; 52). Este procedimiento representa una de las alternativas que existen para fijar los valores inferiores al LDD (53).

El análisis inicial de las concentraciones de los parámetros, algunos de los cuales se encuentran incluidos en los estándares nacionales de calidad de agua (ECA), en cada estación de muestreo se evaluaron mediante los estadísticos descriptivos iniciales de las distribuciones Gaussiana o normales, a saber, media, desviación estándar, sesgo y curtosis, además de los valores mínimo y máximo (Tabla N° 4-2). Estos valores, los cuales provienen de la estadística paramétrica, examinaron el comportamiento de cada una de las curvas de distribución de los datos de los parámetros y la tendencia de cada gráfica a una distribución Gaussiana o normal. Esta tendencia es fundamental para los análisis multivariados posteriores (55).

En las curvas de distribuciones normales, el valor del sesgo es cero. Una desviación de la curva de distribución hacia la derecha resulta en un valor del sesgo mayor a cero y una desviación hacia la izquierda resulta en un sesgo menor a cero. A medida que el valor del sesgo aumenta, mayor es la desviación de la curva de distribución de los datos. Las distribuciones normales además tienen una concentración media de los datos alrededor de sus valores centrales (medias), y por esto el valor de la curtosis es cero. Valores positivos y negativos de la curtosis, indican una mayor y menor concentración de los datos alrededor de las medias respectivamente (20).

Tabla N° 4-1. Base de datos originales de los 20 parámetros ambientales en las 7 estaciones de monitoreo del agua del rio Rímac (Julio 2008–Junio 2009)

FECHA	ESTACION	T	pH	CE	SST	ST	NO_3-N	NO_2-N	SO_4	PO_4-P	Cl	Alc	DT	DCa	Fe	Pb	Zn	Cu	Cr	Al	Cd
22-jul	Ricardo Palma	22.2	8.54	827	28	608	0.98	0.012	325	0.07	33.2	153	515	382	0.431	0.116	0.153	0.023	< *LDD*[3]	< *LDD*[4]	0.010
22-jul	Santa Eulalia	20.5	8.34	592	7	620	0.56	0.017	166	0.02	16.6	122	333	230	0.229	0.067	0.028	0.017	< *LDD*[3]	< *LDD*[4]	0.005
22-jul	Los Ángeles	21.3	8.95	644	13	676	0.88	0.317	219	0.30	21.5	112	368	245	0.294	0.110	0.045	0.021	< *LDD*[3]	< *LDD*[4]	0.013
22-jul	Huampaní	22.1	9.05	616	8	812	1.31	0.081	223	0.23	20.5	108	333	225	0.108	0.122	0.133	0.010	< *LDD*[3]	< *LDD*[4]	0.007
22-jul	Huachipa	19.2	8.21	609	12	584	0.56	0.160	186	0.23	17.6	108	338	235	0.454	0.104	0.205	0.051	< *LDD*[3]	< *LDD*[4]	0.011
22-jul	Del Ejército	18.9	7.54	695	300	1040	0.91	0.154	235	0.12	39.0	112	363	270	1.281	0.153	0.620	1.004	1.338	5.932	0.013
22-jul	Gambeta	20.3	7.16	1527	600	1750	0.22	0.010	231	5.40	192	271	480	353	4.331	0.141	0.390	0.181	0.065	1.473	0.017
22-ago	Ricardo Palma	22.1	8.29	826	12	812	1.00	0.018	443	0.04	39.0	163	456	338	0.888	0.044	0.225	0.021	0.007	0.084	0.005
22-ago	Santa Eulalia	20.9	8.34	560	11	500	0.56	0.015	203	0.03	20.5	118	314	230	0.219	0.030	0.144	0.021	0.006	0.041	0.003
22-ago	Los Ángeles	22.2	8.78	605	11	624	0.93	0.207	242	0.27	20.5	118	338	255	0.168	0.023	0.058	0.166	0.006	< *LDD*[4]	0.004
22-ago	Huampaní	23.9	9.34	599	13	544	1.30	0.136	262	0.27	24.4	118	377	270	0.096	0.027	0.051	0.004	0.006	< *LDD*[4]	0.008
22-ago	Huachipa	19.4	8.2	567	17	496	0.64	0.112	223	0.13	17.6	108	353	270	0.157	0.023	0.214	0.009	0.007	0.106	0.002
22-ago	Del Ejército	18.8	8.01	601	524	1124	0.84	0.200	274	0.62	22.4	114	368	260	21.97	0.106	2.695	1.136	0.130	8.538	0.019
22-ago	Gambeta	20.5	7.26	1507	576	1548	0.25	0.032	397	4.86	199	277	422	284	4.077	0.054	0.685	0.204	0.323	3.149	0.006
22-set	Ricardo Palma	22.4	8.28	830	11	708	0.97	0.011	329	0.02	36.1	155	431	348	0.362	0.177	0.059	< *LDD*[2]	0.008	< *LDD*[4]	0.008
22-set	Santa Eulalia	21.5	8.43	563	5	490	0.55	0.006	211	0.02	17.6	118	309	235	0.228	0.057	0.097	< *LDD*[2]	< *LDD*[3]	< *LDD*[4]	0.003
22-set	Los Ángeles	22.5	8.96	607	5	540	1.00	0.158	250	0.27	20.5	110	338	260	0.228	0.074	0.022	< *LDD*[2]	< *LDD*[3]	< *LDD*[4]	0.002
22-set	Huampaní	24.1	9.38	592	5	856	1.31	0.172	254	0.28	21.5	108	294	255	0.239	0.126	0.161	< *LDD*[2]	< *LDD*[3]	< *LDD*[4]	0.002
22-set	Huachipa	22.0	8.23	584	8.4	484	0.67	0.114	242	0.18	19.5	106	324	245	0.403	0.100	0.102	< *LDD*[2]	0.009	< *LDD*[4]	0.003
22-set	Del Ejército	22.3	8.5	605	275	512	0.94	0.107	195	0.14	25.4	110	284	240	1.509	0.151	0.075	0.033	0.009	0.969	0.007
22-set	Gambeta	20.9	7.44	1557	275	1380	0.19	0.009	325	7.30	193	300	367	269	1.745	0.194	0.340	0.069	0.241	1.892	0.008

Tabla N° 4-1. (Continuación)

FECHA	ESTACION	T	pH	CE	SST	ST	NO_3-N	NO_2-N	SO_4	PO_4-P	Cl	Alc	DT	DCa	Fe	Pb	Zn	Cu	Cr	Al	Cd
21-oct	Ricardo Palma	23.2	8.47	886	60	736	1.63	0.023	1019	0.06	38.0	151	437	382	0.888	0.047	0.103	0.027	0.004	0.430	0.005
21-oct	Santa Eulalia	22.6	8.64	607	39	620	0.74	0.005	336	0.07	19.5	110	280	243	0.127	0.044	0.064	0.005	0.004	< *LDD*[4]	0.003
21-oct	Los Ángeles	23.1	10.0	655	7	540	1.54	0.134	356	0.33	23.4	108	308	276	0.096	0.034	0.017	< *LDD*[2]	0.006	< *LDD*[4]	0.002
21-oct	Huampaní	23.8	11.2	663	11	716	2.11	0.117	403	0.09	16.6	102	290	257	0.507	0.041	0.030	0.009	0.003	0.376	0.003
21-oct	Huachipa	18.4	8.40	807	10	540	0.74	0.086	380	0.18	20.5	102	278	243	0.435	0.051	0.151	0.009	0.007	0.241	0.003
21-oct	Del Ejército	19.4	7.92	805	51	796	1.61	0.394	235	0.45	51.7	151	304	251	3.028	0.054	0.221	0.103	0.015	1.051	0.004
21-oct	Gambeta	21.1	7.28	1736	1040	2424	0.16	0.008	917	5.74	213	252	461	314	14.63	0.209	1.389	0.402	0.507	7.653	0.009
25-nov	Ricardo Palma	23.0	8.30	915	16	564	1.31	0.013	391	0.04	41.0	187	416	337	0.387	0.043	0.186	0.008	0.014	< *LDD*[4]	0.004
25-nov	Santa Eulalia	22.3	8.51	600	10	316	0.81	0.006	215	0.05	16.6	137	275	249	0.305	0.043	0.269	0.011	0.005	< *LDD*[4]	0.004
25-nov	Los Ángeles	23.1	9.42	609	14	360	0.91	0.041	180	0.22	17.6	125	280	241	0.089	0.029	0.169	0.006	0.003	< *LDD*[4]	0.003
25-nov	Huampaní	25.2	10.2	580	11	328	1.11	0.118	199	0.18	31.2	114	282	243	0.069	0.034	0.424	0.005	0.003	< *LDD*[4]	0.004
25-nov	Huachipa	23.0	8.10	588	19	544	0.75	0.103	164	0.15	18.5	119	290	251	0.316	0.031	0.366	0.214	0.005	< *LDD*[4]	0.002
25-nov	Del Ejército	25.4	7.81	937	348	816	0.72	0.494	180	0.30	73.1	208	310	259	4.211	0.049	0.438	0.131	0.023	1.303	0.003
25-nov	Gambeta	23.6	7.01	1897	360	1532	0.18	0.007	305	3.44	270	264	402	304	1.970	0.056	0.592	0.091	0.082	2.951	0.004
22-dic	Ricardo Palma	18.8	8.32	655	1255	2508	1.02	0.044	760	0.20	15.6	121	549	451	28.27	0.396	4.112	0.227	0.007	12.92	0.011
22-dic	Santa Eulalia	17.1	8.17	473	195	944	0.61	0.015	266	0.16	13.7	75	247	202	3.052	0.260	0.122	0.086	< *LDD*[3]	2.057	0.004
22-dic	Los Ángeles	17.5	7.88	480	485	1248	2.53	0.019	117	0.18	15.6	87	290	255	13.16	0.335	1.171	0.189	0.004	3.018	0.011
22-dic	Huampaní	17.8	8.24	484	506	1488	2.77	0.027	274	0.19	18.5	102	290	251	11.79	0.320	1.258	0.177	0.004	3.351	0.008
22-dic	Huachipa	21.2	8.19	521	299	1828	0.89	0.087	442	0.17	18.5	108	278	237	7.272	0.170	0.559	0.101	0.006	2.390	0.006
22-dic	Del Ejército	26.5	8.04	567	319	1140	4.84	0.275	285	0.18	31.2	106	273	225	7.763	0.076	0.617	0.291	0.022	2.175	0.006
22-dic	Gambeta	25.5	7.38	1068	348	1432	0.31	0.006	285	2.40	125	258	325	263	6.585	0.088	0.472	0.164	0.151	2.312	0.005

Tabla N° 4-1. (Continuación)

FECHA	ESTACION	T	pH	CE	SST	ST	NO_3-N	NO_2-N	SO_4	PO_4-P	Cl	Alc	DT	DCa	Fe	Pb	Zn	Cu	Cr	Al	Cd
21-ene	Ricardo Palma	19.5	7.98	526	540	1104	1.04	0.012	509	0.21	20.5	104	341	284	20.11	5.330	5.559	0.463	< *LDD*[3]	14.93	0.025
21-ene	Santa Eulalia	18.0	7.64	423	7	384	0.89	0.005	340	0.02	11.7	87	224	176	0.483	0.642	0.244	0.036	< *LDD*[3]	0.262	0.006
21-ene	Los Ángeles	18.8	7.58	446	1891	2380	0.52	0.024	250	0.61	13.7	100	382	363	6.650	2.986	25.92	1.171	0.027	33.60	0.056
21-ene	Huampaní	19.7	7.87	440	1182	1588	0.97	0.025	348	0.36	16.6	154	341	275	8.602	5.243	11.85	2.056	< *LDD*[3]	17.60	0.004
21-ene	Huachipa	22.6	7.74	461	508	888	0.71	0.064	348	0.42	18.5	100	292	214	18.88	0.469	4.949	0.473	< *LDD*[3]	4.693	0.025
21-ene	Del Ejército	27.5	8.05	475	99	752	0.85	0.125	211	0.14	18.5	102	225	188	3.772	1.858	0.228	0.060	< *LDD*[3]	2.560	0.005
21-ene	Gambeta	27.4	7.27	843	155	816	0.08	0.005	289	1.35	96.6	179	257	210	1.716	1.424	0.072	0.060	0.110	1.227	0.005
25-feb	Ricardo Palma	17.6	7.07	351	1068	1368	1.14	0.004	130	0.02	10.7	86	304	218	23.45	0.172	0.616	0.112	0.011	13.46	0.008
25-feb	Santa Eulalia	16.1	6.34	207	135	304	0.74	0.003	72	0.03	2.90	56	108	86	0.782	0.047	0.167	0.020	0.005	1.670	0.003
25-feb	Los Ángeles	17.5	6.61	278	423	628	0.80	0.004	76	0.06	7.80	85	216	163	14.42	0.078	0.245	0.050	0.008	4.270	0.003
25-feb	Huampaní	17.9	6.73	277	569	788	1.11	0.005	81	0.05	9.80	74	204	165	12.72	0.113	0.322	0.066	0.008	6.432	0.004
25-feb	Huachipa	20.1	6.88	285	661	968	0.72	0.011	96	0.07	8.80	76	214	169	28.49	0.154	0.539	0.103	0.011	10.68	0.005
25-feb	Del Ejército	22.3	7.22	296	1316	1620	0.78	0.018	106	0.06	12.7	79	269	186	27.56	0.234	0.585	0.158	0.018	5.892	0.006
25-feb	Gambeta	22.5	7.04	339	841	1136	0.71	0.032	111	0.09	23.4	95	249	227	32.50	0.244	1.360	0.183	0.018	12.92	0.007
26-mar	Ricardo Palma	18.4	8.01	424	146	444	1.30	0.006	147	0.08	10.1	256	203	151	12.85	0.019	0.510	0.033	< *LDD*[3]	1.768	0.011
26-mar	Santa Eulalia	16.6	7.20	293	23	200	0.57	0.005	74	0.03	5.00	189	133	119	2.248	0.013	0.313	0.034	< *LDD*[3]	0.488	0.002
26-mar	Los Ángeles	18.0	7.15	358	164	392	0.72	0.004	101	0.08	12.1	229	171	123	11.05	0.017	0.448	0.037	< *LDD*[3]	1.608	0.002
26-mar	Huampaní	18.5	7.30	366	132	404	0.67	0.005	107	0.08	10.1	236	183	155	10.10	0.013	0.657	0.034	< *LDD*[3]	1.768	0.002
26-mar	Huachipa	21.2	7.30	364	296	496	0.74	0.011	101	0.11	11.1	186	179	155	12.43	0.017	0.412	0.036	< *LDD*[3]	2.488	0.001
26-mar	Del Ejército	22.9	7.48	360	217	552	0.77	0.021	107	0.12	8.10	191	183	139	15.50	0.017	0.418	0.055	< *LDD*[3]	2.808	0.001
26-mar	Gambeta	23.6	7.04	445	357	636	0.58	0.035	141	0.11	22.2	201	203	171	15.93	0.033	0.501	0.061	< *LDD*[3]	2.488	0.004

Tabla N° 4-1. (Continuación)

FECHA	ESTACION	T	pH	CE	SST	ST	NO_3-N	NO_2-N	SO_4	PO_4-P	Cl	Alc	DT	DCa	Fe	Pb	Zn	Cu	Cr	Al	Cd
23-abr	Ricardo Palma	19.8	7.92	509	19	420	1.24	0.003	202	0.04	21.2	250	248	193	0.764	0.037	0.349	0.033	0.009	0.408	0.004
23-abr	Santa Eulalia	18.3	7.60	403	9	348	0.45	0.001	174	0.02	12.1	225	203	167	0.445	0.021	0.287	0.029	< *LDD*[3]	0.368	0.003
23-abr	Los Ángeles	19.1	7.84	428	13	356	0.66	0.004	185	0.04	22.2	231	211	185	0.339	0.015	0.279	0.025	< *LDD*[3]	0.368	0.002
23-abr	Huampaní	19.7	7.78	434	20	332	0.64	0.004	177	0.04	19.1	242	215	179	0.657	0.009	0.301	0.023	< *LDD*[3]	0.328	0.001
23-abr	Huachipa	22.1	7.76	440	19	376	0.71	0.033	175	0.08	20.1	238	207	171	0.445	0.015	0.219	0.020	< *LDD*[3]	0.408	0.001
23-abr	Del Ejército	23.8	7.80	430	29	352	0.8	0.057	147	0.07	23.2	232	244	175	0.657	0.023	0.192	0.026	< *LDD*[3]	0.648	0.001
23-abr	Gambeta	23.9	7.20	555	79	476	0.69	0.095	134	0.24	46.3	271	228	199	0.657	0.029	0.277	0.040	< *LDD*[3]	0.528	0.002
25-may	Ricardo Palma	20.9	8.20	580	20	616	1.69	0.008	261	0.03	32.2	655	302	258	0.660	0.019	0.151	0.007	< *LDD*[3]	0.046	0.003
25-may	Santa Eulalia	20.2	8.12	503	5	536	0.59	0.004	107	0.01	21.2	294	276	238	0.098	0.021	0.171	0.007	< *LDD*[3]	0.105	0.002
25-may	Los Ángeles	22.6	8.67	600	17	648	1.01	0.336	367	0.20	38.3	349	324	280	0.585	0.025	0.360	0.029	< *LDD*[3]	< *LDD*[4]	0.005
25-may	Huampaní	23.7	9.20	553	6	596	0.74	0.126	308	0.13	30.2	302	300	260	0.129	0.005	0.097	0.024	< *LDD*[3]	0.105	0.001
25-may	Huachipa	21.0	8.34	548	10	592	0.80	0.105	313	0.12	24.2	320	296	264	0.469	0.012	0.317	0.010	< *LDD*[3]	< *LDD*[4]	0.002
25-may	Del Ejército	22.9	8.30	558	37	608	1.13	0.050	325	0.08	33.2	302	306	246	3.374	0.019	0.514	0.103	0.228	0.152	0.004
25-may	Gambeta	22.1	7.25	1107	174	1088	0.68	0.006	345	2.81	171	570	334	290	1.052	0.036	0.227	0.057	< *LDD*[3]	0.248	0.003
22-jun	Ricardo Palma	21.2	8.34	700	24	716	1.91	0.007	302	0.43	40.3	309	372	302	1.484	0.025	0.196	0.012	< *LDD*[3]	0.160	0.002
22-jun	Santa Eulalia	20.3	8.17	570	7	496	0.89	0.013	202	0.27	27.2	309	274	246	0.124	<*LDD*[1]	0.213	0.007	< *LDD*[3]	0.110	0.001
22-jun	Los Ángeles	22.3	8.76	620	10	572	1.49	0.323	268	0.50	33.2	312	320	276	0.103	0.010	0.252	0.011	< *LDD*[3]	0.110	0.003
22-jun	Huampaní	22.5	9.15	605	13	596	2.22	0.100	235	0.57	35.3	327	316	272	0.146	<*LDD*[1]	0.125	0.017	< *LDD*[3]	0.160	0.002
22-jun	Huachipa	19.5	8.25	570	18	564	1.84	0.082	202	0.15	43.3	311	306	240	0.574	0.016	0.343	0.293	< *LDD*[3]	0.210	0.002
22-jun	Del Ejército	18.1	8.40	670	49	636	1.62	0.083	218	0.20	42.3	311	306	268	6.028	0.027	0.559	1.541	< *LDD*[3]	0.460	0.004
22-jun	Gambeta	20.8	7.10	1430	194	1076	0.18	0.006	218	4.75	141	283	338	298	1.173	0.043	0.381	0.033	0.192	0.160	0.002

[1] LDD: 0.004, [2] LDD: 0.002, [3] LDD: 0.002, [4] LDD: 0.03. T (°C), ´pH (unidades de pH), CE (us/cm), SST- Cd (ppm).

Los resultados mostraron que los parámetros, a excepción de la temperatura (T), tienen una gran diferencia entre los valores mínimos y máximos, y altos valores de la desviación estándar, lo cual indica la gran dispersión de las concentraciones de los parámetros en el río Rímac. Por otro lado, la mayoría de los valores de sesgo y curtosis fueron muy superiores a cero, esto es, los parámetros tuvieron curvas de distribución con altas concentraciones de los datos en la media y desviados hacia la derecha.

Por lo tanto, estos análisis iniciales de los estadísticos paramétricos identificaron que la mayoría de los parámetros se encuentran en un rango amplio de valores y con curva de los datos no tiene una distribución normal.

Tabla N° 4-2. Estadísticos de los 20 parámetros del agua del río Rímac

PARAMETRO	UNIDAD	MINIMO	MAXIMO	MEDIA	DESVIACION ESTANDAR	SESGO	CURTOSIS
T	^{0}C	16.1	27.5	21.2	2.461	0.207	-0.213
pH	Unid. De pH	6.34	11.16	8.05	0.818	0.854	1.951
CE	µS/cm	207	1897	639	322.9	2.068	4.654
SST	mg/L	5	1891	233	362.2	2.316	5.975
ST	mg/L	200	2508	809	487.9	1.682	2.820
NO_3-N	mg/L	0.08	4.84	0.98	0.663	2.900	13.57
NO_2-N	mg/L	0.001	0.494	0.071	0.098	2.201	5.290
SO_4	mg/L	72	1019	259	157.8	2.577	9.470
PO_4-P	mg/L	0.01	7.30	0.60	1.398	3.288	10.36
Cl	mg/L	2.9	270	39.7	52.00	2.810	7.412
Alc	mg/L	56	655	184	105.8	1.754	4.856
DT	mg/L	108	549	302	81.76	0.456	0.736
DCa	mg/L	86	451	243	62.73	0.353	1.075
Fe	mg/L	0.069	32.497	5.293	7.946	1.788	2.451
Pb	mg/L	0.004	5.330	0.285	0.886	4.877	24.54
Zn	mg/L	0.017	25.92	0.964	3.160	6.651	49.21
Cu	mg/L	0.002	2.056	0.154	0.339	3.834	15.83
Cr	mg/L	0.002	1.338	0.045	0.163	6.586	49.63
Al	mg/L	0.030	33.60	2.545	5.078	3.690	17.35
Cd	mg/L	0.001	0.056	0.006	0.007	4.607	27.74

Un complemento al análisis de los estadísticos descriptivos anteriores es realizado con los métodos inferenciales mediante las pruebas de normalidad de Shapiro–Wilks y Kolmogorov-Smirnov (42). Estos métodos, ambos reportado en la Tabla N° 4-3, registran valores estadísticos de los respectivos métodos y valores de probabilidad (p-valor), el cual indican el valor del nivel de confianza con el cual el método inferencial se está llevando a cabo para las correspondientes distribuciones normales de los parámetros.

Los niveles de confianza o probabilidad más usados son del 95% y 99%, lo cual lleva a p-valor de 0.05 y 0.01 respectivamente. Así, si el p-valor es menor o igual a los valores anteriores, entonces la hipótesis de que los parámetros tienen una distribución normal es rechazada al nivel de confianza evaluado (20).

Tabla N° 4-3. Valores de las pruebas de normalidad, Shapiro-Wilks y Kolmogorov-Smimov de los parámetros.

	Prueba de Shapiro-Wilk		Prueba de Kolmogorov-Smimov	
PARAMETRO	**Valor Wilk**	**p-valor**	**Valor d**	**p-valor**
T	0.9808	0.2431	0.09457	>0.20
pH	0.9542	0.0461	0.10047	>0.20
CE	0.7761	0.0000	0.23834	<0.01
SST	0.6770	0.0000	0.26402	<0.01
ST	0.8221	0.0000	0.19685	<0.01
NO_3-N	0.7612	0.0000	0.19127	<0.01
NO_2-N	0.7091	0.0000	0.23876	<0.01
SO_4	0.7706	0.0000	0.14307	<0.01
PO_4-P	0.4348	0.0000	0.38730	<0.01
Cl	0.5576	0.0000	0.32936	<0.01
Alc	0.8235	0.0000	0.19011	<0.01
DT	0.9765	0.1277	0.10204	>0.20
DCa	0.9690	0.0405	0.10435	>0.20
Fe	0.6966	0.0000	0.26921	<0.01
Pb	0.3138	0.0000	0.39226	<0.01
Zn	0.2665	0.0000	0.41614	<0.01
Cu	0.4608	0.0000	0.32804	<0.01
Cr	0.2761	0.0000	0.41371	<0.01
Al	0.5398	0.0000	0.31043	<0.01
Cd	0.5463	0.0000	0.24842	<0.01

Estas pruebas de normalidad mostraron que todos los parámetros, excepto la T y DT con un p-valor de Shapiro-Wilks de 0.2431 y 0.1277 respectivamente y un p-valor de Kolmogorov-Smirnov mayor de 0.20, tuvieron al menos un p-valor mayor de 0.05 ó nivel de confianza del 95% para ambas pruebas de normalidad. Por lo que, finalmente la mayoría de los parámetros rechazaron la hipótesis nula, el cual indica que los grupos de sus datos no tengan una curva de distribución normal.

4.2.2 Pretratamiento de datos

Siendo necesario que estos grupos de datos de cada parámetro ambiental tengan aproximadamente una distribución normal se realizó la transformación de Box–Cox (15). Esta transformación permite que estos nuevos grupos de datos tengan una aproximación a una distribución normal, y de esa manera los análisis posteriores (análisis multivariado) serán desarrollados de manera adecuada; puesto que la mayoria de estos análisis se basan en la hipótesis de la distribución normal multivariados o distribución gaussiana de los datos de los parámetros de estudio. De esa manera, el resultado son nuevos valores de estos parámetros que tienen una distribución normal para algunos parámetros y una aproximación a esta distribución para otros (ver anexo 5).

Se determinó los métodos descriptivos de normalidad para estos nuevos valores (Tabla N° 4-4), el cual mostró que los valores de la curtosis y el sesgo de los parámetros fueron reducidos a valores más cercanos a cero debido a la transformación realizada sobre los datos originales. Asimismo, se reportó las dos pruebas de normalidad para cada parámetro ambiental, en donde hay un aumento en la cantidad de los parámetros que tienen un p-valor mayor de 0.05 para ambas pruebas de normalidad.

No obstante, los resultados de la prueba de normalidad de la mitad de los parámetros ambiental rechazaron la hipótesis nula, pero los valores estadístico y por ende el valor de probabilidad (p-valor) para ambas pruebas de normalidad mejoran con respecto a los valores iniciales, con lo cual los diferentes análisis estadísticos multivariados, basados en la asunción de que los parámetros tienen normalidades multivariadas, serán desarrolladas de forma adecuada. Luego se normalizaron los datos transformados

usando el autoescalado o transformación Z para evitar la pérdida de clasificación debido a la magnitud y rango de variación de los parámetros. Los valores de cada parámetro ambiental transformado tienen una media de cero y varianza uno (31).

Tabla N° 4-4. Valores de sesgo y curtosis y las pruebas de normalidad, Shapiro-Wilks y Kolmogorov-Smirnov de los parámetros transformados (transformación Box-Cox)

	Prueba de Shapiro-Wilk		Prueba Kolmogorov-Smirnov		Sesgo	Curtosis
PARAMETRO	**Valor de Wilk**	**p-valor**	**Valor d**	**p-valor**		
T	0.9835	0.3583	0.10838	>0.20	-0.004	-0.391
pH	0.9862	0.5124	0.07743	>0.20	-0.013	0.367
CE	0.9618	0.0137	0.12384	<0.20	-0.035	0.842
SST	0.9082	0.0000	0.15425	<0.05	0.101	-1.518
ST	0.9834	0.3554	0.09036	>0.20	0.013	-0.047
NO_3-N	0.9489	0.0022	0.13183	<0.15	0.088	2.016
NO_2-N	0.9512	0.0030	0.13172	<0.15	0.031	-1.202
SO_4	0.9740	0.0864	0.07955	>0.20	-0.003	0.502
PO_4-P	0.9799	0.2121	0.06769	>0.20	0.025	-0.112
Cl	0.9469	0.0017	0.11035	>0.20	-0.084	1.740
Alc	0.9512	0.0030	0.15030	<0.05	0.057	-0.875
DT	0.9855	0.4666	0.09229	>0.20	0.026	0.576
DCa	0.9741	0.0878	0.11111	>0.20	0.040	0.839
Fe	0.9432	0.0010	0.09999	>0.20	0.056	-1.254
Pb	0.9843	0.3993	0.05801	>0.20	-0.016	0.427
Zn	0.9690	0.0406	0.11133	>0.20	-0.060	1.157
Cu	0.9822	0.2979	0.05637	>0.20	0.015	-0.388
Cr	0.8447	0.0000	0.27833	<0.01	0.495	-1.117
Al	0.9094	0.0000	0.15065	<0.05	0.042	-1.333
Cd	0.99467	0.9830	0.03679	>0.20	0.015	-0.154

La visualización de las relaciones entre las variables transformadas y autoescaladas se realizó mediante la matriz de correlación de Pearson (Tabla N° 4-5). Este análisis se realiza en la parte final, puesto que ayudará a visualizar el grado de semejanza necesaria que deben tener estos datos finales para realizar los análisis multivariado. Esto se debe a que los análisis multivariado no son adecuados si existen muy pocas correlaciones o si existe una gran concordancia entre los datos, el cual conlleva a que no sea necesaria la realización de esos análisis. La Tabla N° 4-5 reportó la matriz de los

coeficientes de correlación de Pearson *r*, de los parámetros, donde los coeficientes de correlación r>0.70 resaltados son considerados significantes a un 95 % de confianza (p-valor > 0. 05) (15, 28).

De esta manera se obtuvo la existencia de las correlaciones positivas más resaltantes entre los parámetros, tales como:

1. pH y NO_3-N, NO_2 con r de 0.77, 0.74 respectivamente, lo que indica la estrecha relación entre estos aniones y el pH.
2. CE y SO4, PO4, Cl, DT, DCa con r de 0.70, 0.81, 0.91, 0.76, 0.74 respectivamente, lo que indica que estos iones pueden contribuir directamente a la conductancia eléctrica del agua.
3. ST y Pb, Cr, Cd con r de 0.80, 0.74, 0.75 respectivamente, lo que indica la relación entre estos metales y los sólidos totales.
4. SST Y ST, Fe, Zn, Cu, Al con r de 0.84, 0.89, 0.78, 0.76, 0.90 respectivamente, lo que indica la relación entre los metales, los sólidos totales y los sólidos suspendidos.

4.2 Análisis multivariado

Todos los cálculos realizados y los resultados dispuestos en tablas y gráficos, en cada análisis estadístico multivariado, se realizó con el programa estadístico Statistica 8.0 (2007).

El análisis de clúster y el análisis de componentes principales se utilizaron para poder visualizar los agrupamientos naturales y las relaciones que tienen las estaciones de muestreo y los parámetros del agua del río Rímac. El análisis de factor sirve para determinar las fuentes de contaminación de acuerdo a los parámetros monitoreados y la influencia sobre cada estación de muestreo, y finalmente el análisis discriminante para calcular el porcentaje de eficiencia de las agrupaciones de las estaciones de muestreo en los análisis multivariado anteriores y determinar cuáles son los parámetros más discriminantes o importantes para realizar tal clasificación.

Tabla N° 4-5. Matriz de correlación (r de Pearson) de los 20 parámetros ambientales

PARAMETRO	T	pH	CE	SST	ST	NO_3-N	NO_2-N	SO_4	PO_4-P	Cl	Alc	DT	DCa	Fe	Pb	Zn	Cu	Cr	Al	Cd
T	1.00																			
pH	0.36	1.00																		
CE	0.41	0.28	1.00																	
SST	-0.17	-0.65	-0.08	1.00																
ST	0.07	-0.17	0.40	**0.84**	1.00															
NO_3-N	-0.02	**0.77**	-0.25	-0.18	-0.09	1.00														
NO_2-N	0.39	**0.74**	0.20	-0.17	0.13	**0.67**	1.00													
SO_4	0.29	0.46	**0.70**	-0.08	0.43	0.04	0.25	1.00												
PO_4-P	0.28	0.00	**0.81**	0.35	0.53	-0.25	0.36	0.38	1.00											
Cl	0.48	0.12	**0.91**	0.05	0.44	-0.26	0.18	**0.71**	**0.74**	1.00										
Alc	0.19	0.00	0.64	-0.16	-0.11	-0.17	-0.12	0.17	0.23	**0.71**	1.00									
DT	0.22	0.34	**0.76**	0.00	0.58	-0.03	0.27	**0.73**	0.40	**0.64**	0.12	1.00								
DCa	0.24	0.40	**0.74**	-0.02	0.54	0.06	0.27	**0.76**	0.39	**0.62**	0.17	**0.95**	1.00							
Fe	-0.27	-0.63	-0.23	**0.89**	0.52	-0.07	-0.20	-0.14	0.16	-0.11	-0.14	-0.12	-0.15	1.00						
Pb	-0.04	-0.20	0.05	0.52	**0.80**	-0.14	-0.02	0.23	0.17	0.01	**-0.77**	0.27	0.22	0.46	1.00					
Zn	-0.29	-0.66	-0.13	**0.78**	0.43	-0.08	-0.13	-0.01	0.21	-0.01	0.04	0.01	0.04	**0.69**	0.33	1.00				
Cu	-0.24	-0.66	-0.01	**0.76**	0.56	-0.09	0.03	0.03	0.32	0.11	-0.04	0.08	0.04	**0.73**	0.41	**0.74**	1.00			
Cr	0.09	-0.37	0.42	0.49	**0.74**	-0.33	-0.05	0.17	0.42	0.44	-0.10	0.37	0.33	0.29	0.35	0.20	0.32	1.00		
Al	-0.30	-0.68	-0.25	**0.90**	0.52	-0.13	-0.21	-0.14	0.20	-0.10	-0.18	-0.17	-0.18	**0.88**	0.45	**0.70**	**0.75**	0.34	1.00	
Cd	-0.12	-0.04	0.18	0.43	**0.75**	0.00	0.07	0.31	0.16	0.10	**-0.68**	0.46	0.37	0.38	**0.72**	0.35	0.46	0.36	0.33	1.00

4.2.1 Análisis de clúster (AC)

El análisis de agrupaciones entre las estaciones de muestreo y los parámetros se desarrolló mediante el análisis de clúster (AC) por dos diferentes procedimientos (35; 43).

El primero se realizó para el análisis de los valores medios autoescalados de las estaciones de muestreo con respecto a los 20 parámetros. Así, el uso de estas medias permite desarrollar el dendograma en donde se muestren cada una de las 7 estaciones de muestreo a lo largo de todo el año de monitoreo, con lo cual se tendrá interpretaciones inmediatas y descripciones más simples de las agrupaciones que existen entre estos (57). El segundo se realizó para el análisis de los parámetros realizado sobre toda la matriz de los datos transformados y autoescalados (7 estaciones de muestreo y 20 parámetros), puesto que únicamente se tiene un solo valor de cada parámetro ambiental a lo largo de todo el periodo de análisis.

Ambos procedimientos se realizaron mediante el algoritmo de agrupamiento jerárquico utilizando el método de Ward como algoritmo de agrupamiento de los clústeres y la distancia euclidiana y la basada en la correlación de Pearson como medida de semejanza entre las estaciones de muestreo y los parámetros respectivamente. El desarrollo del método de Ward, usando la distancia euclidiana, permite que se tenga más información sobre el contenido de los clústeres, además de ser un mecanismo de agrupación que produce una gran proporción de individuos clasificados correctamente con respecto a la mayoría de los otros métodos de agrupación (43; 45; 54).

Los resultados finales del agrupamiento de los parámetros y estaciones de muestreo son visualizados gráficamente por el dendograma, el cual da un resumen visual de los pasos de agrupación, además de encontrar las semejanzas que tienen(31).

El dendograma del AC agrupó a las estaciones de muestreo en un sólo clúster (Figura N° 4-2), a saber, GRUPO I: estación Los Ángeles, Huampaní y Huachipa, el cual tiene un bajo porcentaje de la distancia de agrupamiento, menor a 25%, el cual significa que tiene un alto grado de agrupamiento.

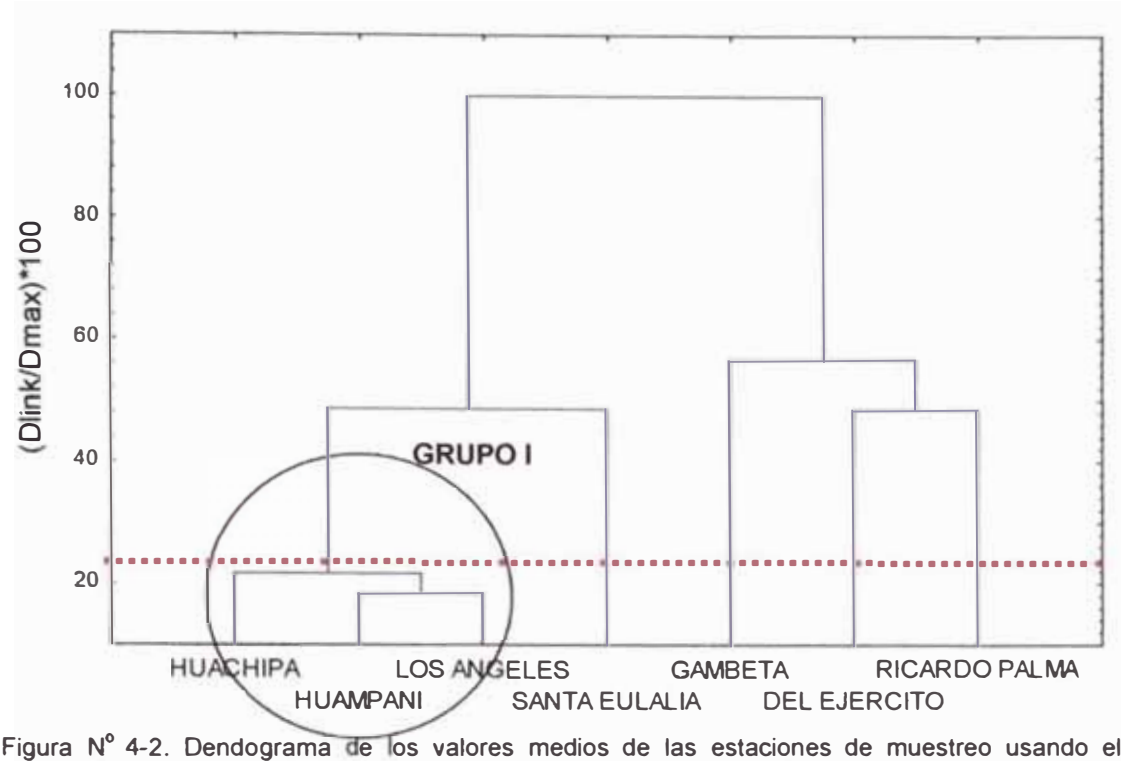

Figura N° 4-2. Dendograma de los valores medios de las estaciones de muestreo usando el método de Ward y la distancia euclidiana (julio 2008–junio 2009).

Las demás estaciones no se encuentran agrupadas, a saber, Ricardo Palma, Santa Eulalia, Del Ejército, y Gambeta. La no formación de grupos entre estas estaciones de muestreo fue principalmente por las diferentes influencias que existen sobre el agua en cada estación, debidos principalmente a la diferencia geográfica y demográfica.

Así, las estaciones Ricardo Palma y Santa Eulalia, las cuales se encuentran en las cuencas del alto Rímac y Santa Eulalia respectivamente (Ver Figura N° 5 y 6 del Anexo 4) no se encuentran agrupadas debido a la diferencia demográfica, influencias antropogénicas (minera y domestica) y las diferentes geología de los ecosistemas que hay en cada cuenca. Mientras que las estaciones Del Ejército y Gambeta, aun estando dentro de la ciudad de Lima, no se encuentran agrupadas a un alto porcentaje de agrupamiento; puesto que entre estas dos estaciones existen cinco distritos ribereños (Ver Figura N° 1 del Anexo 4).

Estos distritos influyen notablemente la calidad del agua, debido a la poca disponibilidad y un mal manejo de sus habitantes para desechar sus residuos sólidos y de las diferentes fuentes de aguas residuales que se descargan directamente al río.

No obstante, dos de las estaciones del GRUPO I, Los Ángeles y Huampaní, a una distancia de agrupamiento de 8.32%, fueron las estaciones de muestreo de aguas de características muy similares, esto debido a que se encuentran en áreas demográfica y geográficamente similares (ambas estaciones se encuentran entre los distritos de Chaclacayo y Lurigancho-Chosica), con una distancia pequeña de 3.5km (Ver Figura N° 4 del Anexo 4).

Por otro lado, el dendograma del AC (Figura N° 4-3) agrupó a los parámetros en 5 clústeres. Clúster I: Cd, Pb, Cr, ST, clúster II: Cu, Zn, Fe, Al, SST, clúster III: NO_3-N, NO_2-N, pH clúster IV: DCa, DT, SO_4, clúster V: Alc, PO_4-P, Cl, CE. La formación de todos los clústeres, cuya numeración se establece de izquierda a derecha, tiene un bajo porcentaje de la distancia de agrupamiento, esto es, a 20%.

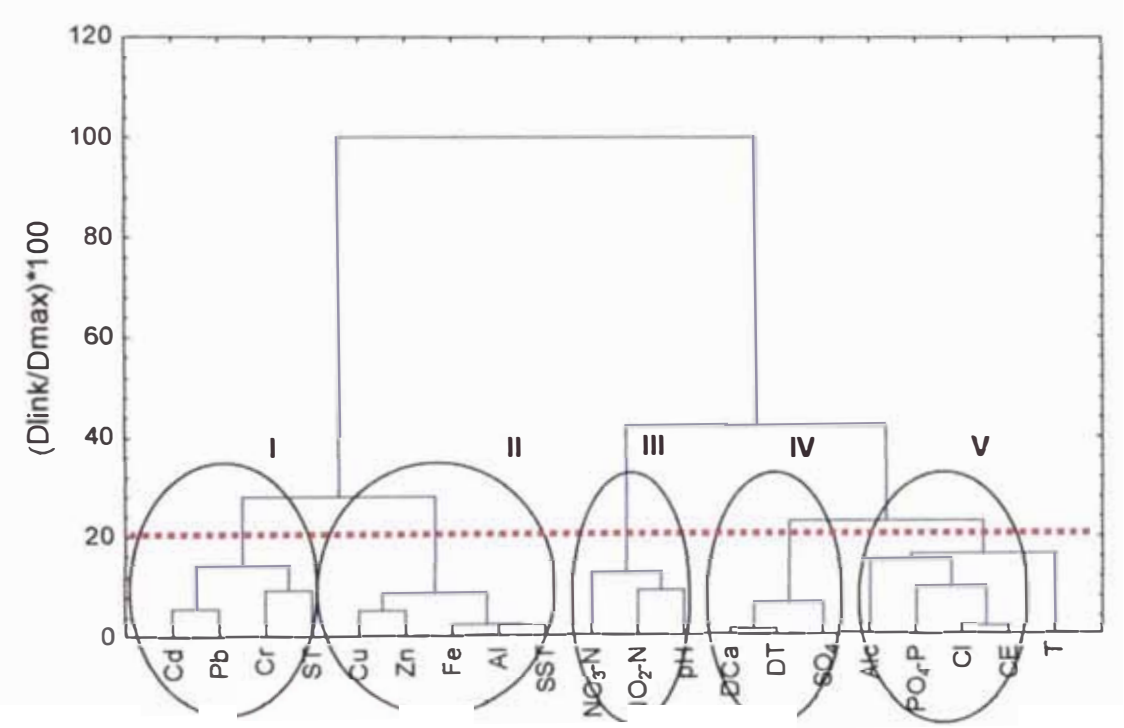

Figura N° 4-3. Dendograma basado en los datos autoescalados de los 20 parámetros usando el método de Ward y la distancia r de Pearson (periodo de julio 2008–junio 2009).

Así, el clúster I agrupó a los metales que no son esenciales para la mayoría de las especies acuáticas, a saber, Cd, Pb y Cr. Estos metales, los cuales tienen un posible origen en las diversas actividades industriales mineras y de los minerales de la corteza terrestre, están agrupados con los sólidos totales (ST) a un porcentaje de agrupamiento de 14.27%. Esta agrupación permitió determinar que estos metales se encuentran en el

agua del río Rímac en forma de materia suspendida y disuelta, debido a la interacción entre el sedimento y el agua, el cual aumenta en épocas de caudales altos. Esta materia suspendida son complejos de los metales con especies orgánicas e inorgánicas.

El clúster II agrupó a los metales de mayor concentración en la naturaleza y algunos esenciales para mantener la vida acuática en concentraciones trazas, a saber, Cu, Zn, Fe, Al. Estos metales, los cuales se encuentran en la naturaleza en diversas formas, tales como Fe y Al con 7.4 y 4.7% de la corteza terrestre respectivamente (1), y en altas concentraciones en los sedimentos de los ríos, se encuentran agrupados con los sólidos suspendidos totales (SST). Esta agrupación, realizada a un porcentaje de agrupación de 8.69%, permitió determinar que estos metales se encuentran en el agua del río Rímac en mayor medida en forma de materia suspendida, tales como los compuestos de oxido hidratado de Fe (III), normalmente representado como (Fe_2O_3.x (H_2O), Fe ($OH)_3$, el cual enlaza y precipita a la mayoría de los compuestos de los metales en los sedimentos (1).

La presencia de los metales en los clústeres I y II en el agua del río Rímac en los diferentes compuestos de la materia suspendida en aguas no contaminadas, tiende a correlacionarse bien con los minerales que originaron los sólidos suspendidos debido a la interacción con los sedimentos del agua y el traslado de los diversos compuestos de la corteza terrestre mediante las lluvias y aguas subterráneas. En las aguas contaminadas, aparecen anomalías donde las fuentes externas de influencia, tales como las fuentes mineras e industriales se añaden al contenido de metales del agua (1).

El clúster III agrupó el pH con el NO_2-N y NO_3-N, a un porcentaje de agrupación de 12.66%. Esta agrupación se debe a la estrecha dependencia de los compuestos de nitrógeno con el pH debido a la oxidación en la naturaleza de las formas reducidas de compuestos de nitrógeno, principalmente el amoniaco (NH_3), a nitritos en condiciones aeróbicas por las bacterias nitrificantes autótrofas (grupo de *nitrosomona*) (ecuación 49). Estas bacterias nitrificantes se encuentran tanto en la superficie terrestre y en el agua, y cuyo pH óptimo para su crecimiento se da entre 7.1 y 8. Siendo por esto que a pH menores de 6 no ocurre la nitrificación (1; 4; 58).

$$2NH_3+3O_2 \xrightarrow{\text{bacterias}} 2NO_2^-+2H^++2H_2O \quad (49)$$

Estos nitritos formados son inestables en la naturaleza, por lo cual son oxidados rápidamente por el bacterias nitrificante del grupo *nitrobacteria*, conocido como formador de nitrato (30, 26), en condiciones aeróbicas a nitratos (ecuación 50).

$$2NO_2^- + O_2 \xrightarrow{\text{bacterias}} 2NO_3^- \quad (50)$$

El clúster IV agrupó a los parámetros DCa, DT, SO_4, a un porcentaje de agrupamiento de 6.53%, los cuales están unidos de acuerdo a la relación química que existen entre los cationes Ca^{2+} y Mg^{2+} con el anión SO_4^{2-}. Esta unión se debe principalmente a la solubilización de los sedimentos de formación paleógenas del río Rímac, el cual se caracteriza por tener en su composición rocas calcáreas y areniscas de la corteza terrestre(47; 59). Esta composición del sedimento está constituido principalmente de yeso natural ($CaSO_4.2H_2O$), que contiene al anión SO_4^{2-}, el cual se encuentra en concentraciones variables prácticamente en todas las aguas naturales (1; 4). Asimismo, el sedimento también contiene a los cationes Ca^{2+} y Mg^{2+} que se encuentran en los minerales menos soluble como la calcita ($CaCO_3$), magnesita ($MgCO_3$), dolomita ($CaCO_3.MgCO_3$) entre otros minerales que contiene iones Al^{3+}, Fe^{3+}, Si^{2+}, Cl^-, Na^+ y K^+ (1; 9).

Los minerales son depositados en los sedimentos del agua a partir del Intemperismo, el cual ocurre debido a las corrientes de agua que erosionan los compuestos de las rocas para depositarlo como sedimentos en el agua. La solubilización de los compuestos que forman los minerales de las rocas para la formación de los iones correspondientes se producen por la interacción del agua con los sedimentos, el cual se solubiliza a medida que aumenta la concentración de gases disueltos ácidos como el gas carbónico, CO_2, normalmente presente como resultado de la acción bacteriana (ecuaciones 51, 52 y 53).

$$CaSO_4.2H_2O_{(s)} + H_2O_{(l)} \rightarrow Ca^{2+}_{(ac)} + SO_{4\,(ac)}^{2-} + 2H_2O_{(l)} \quad (51)$$

$$CaCO_{3(s)} + CO_{2\,(ac)} + H_2O_{(l)} \rightarrow Ca^{2+}_{(ac)} + 2HCO_{3\,(ac)}^- \quad (52)$$

$$MgCO_{3(s)} + CO_{2\,(ac)} + H_2O_{(l)} \rightarrow Mg^{2+}_{(ac)} + 2HCO^-_{3\,(ac)} \quad (53)$$

La disminución del pH del agua por influencia antropogénica, especialmente por los relaves de las actividades mineras libera en mayor concentración los iones Ca^{2+}, Mg^{2+} y SO_4^{2-} (ecuaciones 54 y 55) (1; 4). Esta concentración proviene de los minerales calcita magnesita y las rocas calcáreas de las riberas y lecho de los ríos, y se da principalmente en épocas de lluvia (12; 60).

$$CaCO_{3(s)} + 2H^+ + SO_4^{2-} \rightarrow Ca^{2+} + SO_4^{2-} + H_2O + CO_{2\,(g)} \quad (54)$$

$$MgCO_{3(s)} + 2H^+ + SO_4^{2-} \rightarrow Mg^{2+} + SO_4^{2-} + H_2O + CO_{2\,(g)} \quad (55)$$

El clúster V agrupó a la Alc, PO_4-P, Cl con la CE a un porcentaje de agrupamiento de menor a 20%. Estos aniones están relacionados con los valores de la conductancia eléctrica del agua de río, siendo los iones cloruros de mayor influencia.

La alcalinidad del agua está relacionada a la interacción con las áreas geográficas y estratos minerales alcalinos. La actividad minera aumenta la alcalinidad por la exposición de la capa superficial alcalina de las minas a cielo abierto a las aguas del río o subterráneas (1). El aumento de los iones Cl^- es por los desechos humanos, especialmente la orina, y de los iones PO_4^{2-} por la existencia de efluentes que contienen detergentes sintéticos con altas cantidades de fosforo inorgánico, siendo el principal aditivo el tripolifosfato de sodio, $Na_5P_3O_{10}$, el cual se hidroliza gradualmente en solución acuosa y se revierte en la forma de ortofosfato como: HPO_4^{2-}, PO_4^{3-}, $H_2PO_4^-$ (2; 4).

Finalmente, la temperatura no se agrupa con ningún parámetro ambiental, con lo cual se indica que las diferentes fuentes de contaminación no varían notablemente la temperatura del agua del río Rímac, pudiendo tener otro tipo de influencia como el clima.

4.2.2 Análisis de componentes principales

El análisis de clúster implica solamente agrupaciones de los clústeres de acuerdo a ciertas características fisicoquímicas de los parámetros, pero no relaciones que puedan

tener (56). Por esto, con el objetivo de visualizar los agrupamientos, diferencias y posible influencias sobre las estaciones de muestreo, las interrelaciones de los parámetros, y al mismo tiempo, simplificar la estructura original de la base de datos (reducción de las múltiples dimensiones) (36; 53), se aplicó el análisis de componentes principales (ACP) mediante la descomposición de valores singulares (DVS) a la matriz de datos transformados (transformación Box-Cox) y autoescalados.

Tabla N° 4-6. Componentes principales de los datos transformados y autoescalados

COMPONENTE PRINCIPAL	AUTOVALOR	% VARIANZA EXPLICADA	AUTOVALOR ACUMULADO	% VARIANZA ACUMULADA
1	6.13	30.65	6.13	30.65
2	5.44	27.21	11.57	57.86
3	2.27	11.37	13.85	69.23
4	1.43	7.18	15.28	76.41
5	1.18	5.90	16.46	82.31
6	0.63	3.16	17.09	85.47
7	0.61	3.05	17.71	88.53
8	0.44	2.20	18.15	90.73
9	0.35	1.75	18.49	92.48
10	0.31	1.53	18.80	94.01
11	0.26	1.33	19.07	95.34
12	0.21	1.04	19.27	96.37
13	0.17	0.83	19.44	97.21
14	0.14	0.69	19.58	97.90
15	0.12	0.61	19.70	98.51
16	0.10	0.49	19.80	98.99
17	0.08	0.40	19.88	99.39
18	0.05	0.27	19.93	99.66
19	0.04	0.18	19.97	99.84
20	0.03	0.16	20.00	100.00

El resultado fue la extracción de veinte autovalores de los componentes principales, y el respectivo porcentaje de la varianza explicada y acumulada (Tabla N° 4-6). La suma de

todos los autovalores es igual al número de parámetros evaluados, y de una manera de reducir la cantidad de parámetros mediante el conocimiento de las relaciones entre estos, los componentes principales toman un cierto autovalor de acuerdo al grado de explicación que tienen sobre estas relaciones. Mientras más alto sea el autovalor, mayor es el grado de explicación que tiene de cada componente de las relaciones de los parámetros, lo cual indican un mayor porcentaje de la varianza explicada, calculada a partir de la división entre el autovalor correspondiente y el número de parámetros.

Sin embargo, existen algunos componentes principales con bajo autovalor que tienen información redundante e innecesaria entre los parámetros (bajo porcentaje de varianza explicada). Por tanto, se tomó en cuenta dos de los varios criterios que existen en el ACP, esto para tomar en cuenta los CPs óptimos que expliquen las relaciones importantes y sin tener una pérdida de información necesaria.

El primer criterio está basado en el criterio de Kaiser o del autovalor > 1, el cual considera sobre la matriz de datos autoescalados que cada CP debe explicar al menos un parámetro ambiental. De esta manera, este criterio considera que los cuatro primeros CPs son los más óptimos, debido a que después del CP4, los CPs restantes tienen autovalores menores o cercanos a 1. El segundo criterio es una manera gráfica de observar los componentes principales óptimos. Este criterio está basado en el gráfico de sedimentación (Figura N° 4-4), el cual determina gráficamente el número de CPs óptimos que tienen la pendiente del grafico más vertical.

De esa manera, luego del CP4, el gráfico tienen una menor pendiente considerando por lo tanto a estos cuatro primeros CPs como los óptimos, los cuales explican el 76.41% de la información o varianza explicada de las relaciones entre los parámetros transformados. La adición de más CPs contribuye muy poco a la varianza total.

Debido a los resultados de los dos criterios anteriores, toda la información inicial de los parámetros y estaciones de muestreo pudieron ser explicadas y desarrolladas en el espacio reducido de las cargas y puntuaciones de los 4 primeros CPs. Por lo tanto, toda la información se pudo observar en dos graficas bidimensionales, correspondiente a los

primeros cuatro CPs. De esta manera, se explicó el 76.41% de la información inicial de las relaciones de los parámetros.

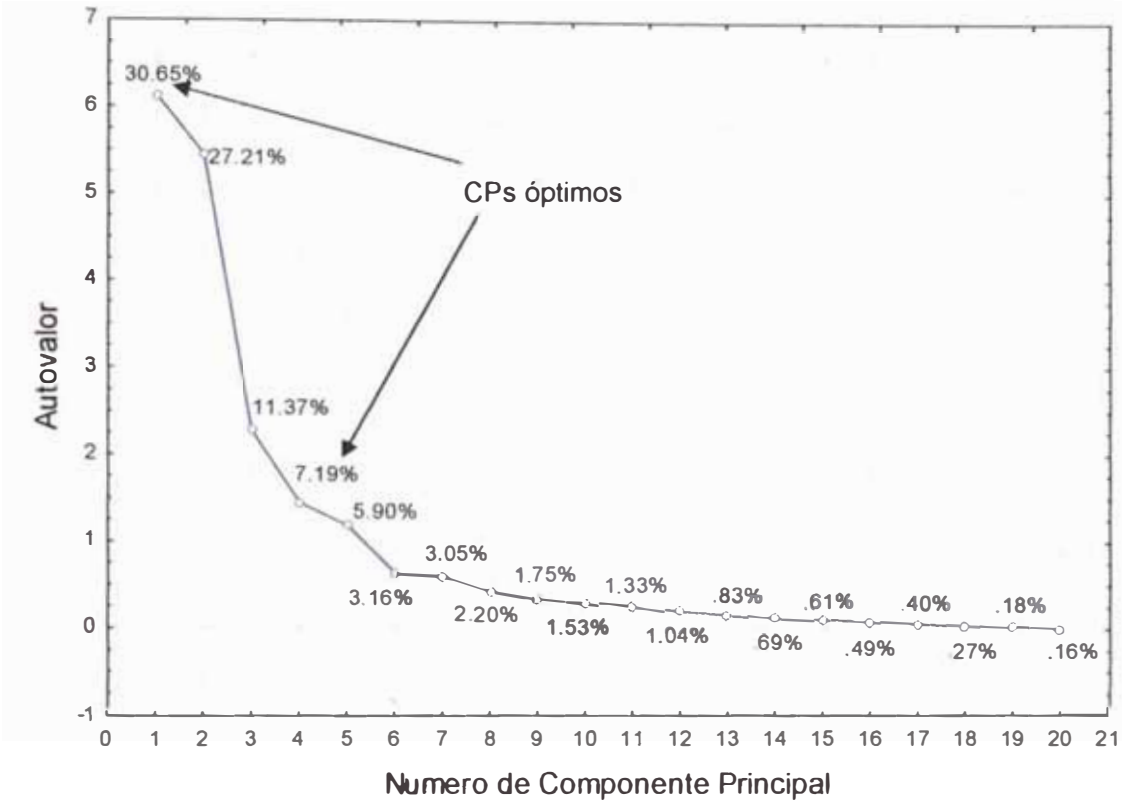

Figura N° 4-4. Gráfico de sedimentación de los autovalores de los 20 componentes principales.

En primer lugar, las Figuras N° 4-5 (a) y (b) mostraron las coordinaciones de las cargas de los primero cuatro CPs. Estos valores son visualizados como vectores, el cual tienen una norma o valor que varía en un rango de -1 a +1. De esta manera, la interpretación o peso de los componentes principales sobre cada parámetro aumenta si la norma de cada vector sea más cercana a los valores extremos de la norma, esto es más cercanos al valor de 1; tanto de manera negativa como positiva. Los parámetros más correlacionados son los que tuvieron los ángulos más agudos entre sus vectores de coordinación de las cargas y los que tienen ángulos cercanos a 180 tienen una relación inversa.

Así, la Figura N° 4-5 (a) señaló que:

Los componentes principales 1 y 2 visualizaron el agrupamiento entre los parámetros del clúster I y el clúster II realizados por el AC, los cuales son nombrados como clase I y clase II. Sin embargo, el ACP visualiza ciertas relaciones, tales como:

Los metales Cu, Zn, Al y Fe se encuentran relacionados negativamente al pH. Esto es debido a que valores más altos de pH en el agua, los metales tienden a precipitar sobre el sedimento de los ríos, en donde los metales son retenidos, siendo por esto que la concentración de los pocos metales disueltos y los sólidos suspendidos disminuyen en el agua del río Rímac (48). Sin embargo, a medida que el valor de pH disminuye, la disolución de los compuestos sólidos de los metales, principalmente en forma de sólidos suspendidos aumenta, debido a que las aguas de menor pH de lo normal, excede la capacidad de tamponamiento del suelo (11).

La gráfica también mostró el agrupamiento entre los parámetros DT, DCa, Cl y SO_4 con la CE, lo cual indica que existe un aporte importante además de los cloruros de los cationes Ca^{2+}, Mg^{2+} y SO_4^{2-} a la conductancia eléctrica, dejando en menor aporte a los iones PO_4^{3-}. Por lo tanto, los clústeres IV y V hallados por el análisis de clúster forman un solo grupo (clase IV), esto es, el grupo de los iones que más contribuyen a la conductancia eléctrica del agua del río.

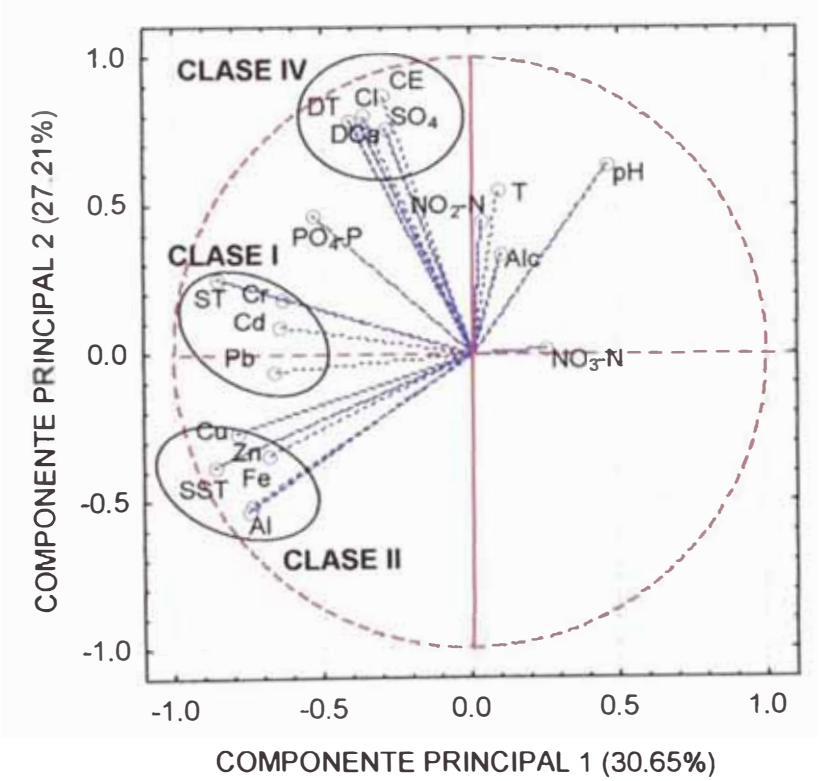

Figura N° 4-5. (a) Gráfico de dispersión de las coordinaciones de las cargas de los CP1 y CP2 correspondientes a los parámetros (julio 2008-junio 2009).

Finalmente, el componente principal 1 está influenciado principalmente por los parámetros de la clase I y II, es decir los metales que provienen tanto de las actividades antropogénicas (mineras e industriales) y fuentes naturales (minerales de la corteza terrestre), mientras que el componente principal 2 por los parámetros de la clase IV, que provienen de los principales compuestos solubles del agua de influencia geológica y actividades antropogénicas (domesticas y minera).

La Figura N° 4-5 (b) señaló que:

Los componentes principales 3 y 4 comprobaron el agrupamiento del clúster III por el AC. Sin embargo, el ACP visualiza una relación negativa entre el Cd y Pb con la Alc. Esto es debido que la velocidad de precipitación de ambos parámetros, que se encuentra mayormente en sólidos disueltos y suspendidos, se incrementa sobre el lecho de los ríos a una alta alcalinidad. Puesto que cuanto mayor es la alcalinidad del agua mayor es su capacidad de mantener un valor de pH fijo básico frente a los cambios de pH (12).

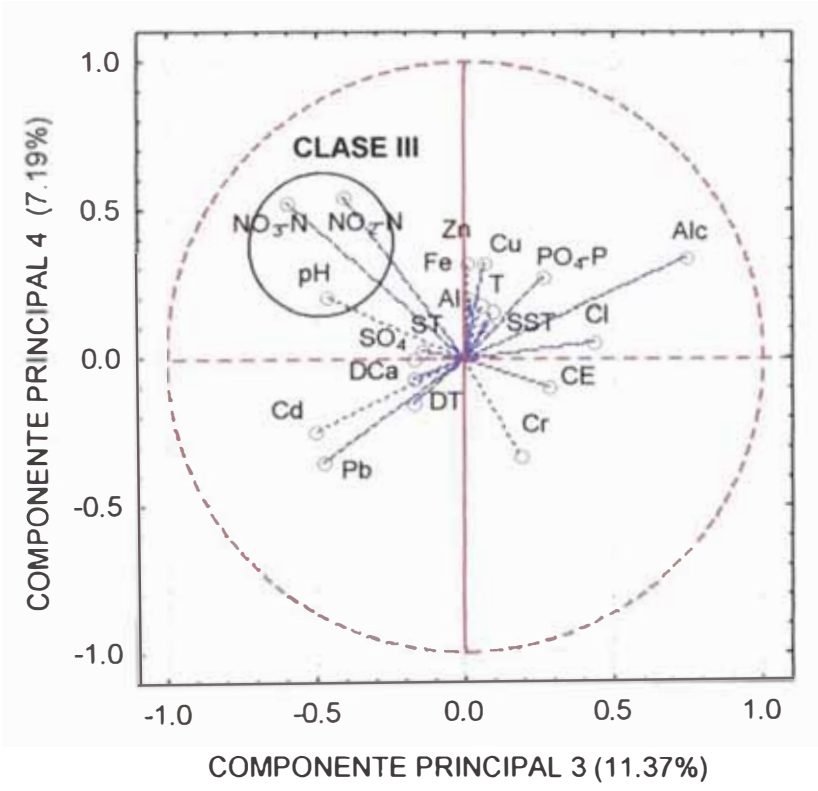

Figura N° 4-5. (b) Gráfico de dispersión de las coordinaciones de las cargas de los CP3 y CP4 correspondientes a los parámetros (julio 2008-junio 2009).

Estos cambios de alcalinidad afectan principalmente a los compuestos de Pb ($PbCO_3$) y cadmio ($CdCO_3$) en la naturaleza, los cuales son solubles en agua superficial cercanas a la neutralidad (60). A baja alcalinidad, que está relacionado a pH bajos, la concentración de Cd y Pb en forma disuelta en el agua es más alta (11).

Finalmente, el componente principal 3 está influenciado principalmente por la alcalinidad, proveniente de la interacción del CO_2 de la atmosfera y los compuestos de carbono del agua (ciclo biogeoquímico del carbono) y en una mayor concentración de las actividades urbanas e industriales, mientras que el componente principal 4 por los compuestos de nitrógeno (nitratos y nitritos), los cuales provienen de las fuentes naturales (ciclo biogeoquímico del nitrógeno), e influencia antropogénica (actividades agrícolas e industriales).

Las Figuras N° 4-6 (a) y (b) mostraron los gráficos biplot, tanto de las medias de las coordinaciones de las puntuaciones y las medias estandarizadas de las cargas de los parámetros de los primero 4 CPs. Estos valores tienen valores negativos y positivos y no se encuentran en un rango de valores específico. La influencia de los componentes principales sobre cada estación de muestreo aumenta si tienen mayores valores tanto de manera negativa como positiva.

Así la Figura N° 4-6 (a) señalo que:

El componente principal 1, con 30.65 % de varianza explicada, separa a las estaciones Gambeta y Santa Eulalia debido a la posible influencia de los parámetros de los metales en forma de sólidos suspendidos y totales (clase I y II), de origen antropogénico y geológico.

El componente principal 2 con 27.21% de varianza explicada indica que existe una diferencia entre las estaciones Gambeta y Ricardo Palma con Santa Eulalia por la posible influencia de metales que están en mayor concentración en la naturaleza (clase IV) y componentes de la conductancia eléctrica del agua del río de origen geológico y antropogénico (clase II).

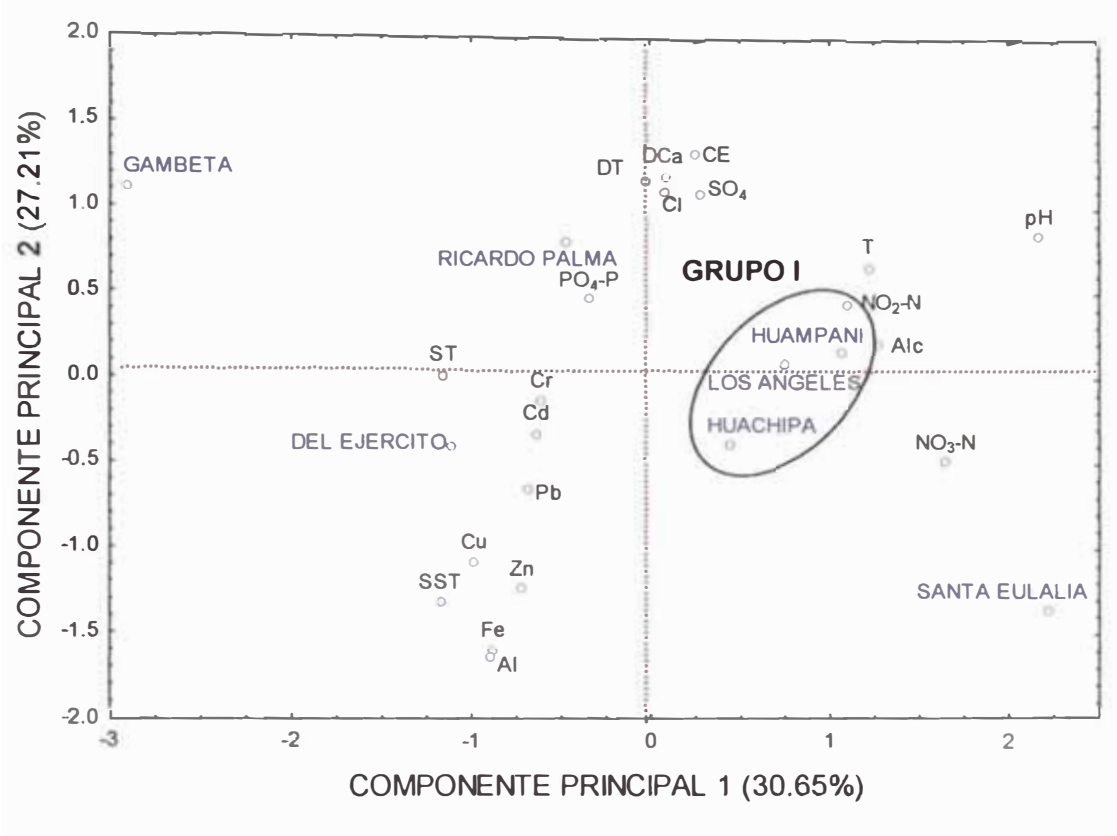

Figura N° 4-6. (a) Gráfico Biplot de las medias de las coordinaciones de las puntuaciones correspondientes a las estaciones de muestreo y las medias de las cargas estandarizadas correspondientes a los parámetros de los CP1 y CP2 (julio 2008-junio 2009).

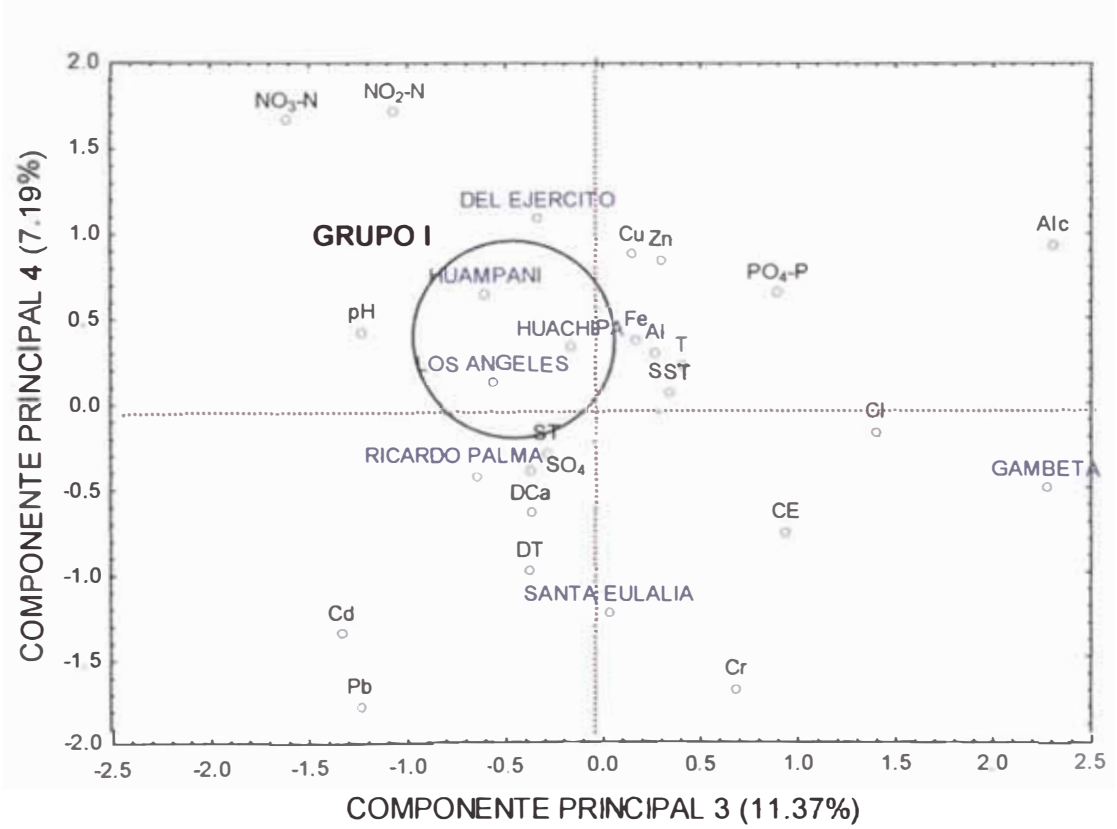

Figura N° 4-6. (b) Gráfico Biplot de las medias de las coordinaciones de las puntuaciones correspondientes a las estaciones de muestreo y las medias de las cargas estandarizadas correspondientes a los parámetros de los CP3 y CP4 (julio 2008-junio 2009).

De la misma manera, la figura 4-6 (b) señalo que:

El componente principal 3, con 11.37% de varianza explicada, indica que la estación Gambeta se separa de las demás estaciones de muestreo debido a la influencia de la alcalinidad, el cual pueda deberse a la diferencia geología e influencia de las actividades industriales y domesticas, mientras que el componente principal 4 con 7.19% de varianza explicada indica que hay una diferencia entre las estaciones Del Ejercito y Santa Eulalia debido a la influencia de los iones nitrato y nitrito y los metales menos abundantes y más tóxicos en la naturaleza (Cd, Pb y Cr), provenientes de las actividades agrícolas, mineras y de la geología del lugar.

El análisis de componentes principales (ACP) además agrupó a las estaciones de muestreo en un solo clúster, GRUPO I, Los Ángeles, Huampaní, Huachipa; de poco peso o valor de las coordinaciones de las puntuaciones en los 4 componentes principales, estando las demás estaciones de muestreo no agrupadas, a saber, estación Ricardo Palma, Santa Eulalia, Del Ejército y Gambeta. Así, estos resultados confirmaron la agrupación de las estaciones de muestreo realizada por el análisis de clúster.

4.2.3 Análisis de factor (AF)

El análisis de factor implica la eliminación de la información redundante tomada en cuenta en el ACP con el objetivo de visualizar los parámetros que se encuentren más relacionados y de esa manera obtener sus fuentes de origen. Además, el AF también visualiza la influencia que tienen las fuentes de origen de los parámetros sobre las estaciones de muestreo en cada mes del año. Para esto, se aplicó la técnica de los factores principales mediante el análisis de las comunalidades = R^2 a la matriz de datos transformados (transformación Box-Cox) y autoescalados. Finalmente se aplicó la rotación varimax a los ejes de los factores (rotación de las cargas de los parámetros).

El criterio de Kaiser o del eigenvalor > 1 fue tomado en cuenta por el AF para desarrollar las estructuras de los parámetros con fuerte correlaciones y las influencias de estas estructuras sobre cada estación de muestreo, facilitando la interpretación de los resultados y la eliminación de información redundante (31).

Los análisis fueron desarrollado por medio de la técnica de los factores principales llamado método de las comunalidades= R^2, donde R es la correlación multivariada entre los parámetros. Esta técnica calculó la matriz de correlación reducida de los parámetros, la cual tiene en su diagonal principal a las llamadas comunalidades, y los respectivos cálculos de los autovalores y autovectores por medio de ecuaciones lineales múltiples. De esa manera, el AF tomó en cuenta solamente los cuatro primeros componentes principales, que ahora son llamados factores principales (Tabla N° 4-7), los cuales explican el 72.26% de la varianza explicada o información entre las relaciones de los parámetros. Este valor fue menor al calculado por el ACP (76.41 % de la varianza), indicando que esta diferencia de varianza es debido a las varianzas especificas propias de cada parámetro ambiental y a los errores aleatorios.

Tabla N° 4-7. Factores principales de los datos transformados y autoescalados

FACTOR PRINCIPAL	AUTOVALOR	% VARIANZA EXPLICADA	AUTOVALOR ACUMULADO	% VARIANZA ACUMULADA
1	5.95	29.74	5.95	29.74
2	5.29	26.47	11.24	56.20
3	2.06	10.29	13.30	66.49
4	1.15	5.769	14.45	72.26

Los valores de las cargas de los factores fueron luego modificados mediante una de las rotaciones ortogonales del AF, esto es, la rotación varimax. Esta rotación reduce el traslape y facilita la interpretabilidad de las cargas de los parámetros sobre cada factor (45; 51). Estos nuevos factores con cargas modificadas por la rotación varimax son llamados varifactores y pueden ser interpretados como influencias de origen común (15; 42). De esta forma, los valores de las cargas rotadas, igual que las coordinaciones de las cargas para el ACP, reflejan las correlaciones entre las variables y los varifactores (52).

Estos valores varían entre un rango de -1 a +1, indicando estos valores las correlaciones mas fuertes (positiva o negativa). Sin embargo, la rotación varimax permite que la dispersión de los factores sea maximizada, reduciendo el número de coeficientes con cargas altas y bajas (32). Dejando por lo tanto que los primeros factores tengan la mayoría de los parámetros con los valores más altos (valores cercanos a +1 y -1),

mientras que los valores de las cargas más bajos (más cercanos a 0) se encuentren en los restantes factores.

Por esto, se considera como "fuerte", "medio", y "débil" a los correspondiente valores absolutos de las cargas de > 0.75, 0.55 a 0.75, y 0.30 a 0.50 respectivamente (7; 29; 31; 32), con lo cual, no se considera significantes los parámetros con cargas de valor absoluto de los factores <0.50 (52). Este valor límite fue indicado en las graficas por un cuadrado interno, cuyos límites dividen a los parámetros significantes de los que no son para el correspondiente varifactor.

Así, la Figura N° 4-7(a) señaló que:

El varifactor 1 agrupó a los parámetros: Zn, Cu, Fe, Al y SST, por lo cual describe el factor de origen metálico del agua del río Rímac, el cual contiene a los metales agrupados mayormente en forma de materia suspendida. Este factor se encuentra influenciado, para altas concentraciones de metales adicionadas, por las actividades industriales y mineras; y en menor concentración por las interacciones con los sedimentos y la corteza terrestre debido a los cambios de los factores meteorológicos, tales como las lluvias y la temperatura ambiental.

Por otro lado, tal como lo analizó el ACP, existe una relación inversa entre los parámetros de la clase II y el pH, con lo cual confirma la relación fisicoquímica que existe entre estos evaluada por el ACP, puesto que a valores de pH básicos del agua, los iones y materia suspendida de los metales tienden a precipitar sobre el sedimento de los ríos y a pH ligeramente ácidos, hay una disolución de los metales de la materia suspendida.

El varifactor 2 agrupó a los parámetros PO_4-P, SO_4, DCa, DT, Cl y CE, por lo cual describe al factor de origen geológico del agua del río Rímac en bajas concentraciones, el cual es debido principalmente a la naturaleza del area de estudio y por ende los minerales que se encuentran en los sedimentos y material suspendido en el agua del río (ecuaciones 51, 52 y 53) (4).

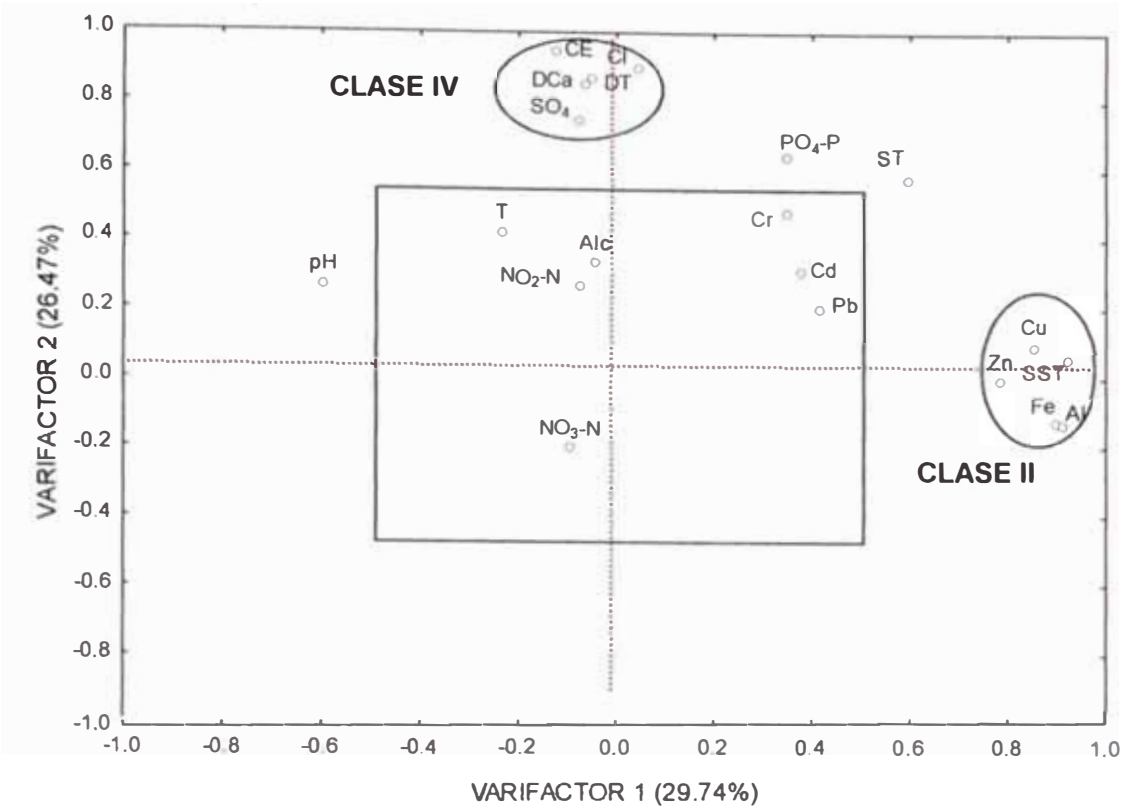

Figura N° 4-7. (a) Grafico de dispersión de las cargas de los varifactores 1 y 2 (rotación varimax) correspondientes a los parámetros.

En altas concentraciones, la influencia se debe a la contaminación de las aguas residuales domestica (desechos humanos, detergentes, etc.) y por las actividades mineras que lixivian a los iones Ca^{2+}, Mg^{2+} y SO_4^{2-} de los suelos y los transportan hacia al río, interaccionando con los sedimentos y material suspendido del agua.

Asimismo, el parámetro ST (sólidos en forma de disuelta y suspendida) está relacionado por ambos varifactores, varifactor 1 y varifactor 2, por lo cual involucra a la mayoría de los metales (clase II) y aniones (clase IV) y PO_4-P en forma de materia suspendida y disuelta provenientes de diversos factores de influencia (urbano, industrial y geológico).

La Figura N° 4-7 (b) muestra que:

El varifactor 3 agrupó los parámetros Cd y Pb y la Alcalinidad. Este varifactor describe la influencia de estos metales tóxicos provenientes de la corteza terrestre y característica de zonas mineras, además de la mayor alcalinidad del agua sobre el río Rímac debido a las actividades mineras a tajo abierto. Además se confirma la relación fisicoquímica entre estos parámetros tal como lo analizó el ACP.

El varifactor 4 agrupó los parámetros NO_2-N, NO_3-N, pH. Este varifactor describe el factor de origen agrícola debido a la aplicación de fertilizantes en las tierras de cultivo en forma de sales de amonio anhidro o sales de amonio, los cuales por la transformación microbiana son oxidados a nitritos y luego a nitratos y los cuales son asimilados por las plantas (ecuación 49 y 50). Una alta concentración de estos parámetros se podría originar por el desplazamiento de los compuestos nitrogenados de los fertilizantes por las escorrentías agrícolas desde los campos de cultivo al río, debido al riego de los campos agrícolas o las lluvias en épocas de verano (1; 60).

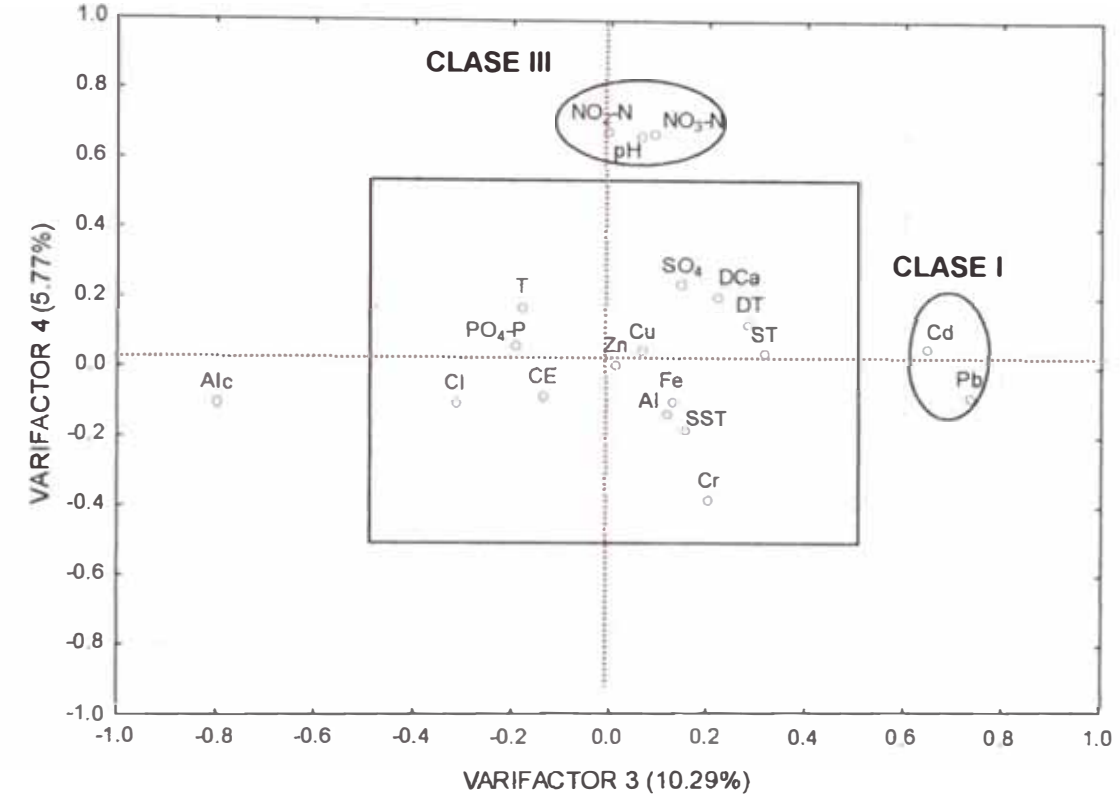

Figura N° 4-7. (b) Grafico de dispersión de las cargas de los varifactores 3 y 4 (rotación varimax) correspondientes a los parámetros.

Por otro lado, se graficaron los valores de las puntuaciones de los varifactores correspondientes a cada estación de muestreo durante todo el periodo de estudio (Julio 2008-Junio 2009), ver Figuras N° 4-8 (a), (b), (c) y (d). Estos gráficos miden la influencia que tiene cada varifactor, y por ende cada tipo de influencia sobre cada estación de muestreo. De esta manera, los valores de las puntuaciones más positivas corresponden a una alta influencia del varifactor sobre las estaciones de muestreo, mientras valores negativos corresponden a una influencia del varifactor correspondiente a los valores

negativos de las cargas de los parámetros sobre las estaciones de muestreo. Valores de las puntuaciones cercanas a cero reflejan una nula influencia de los varifactores sobre las estaciones (31).

Así, la Figura N° 4-8 (a) muestra que:

Las estaciones de muestreo que tuvieron un rango mayor de influencia de los parámetros que provienen de las actividades industriales y los sedimentos en el agua de río, esto es, un rango de mayores puntuaciones del varifactor 1 fueron: Gambeta y Del Ejército, y de menor influencia la estaciones de muestreo: Santa Eulalia.

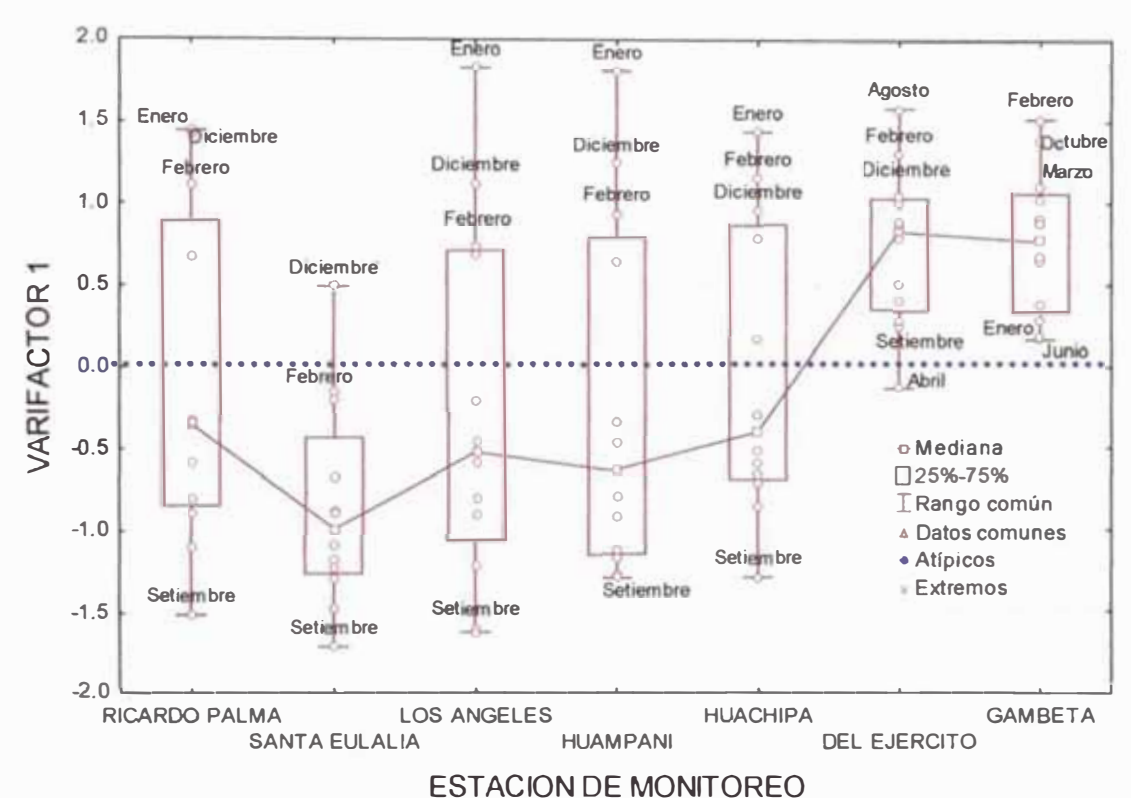

Figura N° 4-8. (a) Gráfico de cajas y bigotes de las puntuaciones del varifactor 1 (rotación varimax) correspondientes a las estaciones de muestreo.

Las estaciones de Gambeta y Del Ejército tienen estos mayores valores de las cargas del varifactor 1 en diferentes meses del año (meses de caudales altos y bajos), por lo cual existe sobre estas estaciones la influencia tanto del aumento de los sedimentos por el aumento del caudal, como de las actividades industriales, puesto que recorren zonas de gran actividad industrial (Ver Figuras N° 1 y 2 del Anexo 4). Mientras que el resto de las estaciones de muestreo, el agua recorre pocas zonas de actividad industrial (Ver

Figuras N° 3 a 6 del Anexo 4). No obstante, los mayores valores de las cargas del varifactor 1 sobre estas estaciones restantes de muestreo se da en épocas de Diciembre a Febrero (meses de lluvia y de caudales altos), mientras que los menores valores se dan en meses de caudales bajos (setiembre).

Por lo tanto la principal influencia para el aumento de los metales en las estaciones es el aumento de los sólidos suspendidos y la mezcla de estos con los metales de las actividades mineras de la cuenca alta del río Rímac y el lecho del río debido al aumento del caudal.

La Figura N° 4-8 (b) muestra que:

Las estaciones de muestreo que tuvieron un rango mayor de influencia de origen geológico y de la contaminación de las actividades urbanas o domesticas, esto es, un rango de mayores puntuaciones del varifactor 2 fueron: Gambeta, Ricardo Palma, y de menor influencia las estaciones Del Ejército, Los Angeles, Huampaní, Huachipa, Santa Eulalia. Esto debido a que el agua que llega la estación Gambeta tiene un contacto más directo con distritos de mayor población urbana (Cercado de Lima, San Martin de Porres, Carmen de la Legua y Reynoso y el y El Agustino) Callao) en comparación con la estación Del Ejército (Rímac, Cercado de Lima, San Juan de Lurigancho), el cual recibe influencia de las aguas contaminadas de la quebrada Jicamarca. En ambas estaciones, hay un aumento de la influencia del varifactor 2 en temporadas de bajo caudal, el cual permite que los efluentes domésticos no tengan una mayor distribución a lo largo del río.

Mientras que el resto de las estaciones de muestreo, tienen áreas con una menor cantidad de población ribereña en comparación con las estaciones que se encuentra en la urbe de Lima. Por otro lado, la estación Ricardo Palma tiene una mayor influencia de la geología y actividades mineras del río que la actividad doméstica. Esta influencia es producto de los efluentes mineros y de los pasivos mineros ambientales ubicados en la cuenca alta del río Rímac (48), el cual hay una mayor influencia en temporadas de caudales bajo, el cual permite un menor flujo de agua y por ende una menor distribución de los parámetros del varifactor 2 en el agua.

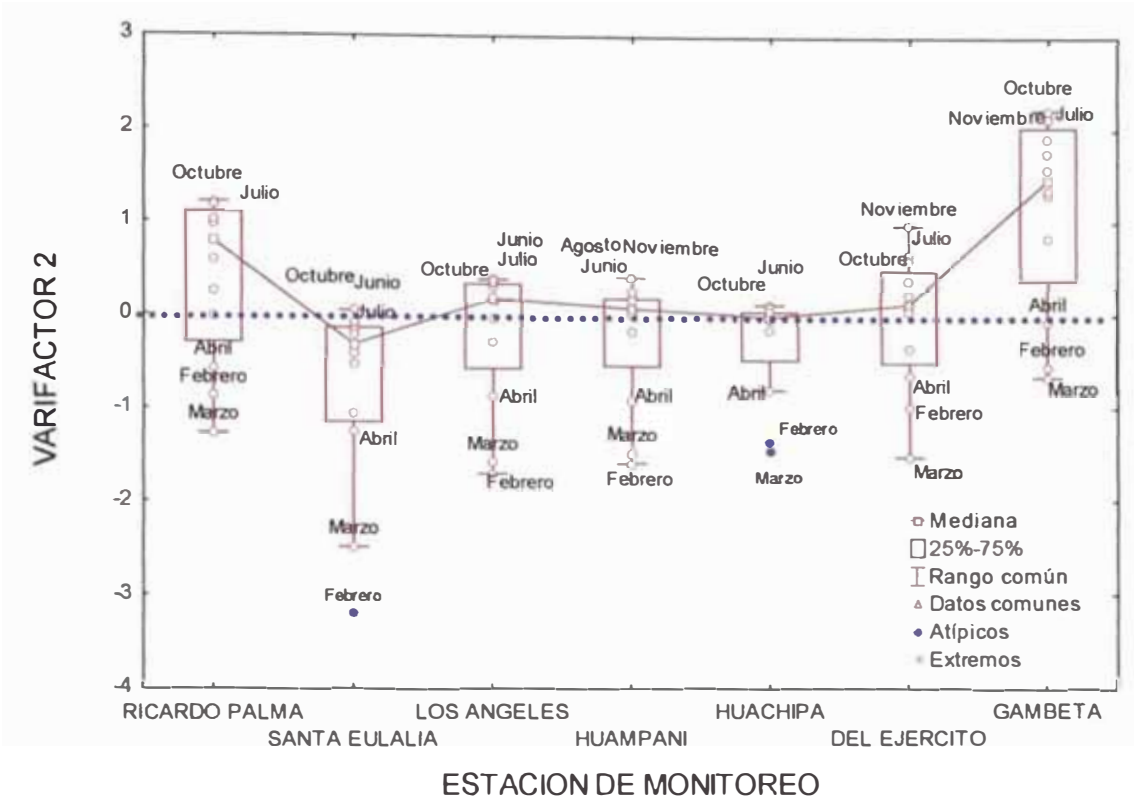

Figura N° 4-8. (b) Gráfico de cajas y bigotes de las puntuaciones del varifactor 2 (rotación varimax) correspondientes a las estaciones de muestreo.

La Figura N° 4-9 (a) muestra que:

Las estaciones de muestreo que tuvieron un rango mayor de influencia de los metales tóxicos (Cd y Pb), esto es, un rango de mayores puntuaciones del varifactor 3 fueron: Ricardo Palma, Santa Eulalia, Los Angeles, Huachipa y Huampaní (estaciones más alejadas de la ciudad de Lima), y de menor influencia la estación Gambeta y Del Ejército (ambas estaciones se encuentran dentro de la ciudad de Lima): La influencia de estos metales es variable y se da tanto en meses de caudales altos y bajos (Enero y Julio).

Sin embargo, las concentraciones del Cd y Pb, las cuales provienen de las descargas de las actividades mineras (1), son bajas en todas las estaciones de muestreo (Tabla N° 4-1), siendo ligeramente mayor en las estaciones que se encuentra en la cuenca alta del río Rímac, debido principalmente a la influencia de la actividad minera a saber, Ricardo Palma y Los Angeles, por lo cual estos parámetros no influyen de manera considerable sobre la calidad del agua del río Rímac.

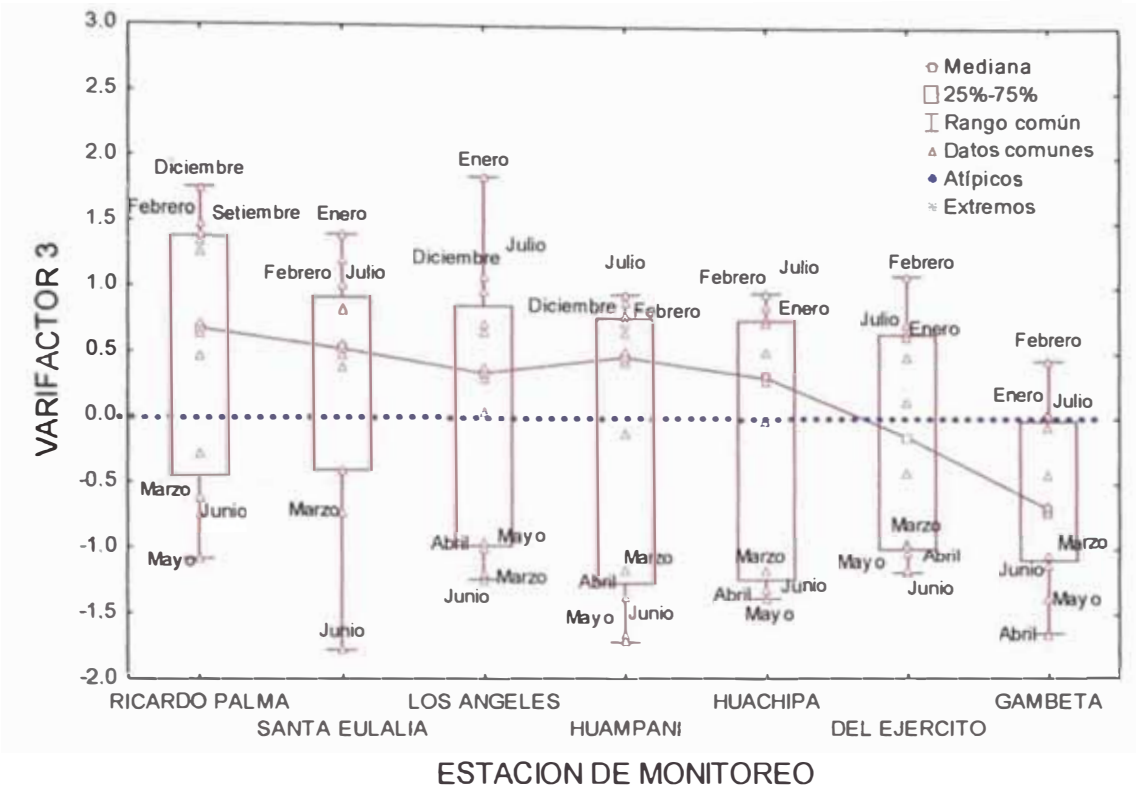

Figura N° 4-9. (a) Gráfico de cajas y bigotes de las puntuaciones del varifactor 3 (rotación varimax) correspondientes a las estaciones de muestreo.

La Figura N° 4-9 (b) muestra que:

Las estaciones de muestreo que tuvieron un rango mayor influencia de origen agrícola, esto es, un rango de mayores puntuaciones del varifactor 3 fueron: Huampaní, Los Angeles, Del Ejercito, de influencia media las estaciones de muestreo Ricardo Palma y Huachipa, y de menor influencia las estaciones Santa Eulalia y Gambeta.

Así, la Figura N° 4 del Anexo 4 mostró que las pequeñas zonas agrícolas cercanas a las estaciones de muestreo Los Angeles y Huampaní son la fuente de influencia del factor agrícola. Esto es debido al fácil acceso de las escorrentías agrícolas al agua del río Rímac. Mientras que la Figura N° 6 mostró que la estación de muestreo Del Ejército está influenciada de las amplias zonas agrícolas que se encuentran en la quebrada de Jicamarca.

Por otro lado, la Figura N° 3 y 5 del Anexo 4 mostraron que las estaciones Huachipa y Ricardo Palma tienen amplias zonas agrícolas. Sin embargo, la influencia es menor debido a la distancia de estas zonas y el difícil acceso de las escorrentías por las construcciones y carreteras que se encuentran alrededor del río.

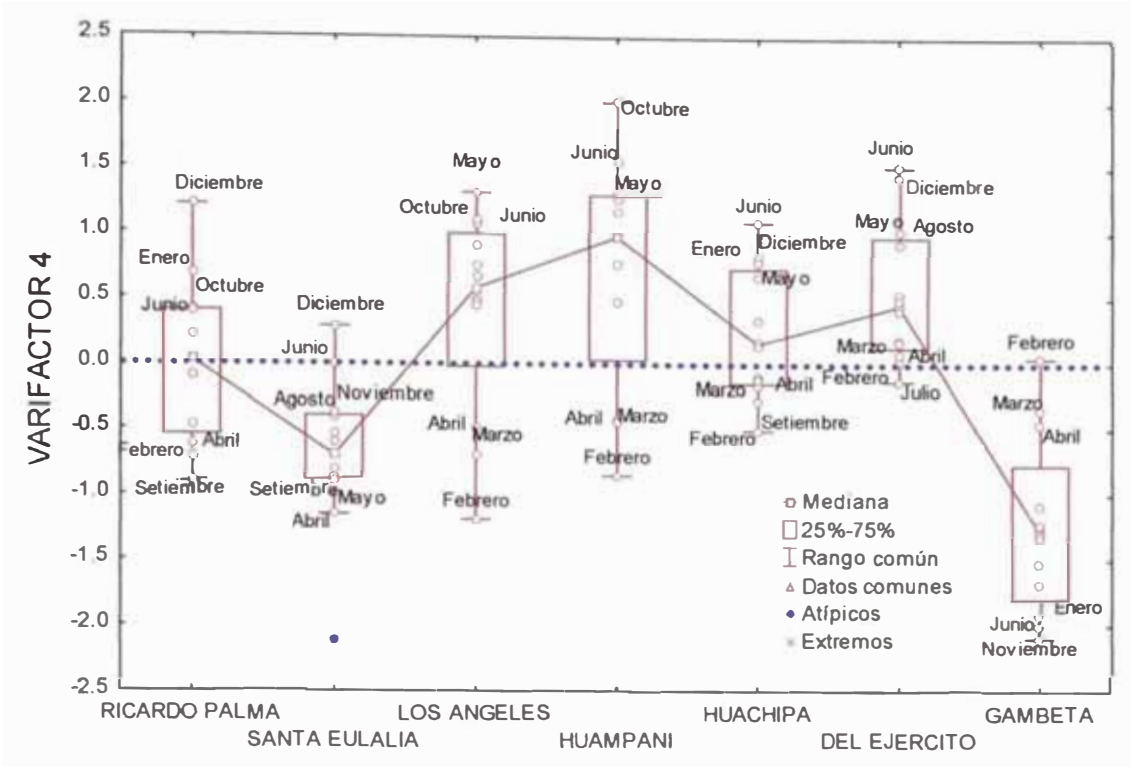

Figura N° 4-9. (b) Gráfico de cajas y bigotes de las puntuaciones del varifactor 4 (rotación varimax) correspondientes a las estaciones de muestreo.

Mientras que la Figura N° 6 del Anexo 4 mostró que la estación Santa Eulalia tiene muy pocas zonas agrícolas y la Figura N° 1 muestra que la estación Gambeta tiene una amplia zona agrícola cercana, sin embargo tiene poca cantidad de oxígeno disuelto debido a que sus aguas contienen grandes cantidades de materia orgánica y componentes microbiológicos (coliformes y bacterias) que consumen grandes cantidades de oxigeno, el cual es de importancia para oxidar los compuestos reducidos del nitrógeno como el nitrógeno amoniacal a nitritos y nitratos (30). La influencia de esta influencia antropogénica es variable y se da tanto en meses de caudales altos y bajos.

4.2.4 Análisis discriminante (AD)

Luego de realizar la clasificación de las estaciones de muestreo de acuerdo a los parámetros desarrollados, así como agrupar y relacionar estos parámetros, se realiza el método supervisado llamado análisis discriminante (AD). Este método, con un conocimiento a priori de la pertenencia de cada estación de muestreo a lo largo del período de análisis a una clase particular y a través de la clasificación Bayesiana, calcula el porcentaje de eficiencia de la clasificación de cada una de las estaciones de muestreo,

llevadas a cabo por los análisis de clúster (AC) y componentes principales (ACP) (15; 42) y determinar el número de parámetros mas discriminantes y/o óptimos para realizar tal clasificación y evaluar la calidad del agua.

La tabla de clasificación bayesiana del análisis discriminante, Tabla N° 4-8, representa las clasificaciones de las estaciones de muestreo, tanto observadas en las columnas y asignadas por el AD en las filas. Los valores entrantes representa el número de estaciones de muestreo que fueron clasificadas de acuerdo al AD.

Si la clasificación de cada una de las estaciones está bien asignada, entonces los valores se encuentran en la diagonal principal de la tabla, quedando fuera de la diagonal principal el número de estaciones que fueron asignadas a otra clase de estación (7; 25). De esta manera, se visualizó que la tabla de clasificación determinó correctamente el 94 % de las estaciones de muestreo evaluadas para todos los datos transformados de los 20 parámetros.

Tabla N° 4-8. Tabla de clasificación de las estaciones de muestreo para los 20 parámetros del agua del río Rímac

Clase	PORCENTAJE (%)	RP p=0.14286	SE p=0.14286	Grupo I p = 0.42857	DE p=0.14286	GA p=0.14286
RP	100.0	12	0	0	0	0
SE	91.7	0	11	1	0	0
Grupo I	97.2	0	1	35	0	0
DE	83.3	0	0	2	10	0
GA	91.7	0	0	0	1	11
Total	94.0	12	12	38	11	11

*Filas: clasificaciones observadas, columnas: clasificaciones asignadas.
Estaciones de muestreo: RP (Ricardo Palma), SE (Santa Eulalia), DE (Del Ejercito), GA (Gambeta), Grupo I: Los Angeles, Huampaní, Huachipa.

Para determinar los parámetros más discriminantes para la clasificación realizada por el AC y ACP, los datos fueron sujetos a los modos discriminantes estándar, forward y backward (Tabla N° 4-9) (15). Para llevar a cabo estos métodos de reducción discriminante, se calcula el valor del estadístico F parcial de cada parámetro ambiental, el cual muestra la significancia de cada uno de los parámetros después de la separación de la matriz de datos en relación a la clasificación de las estaciones de muestreo.

Mientras la lambda de Wilks (λ de Wilks) es un estadístico multivariado para caracterizar la prueba F multivariada de la clasificación de las estaciones de muestreo, y así contrastar la hipótesis nula (H_o) de que todas las estaciones están clasificadas de acuerdo a los análisis de clúster y componentes principales. Este valor toma valores entre 0 a 1 de forma que, cuanto más cerca este de 0, mayor es el poder discriminante de los parámetros consideradas y cuanto más cerca este de 1 menor es dicho poder de los parámetros dadas sobre el análisis espacial del río Rímac.

Tabla N° 4-9. Modos discriminantes de reducción de los parámetros del agua de rio Rimac

PARAMETRO	MODO ESTANDAR λ de Wilks : 0.00789		MODO FORWARD λ de Wilks : 0.00999		MODO BACKWARD λ de Wilks : 0.03332	
	F de salida (4,60)	p-valor	F de salida (4,64)	p-valor	F de salida (4,74)	p-valor
T	2.50	0.05	***2.65***	***0.041***		
pH	***3.22***	***0.02***	2.46	0.054		
CE	***2.91***	***0.03***	***3.28***	***0.016***		
SST	0.27	0.90	0.62	0.650		
ST	***3.28***	***0.02***	2.29	0.069		
NO_3-N	***5.37***	***0.00092***	***4.91***	***0.0016***	**9.097**	**0.0000**
NO_2-N	***7.95***	***0.00003***	***7.99***	***0.00002***	**16.11**	**0.0000**
SO_4	1.36	0.26	1.45	0.23		
PO_4-P	***6.32***	***0.00026***	***7.02***	***0.00009***	**7.14**	**0.0000**
Cl	***4.04***	***0.00572***	***4.61***	***0.00246***	**19.88**	**0.0000**
Alc	***2.56***	***0.047***	1.37	0.25		
DT	***4.09***	***0.00539***	***3.60***	***0.01054***	**15.86**	**0.0000**
DCa	1.30	0.28	0.88	0.48		
Fe	2.50	0.052	***3.43***	***0.01337***	**18.26**	**0.0000**
Pb	1.40	0.25				
Zn	0.99	0.43				
Cu	1.54	0.20	1.84	0.13		
Cr	***2.88***	***0.03***	2.07	0.095		
Al	0.59	0.67				
Cd	1.35	0.26				

En el modo estándar se incluye todos los parámetros, calculándose los valores de F de salida y el respectivo valor de probabilidad (p-valor) de cada parámetro ambiental,

mientras que el modo forward cada parámetro es incluido comenzando con el más discriminante o que tenga el valor de F de entrada más alto hasta que no existan cambios significantes en la λ de Wilks. El modo backward elimina los parámetros menos discriminantes o que tenga el valor de F de salida más pequeño, hasta que no se obtiene algún cambio significante la λ de Wilks. Los parámetros de cada modo discriminante fueron analizados bajo el valor de su F de salida y sus respectivos p-valor (Tabla 4-9). De esta manera, tomando en cuenta que a un valor F de salida pequeña y con un respectivo p-valor mayor de 0.05 se rechaza las variables menos discriminantes hasta obtener finalmente las variables más discriminantes.

El modo estándar determinó que los parámetros resaltados, a saber, pH, CE, ST, NO_3–N, NO_2–N y PO_4–P, Cl, Alc, DT, Cr son los parámetros mas discriminantes tomando en cuenta todos los 20 parámetros para la clasificación de las 7 estaciones de muestreo. El modo forward determinó que los parámetros resaltados, a saber, T, CE, NO_3–N, NO_2–N y PO_4–P, Cl, DT y Fe son los parámetros más discriminantes entre los 16 parámetros. Asimismo, el modo backward determinó que los parámetros NO_3–N, NO_2–N y PO_4–P, Cl, DT y Fe son las parámetros más discriminantes. Por lo cual en cada modo de análisis registro diferentes grupos de parámetros más discriminantes.

Teniendo en cuenta el mecanismo de análisis de los modos discriminantes, se optó por el grupo de parámetros del modo Backward, por tener la menor cantidad de parámetros. De esta manera, el AD sugiere que el NO_3-N, NO_2-N, PO_4-P, Cl, DT y Fe son los parámetros mas discriminantes para la clasificación de las estaciones de muestreo en cinco clases.

Este grupo de parámetros representan a la mayoría de los aniones que se encuentran en forma disuelta y a uno de los metales más abundantes de la tierra que forma compuestos insolubles de hidróxido el cual enlaza a otros metales en condiciones oxidantes en el agua superficial del río. Estos parámetros provienen de las diferentes actividades antropogénicas, influencia geológica y los cuales son variables con respecto

a las condiciones del clima (temperatura y caudal), por lo cual representan de buena manera a los demás parámetros.

La Tabla N° 4-10 visualizó la tabla de clasificación que clasifica correctamente el 82.1% de las estaciones de muestreo evaluadas para todos los datos transformados y autoescalados, tomando en cuenta los 6 parámetros mas discriminantes calculados por el modo Backward.

Tabla N° 4-10. Tabla de clasificación de las estaciones de muestreo para los 6 parámetros mas discriminantes del agua río Rímac (Junio 2008–julio 2009)

CLASE	PORCENTAJE (%)	RP p=0.14286	SE p=0.14286	Grupo I p = 0.42857	DE p=0.14286	GA p=0.14286
RP	83.3	10	0	2	0	0
SE	83.3	0	10	2	0	0
Grupo I	86.1	0	2	31	3	0
DE	75.0	0	0	3	9	0
GA	75.0	0	0	0	3	9
Total	82.1	10	12	38	15	9

*Filas: clasificaciones observadas, columnas: clasificaciones asignadas.
Estaciones de muestreo: RP (Ricardo Palma), SE (Santa Eulalia), DE (Del Ejercito), GA (Gambeta), Grupo I: Los Angeles, Huampaní, Huachipa.

Se evaluó la significancia de las funciones discriminantes para los 6 parámetros más discriminantes (Tabla N° 4-11). Estas funciones de manera similar al análisis de componentes principales, muestra las relaciones entre las estaciones de muestreo de acuerdo a una clasificación conocida (clasificación realizada por el AC y ACP) en un espacio reducido de funciones discriminantes.

Tabla N° 4-11. Funciones discriminantes para los 6 parámetros más discriminantes del agua del río Rímac

FUNCION DISCRIMINANTE	AUTOVALOR	% VARIANZA EXPLICADA	LAMBDA DE WILKS	χ^2	gl	p-valor
1	4.21	61.66	0.03	263.6	24	0.000000
2	1.43	20.96	0.17	135.8	15	0.000000
3	0.99	14.64	0.42	66.97	8	0.000000
4	0.19	2.74	0.84	13.30	3	0.004027

Todas las funciones discriminantes tienen un p-valor menor de 0.05 del respectivo valor estadístico del chi-cuadrado (χ^2), el cual indica que estas funciones discriminantes son

estadísticamente significantes a un 95 % de nivel de confianza. Estas funciones representan toda la información entre las relaciones de las estaciones de muestreo (15). Sin embargo, se tomó en cuenta las tres primeras funciones discriminantes, los cuales explican el 97.26% de la varianza acumulada de discriminación.

Graficando el espacio tridimensional de los ejes principales a estas funciones (Figura N° 4-10) se muestra las medias de las puntuaciones de las estaciones de muestreo. El uso de la media permite una interpretación rápida y simplificación de la gráfica (57). Por lo tanto, las estaciones de muestreo fueron divididas en el Grupo I, formado por las estaciones Los Ángeles, Huampaní y Huachipa, estando las demás estaciones de muestreo no agrupadas.

De esta manera se comprueba mediante el gráfico de las tres funciones discriminantes, la buena discriminación que tienen los 6 parámetros encontrados por el AD sobre el análisis espacial del agua del río Rímac. Estos parámetros podrían considerarse prioritarios en los posteriores monitoreos para evaluar la calidad ambiental del agua del río Rímac en las estaciones de muestreo, puesto que se reducen el número de parámetros que se puedan considerar, optimizando costos y tiempo en los análisis del laboratorio y teniendo la misma interpretación de los resultados.

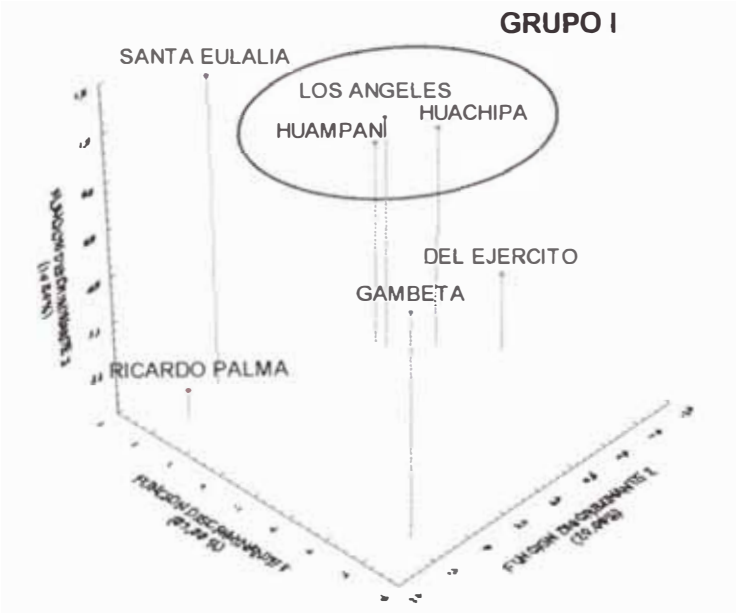

Figura N° 4-10. Gráfico de dispersión de las medias de las puntuaciones de las tres primeras funciones discriminantes para las estaciones de muestreo del río Rímac (Junio 2008–Julio 2009).

CONCLUSIONES

Este trabajo muestra que el conjunto de las técnicas multivariadas, también conocidas como técnicas de reconocimientos de patrones, es un instrumento eficiente para el análisis de los parámetros y estaciones de muestreo del programa de monitoreo del agua del rio Rímac. De esta manera, los resultados mostraron que:

Las estaciones de muestreo del Grupo I: Los Angeles, Huampaní y Huachipa, mediante el análisis de clúster (AC) y el análisis de componentes principales (ACP), se agruparon de acuerdo a las semejanzas en la composición fisicoquímica de sus aguas, mientras las demás estaciones de muestreo no estuvieron agrupadas. Estos resultados nos permiten desarrollar la técnica supervisada, a saber, Análisis Discriminante.

Asimismo, los parámetros se agrupan de acuerdo a sus características y relaciones fisicoquímicas, a saber, metales no esenciales para la vida acuática (clase I); metales de mayor concentración y esenciales a nivel traza (clase II); componentes del ciclo del nitrógeno (clase III); componentes de la conductancia eléctrica del Río Rímac (clase IV).

El análisis de factor (AF) permitió designar a las fuentes de contaminación principal en el río Rímac, a saber, urbana y geológica, industrial y minera, agrícola y geológica se relacionen con los grupos de parámetros ambientales evaluados Clase IV, Clase I y II, Clase III respectivamente. Por otro lado, cada una de las fuentes de contaminación influenció de un modo particular a cada estación de muestreo de acuerdo a las características geológicas e influencias de las actividades antropogénicas del lugar de muestreo.

De esa manera, la estación Ricardo Palma tuvo mayores influencias de las actividades mineras y la geología del área, Santa Eulalia pocas influencias de todas las actividades antropogénicas y naturales, Los Angeles y Huampaní de mayores influencias agrícolas y la geología del área, Huachipa de relativa influencia agrícola, Del Ejercito de mayores influencias de las actividades agrícolas e industriales; y Gambeta de mayores influencias de las actividades urbanas e industriales.

Con los resultados de la clasificación de las estaciones de muestreo por el AC y ACP, se realizó el análisis discriminante (AD) para evaluar la correcta clasificación de cada una de las estaciones de muestreo, con lo cual el 94 % de las estaciones de muestreo fueron bien asignadas, lo cual permite considerar el buen desarrollo de los métodos no supervisados (AC y ACP).

Por otro lado, el AD ha servido para optimizar que parámetros nos puede evaluar la calidad del agua, de esa manera se redujo considerablemente el número de parámetros ambientales evaluados, a saber, NO_3-N, NO_2-N, PO_4-P, Cl, DT, Fe, los cuales representan al resto de parámetros ambientales por provenir de diferentes actividades antropogénicas e influencias naturales que afectan al río Rímac.

Por lo tanto, los análisis multivariados evaluados tomaron en cuenta la existencia de las agrupaciones y relaciones en las estaciones de muestreo y parámetros ambientales, disminuyendo el número de estos factores, el cual permitirá la optimización del tiempo y costo de los posteriores programas de monitoreo y evaluación de las influencias en la calidad del agua del río Rímac.

REFERENCIAS BIBLIOGRAFICAS

1. Manahan, S.E., "*Introduccion a la quimica ambiental*", España, Reverté, 2007. 725 p.

2. Dickson, T.R., "*Química*", México, Limusa, 2005.

3. Quevauviller, P., "*Quality Assurance for water analysis*", England, John Wiley & Sons, 2002, 1a ed., 262p

4. Sawyer, C.N., McCarty, P.L., Parkin, G.F., "*Quimica para ingenieria ambiental*", Colombia, Mc Graw Hill, 4a ed., 2001

5. Albert, L.A., "*Curso básico de toxicología ambiental*", Argentina, Limusa-Wiley, 2a ed., 2002.

6. Alloway, B.J. y Ayres, D.C., "*Chemical Principles of environmental pollution*", London, Chapman & Hall, 1997, 2ea ed., 395 p.

7. Shrestha, S., Kazama, F., "*Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river Basin, Japan*", Environmental Modelling & Software, 22, (2007), 464-475.

8. Tecsup, "*Monitoreo y análisis de aguas, programa de capacitación continua*", Peru, 2007.

9. Reeve, R., "*Introduction to environmental analysis", England,* John Wiley & Sons, 2002, 2a ed., 301p

10. Luis Antonio Tortajada Genaro, "*Soluciones quimiométricas para optimizar el análisis de parámetros químicos en aguas*", España, Universidad de Valencia, Departamento de química analítica, 2002.

11. Spiro, T.G., Stigliani, W.M., "*Quimica medioambiental*", España, Pearson Education, 2a ed., 2004.

12. Domenech, X.y Peral, J., "*Quimica ambiental de sistemas terrestres*", España, Reverté, 2006, 1a ed., 256 p.

13. El Sharaawi, A.H., Piegorsch, W.W, "*Encyclopedia of ENVIRONMETRICS*", England, John Wiley & Sons, Volumen 1, 1a ed., 2002, 2800 p.

14. Arriola, J.F., Pepper, I.L and Brusseau, M., "*Environmental monitoring and characterization*", China, Elsevier academic Press, 1a ed., 2004

15. Kannel, P.R., et.al, "*Chemometric application in classification and assessment of monitoring locations of an urban river system*", Anal. Chim. Acta, 582, 2007, 390-399.

16. Einax, J.W, Zwanzinger, H.W, Geib, S., "*Chemometrics in Environmental Analysis*", Germany, Wiley-VCH, 1a ed., 1997.

17. APHA-AWWA-WEF, "*Standard Methods for the Examination of Waster and Waterwater*", American Public Health Association, Washington D.C., 21a ed., 2005.

18. Barnett, V., "*Environmental Statistics: Methods and Applications*", John Wiley & Sons, England, 2004.

19. Berthouex, P.M., Brown, L.C., "*Statistical for Environmental Engineers*", Florida Boca Raton, 2a ed., 2002.

20. Otto, M., "*Chemometrics: Statistics and computer Application in analytical Chemistry*", Germany, Wiley-VCH, 2007.

21. Ramis Ramos, G., "*Quimiometria*", España, Síntesis, Vol. 1., 2001.

22. Varmuza, K., Filzmoser, P., "*Introduction to Multivariate Statistical Analysis in Chemometrics*", USA, Taylor & Francis Group, LLC, 2008.

23. Brereton, R., "*Chemometrics: Data Analysis for the laboratory and Chemical Plant*", England, John Wiley & Sons, 2003.

24. Beebe, K.R., Pell, K.J., Seasholtz, M.B., "*Chemometrics: A practical Guide*", USA, John Wiley & Sons, 1a ed., 1998.

25. Gotelli, N.J., Ellison, A.M., "*A primer of Ecological Statistics*", USA, Sinauer Associates, 1a ed., 2004.

26. Render, A., "*Methods of Multivariate Analysis*", USA, John Wiley & Sons, 2a ed., 2002.

27. Pemperline, P., "*Practical Guide to Chemometrics*", USA, Taylor & Francis Group, LLC, 2a ed., 2006.

28. Zhou, F., Guo, H., Liu, Y., et al., "*Chemometrics data analysis of marine water quality and source identification in Southern Hong Kong*", Marine Pollution Bolletin, 54, (2007), 745-756.

29. Li, R., Dong, M., Zhao, Y., et al., "*Assessment of Water Quality and Identification of Pollution Sources of Plateau Lakes in Yunnan (China)*", J. Environ. Qual., 36, 2007, 291-297.

30. Chandra Pandra, U., Kumar Sundaray, S., Rath, P., et al., "*Application of factor and clúster analysis for characterization of river and estuarine water systems – A case study: Mahanadi River (India)*", Journal of Hydrology, 331, 2006, 434 445.

31. Kumar, A.R, Riyazuddin, P., "*Application of chemometric techniques in the assessment of groundwater pollution in a suburban area of Chennai city, China*", Current Science, 94, 2008, 1012-1022.

32. Singh, K.P., Malik, A., Mohan, D., et al., "*Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) – a case study*", Water Res., 38, (2004), 3980-3992.

33. Ferreira, M., Gomes F., C., Paes, E., "*Oceanographic characterization of northern Sao Paulo Coast: a chemometric study*", Chemometric and intelligent laboratory systems, 47, 1999, 289-297.

34. Morales, M.M., et.al, "*An environmental study by factor analysis of surface seawaters in the Gulf of Valencia (Western Mediterranean)*", Anal. Chim. Acta, 1999, 394, 109-117.

35. Principi, P., et.al, "*Metal toxicity in municipal wastewater activated sludge investigated by multivariate analysis and in situ hybridization*", Water Res., 40, 2006, 99-106.

36. Simeonov, V., et.al, "*Multivariate statistical study of simultaneously monitored cloud water aerosol and rainwater data from different elevation levels in an alpine valley (Achenkirch, Tyrol, Austria)*", Talanta, 61, 2003, 519-528.

37. Kazi, T.G., et.al, "*Assessment of water quality of polluted lake using multivariate statistical techniques : A case study*", Ecotoxicology and Environmental Safety, 72, 2009, 301-309.

38. Pantsar-Kallio, M., Mujunen, S., Hatzimihalis, G., et al., "*Multivariate data analysis of key pollutants in sewage samples: a case study*", Anal. Chim. Acta, 393, (1999), 181-191.

39. StatSoft, Inc. (2007). STATISTICA (data analysis software system), version 8.0. www.statsoft.com.

40. Tabachnick, B.G., Fidell, L.S., "*Using Multivariate Statistics*", USA, Allyn & Bacon, 5ta ed., 2006.

41. Chen, K., et.al, "*Multivariate statistical evaluation of trace elements in groundwater in a coastal area in Shenzhen, China*", Environmental Pollution, 147, 2007, 771-780.

42. Kowalkowski, T., et.al, "*Application of chemometrics in river water classification*", Water Research, 40, 2006, 744-752.

43. Reghunath, R., Sreedhara Murthy, T.R., Raghavan, B.R., "*The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India*", Water Res., 36, 2002, 2437-2442.

44. Hardle, W., Simar, L., "*Applied Multivariate Statistical Analysis*", New York, Springer, 1era ed., 2003.

45. Mencio, A., Mas-Pla, J., "*Assessment by multivariate analysis of groundwater-surface interactions in urbanized Mediterranean streams*", Journal of Hidrology, 352, 2008, 355-366.

46. Miller, N.J., Miller, J.C., "*Estadística y Quimiometria para Quimica Analitica*", Madrid, Prentice Hall, 4a ed., 2002, 296 p.

47. Ministerio de Energia y Minas, Direccion General de Estudios Ambientales, "*Evaluacion ambiental territorial de la cuenca del río Rímac*", Peru, MINAM; 1997, 48.

48. Juarez S., H.R., "*Contaminacion del río Rímac por metales pesados y efecto en la agricultura en el cono este de Lima Metropolitana*", Peru, Universidad Nacional Agraria La Molina, Escuela de Post Grado, 2006

49. Ministerio de salud, Dirección General de salud ambiental, Sedapal. "*Rio Rímac*", Peru, MINSA/DIGESA, 2008, 37.

50. Ministerio de salud, Dirección General de salud ambiental, "*Rio Rímac*", Peru, MINSA/DIGESA, 2008, 23.

51.Peré-Trepat, E., et.al, "*Chemometrics modeling of organic contaminants in fish and sediment river samples*", Science of Total Environment, 371, 2006, 223-237.

52. Zeng, X., Rasmussen, T.C., "*Multivariate Statistical Characterization of Water Quality in Lake Lanier, Georgia, USA*", J. Environ. Qual., 34, 2005, 1980-1991.

53. Marengo, E., et.al, "*Statistical analysis of ground water distribution in Alessandria Province (Piedmont-Italy)*", Microchemical Journal, 88, 2008, 167-177.

54. Vega, M., Pardo, R., Barrado, E., et al., "*Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis*", Water Res., 32, (1998), 3581-3592.

55. Feng, Z., Huai-Cheng, G., Yong, L., Ze-jia, H., "*Identificacion and spatial patterns of coastal water pollution sources based on GIS and chemometric approach*", Journal of Environmental Sciences, 19, 2007, 805-810.

56. Hanrahan, G., Gibani, S., Miller, K., "*Multivariate chemometrical classification and assessment of Lake Tuendae: A mojave desert aquatic Environment housing the endangered Mojave Tui Chub*", Ecological Informatics, 3, 2008, 334-342.

57. Aruga, R., et.al, "*Pollution of a river basin and its evolution with time studied by multivariate statistical analysis*", Anal. Chim. Acta, 310, 1995, 15-25.

58. Cristina Elizabeth Toro Vílchez, "*Estudio del método analítico de determinación espectrofotométrica de nitratos y nitritos, su control mediante técnicas de validación y aplicación en la evaluación de la contaminación de agua superficial de río*", Perú, Universidad Nacional de Ingeniería, Facultad de Ciencias, 2002.

59. Rivera, H., Chira, J., Zambrano, K., et al., "*Dispersión secundaria de los metales pesados en sedimentos de los ríos Chillón, Rímac y Lurín Departameto de Lima*. Inst. Investig. Fac. minas metal cienc. Geogr, 20, (2007), p.19-25.

60. Baird, C., "*Química ambiental*", España, Reverté, 2a. ed., 2001, 622 p.

GLOSARIO DE TERMINOS

AMBIENTE NATURAL: Ambiente conformado por todo aquello que no ha sido creado ni modificado por el ser humano.

ANTROPOGENICO: Cualquier actividad o evento realizado por el hombre.

AUTODEPURACION: Es la capacidad que tiene el agua, el cual recibe o ha recibido una carga contaminante, de recuperar las condiciones fisicoquímicas y biológicas previas a su contaminación.

CAUCE: Parte del fondo de un valle por donde discurren las aguas de un curso. Es el límite físico normal de un flujo de agua, siendo sus confines laterales las riberas.

CORTEZA TERRESTRE: Es la piel exterior de la Tierra, que es accesible a los seres humanos. Es sumamente delgada comparada con el diámetro de la tierra y mide entre 5 a 40 K de espesor.

CICLO BIOGEOQUIMICO: Se describen como ciclos elementales que involucran a los elementos esenciales como el carbono, el nitrógeno, el oxigeno, el fosforo y el azufre.

CURTOSIS: Una medida de la dispersión o concentración de una distribución normal relativo a su centro. Este es calculado a partir del cuarto momento central.

DENDOGRAMA: Resultado gráfico del análisis de clúster, en forma similar al árbol de clasificación, el cual es también ampliamente usado para ilustrar las relaciones entre las especies.

ECOSISTEMAS ACUATICOS: Conjunto de seres vivos y elementos inertes que se relacionan entre sí en el medio acuático.

EFLUENTES: Descargas al ambiente de contaminantes en su estado natural o tratados parcial o totalmente o de aguas residuales a cualquier ambiente acuático.

ESCORRENTIAS: Es la lamina de agua que circula sobre la superficie en una cuenca de drenaje, es decir la altura en milímetros del agua de lluvia escurrida y extendida.

ESFERICIDAD: La asunción de los métodos de las pruebas multivariadas, y el cual indica que las matrices de covarianzas de cada grupo son iguales entre ellos.

ESTADISTICA DESCRIPTIVA: La estadística descriptiva es una parte de la estadística que describe los datos en términos de variables o combinaciones de variables. Este análisis es muy básico pero fundamental en todo estudio

ESTADISTICA INFERENCIAL: Es una parte de la estadística que comprende los métodos y procedimientos para deducir propiedades (hacer inferencias) de una población, a partir de una pequeña parte de la misma (muestra).

HIDROSFERA: Es un subsistema formado por todo el agua que contiene el planeta en su superficie, su interior y su atmosfera. El agua se halla distribuida formando lagos, mares, ríos, océanos, glaciares, vapor de agua, etc.

HOMOCEDASTICIDAD: Es una propiedad fundamental del modelo de regresión lineal general y está dentro de sus supuestos clásicos básicos. Esta se basa en que las varianzas residuales de todas las clases de objetos son iguales.

INTEMPERISMO: Es la descomposición superficial de las rocas, el desgaste físico y alteración química de rocas y minerales en o cerca de la superficie de la Tierra.

LECHO (Río): Es la parte más excavada de los valles o las depresiones drenadas. Es el órgano elemental de las corrientes de agua.

MANEJO SOSTENIBLE: Es el manejo sobre cualquier ambiente que satisfaga las necesidades del presente sin poner en peligro la capacidad de las generaciones futuras para atender sus propias necesidades.

MATRIZ DIAGONAL: Una matriz cuadrada en el cual todos los valores excepto aquellos que están en la diagonal principal son iguales a cero.

MATRIZ TRANSPUESTA: Es una matriz de modo que sus columnas son iguales a las filas de la matriz original y las columnas de la matriz original igual al número de filas de la matriz.

MODELOS DETERMINISTICOS: Corresponden a modelos matematicos diseñados bajo el supuesto que el resultado de un experimento queda determinado por las condiciones bajo los cuales se realiza.

MODELOS ESTOCASTICOS: Son aquellos en el cual los modelos determinísticos no son adecuados, ya que el resultado no es predecible

NIVEL DE CONFIANZA: Un intervalo que incluye la media verdadera de la población en el n% del tiempo. El término de intervalo de confianza siempre es calificado con el 95%. Los intervalos de confianza son calculados usando estadística frencuentista

PECUARIO: Perteneciente o relativo al ganado.

PM_{10}: Partículas sólidas o líquidas cuyo diámetro varían entre 2,5 y 10 micrómetro (µm)

$PM_{2,5}$: Partículas sólidas o líquidas cuyo diámetro es menor a 2,5 micrómetro (µm)

PRUEBA F: Prueba estadística del análisis de la varianza multivariada.

RESERVORIO: Un embalse de agua almacenado en un valle interceptado por una presa.

SESGO: Una descripción del alejamiento de la simetría de una distribución normal, el cual es calculado a partir del tercer momento central.

TEOREMA DEL LIMITE CENTRAL: Indica que, en condiciones muy generales, la distribución de la suma de variables aleatorias tiende a una distribución normal (también llamada *distribución gaussiana* o *curva de Gauss*) cuando la cantidad de variables es muy grande. Así, el teorema permite que cualquier variable aleatoria pueda ser transformada a una variable aleatoria normal.

TRIBUTARIO: Un curso de agua que no desemboca en el mar si no en otro río más importante con el cual se une en un lugar llamado confluencia.

VALOR DE PROBABILIDAD (P-Valor): Valor de la prueba estadística frencuentista a un % de confianza.

VECTOR: Agente que transporta algo material de un lugar a otro.