UNIVERSIDAD NACIONAL DE INGENIERÍA
Facultad de Ciencias
Escuela Profesional de Química

ESTUDIO QUIMIOMÉTRICO DE LOS CONTROLES DE AGUAS DE LA EMPRESA UNIPETRO ABC S.A.C.

TESIS

PARA OBTENER EL TÍTULO PROFESIONAL DE
LICENCIADO EN QUÍMICA

CARLOS ENRIQUE MINAYA AMES

LIMA-PERÚ
2005

A mi hijo Carlos Enrique,
fuerza impelente en mi vida

Agradecimiento

Quisiera que al abrir estas páginas se refleje mi agradecimiento, a aquellos, que de una u otra forma han permitido culminar este trabajo.

Inicio este listado con mi familia que es la parte mas importante de mi ser y que siempre se encuentran a mi lado, apoyándome y alentándome en todo momento para permitir mi desarrollo personal y profesional.

También tengo que agradecer a mi compañero y amigo Carlos Timaná de La Flor por su tiempo e impulso para que se concrete este proyecto y de lo bien que he pasado en todo este tiempo con su compañía.

A mi asesor Lic. Christian Jacinto Hernández por la confianza y apoyo que en todo momento me brindó.

Al Ing, Víctor Cataño Cauti y su esposa, Ing. Maritza Meza por darme las facilidades y confianza para poder lograr realizar la tesis con los datos de la Empresa UNIPETRO ABC S.A.C. y al Ing. Juan Peralta por su paciencia y buen humor, que en todo momento me mostró, y por proveerme toda la información necesaria en el transcurso de estos meses que se necesitaron para la culminación del proyecto.

No podría olvidarme de mi gran amigo Lic. Jorge Breña Oré que en todo momento estuvo en crítica constructiva para poderme dar cuenta de algunos detalles importante en el acabado de la tesis. Gracias colotordoc.

Y a todos mis amigos que estuvieron a mi lado en este trabajo.

RESUMEN

Los estudios de monitoreo ambiental producen una enorme cantidad de datos de concentración esparcidos en diversos sitios geográficos y durante diferentes periodos de tiempo. Todos estos valores son difíciles de abarcar y evaluar de un modo simple y rápido usando las herramientas estadísticas univariantes, debido especialmente a su gran número y a su correlación multivariante. Para descubrir patrones relevantes dentro de un gran número de datos multivariantes, se propone la aplicación de los métodos de la moderna Quimiometría, basados en el análisis de datos mediante la estadística multivariante. Después de aplicar los métodos quimiométricos, las fuentes de contaminación, puntuales y difusas, y su origen (natural, antropogénico, industrial, ...) se identifican y se evalúa su distribución entre las muestras. En cada sitio de muestreo, se estima una distribución o prorrateo de las diferentes fuentes de contaminación en el ambiente. En esta tesis, se probarán diferentes métodos quimiométricos en una serie de datos ambientales. En particular, se muestra la aplicación del análisis de componentes principales y de los métodos de clasificación multivariante como poderosas herramientas para lograr el objetivo del modelamiento quimiométrico de las fuentes de contaminación en un gran conjunto de datos ambientales adquiridos durante el monitoreo.

ABSTRACT

Environmental monitoring studies produce huge amounts of concentration values of chemicals spread at distant geographical sites and during different time periods. All these data values are difficult to cope and evaluate in a simple and fast way using simple univariate statistical tools, specially due to their large number and to their multivariate correlation. In order to discover relevant patterns within large multivariate data sets, the application of modern chemometric methods based in statistical multivariate data analysis is proposed. After applying chemometric methods, point and diffuse sources of contaminants in the environment and their origin (natural, anthropogenic, industrial, ...) are identified and their relative distribution among samples (geographical, temporal, among environmental compartments) evaluated. At each sampling site, relative source quantitative apportionment is estimated allowing a global evaluation of the environmental impact, distribution and evolution of main chemical contamination sources in the environment. In this presentation, different chemometric methods will be tested on a series of environmental data sets. In particular, the application of principal component analysis and multivariate classificatory methods is shown to be a powerful tool for the goal of chemometrics modelling of contamination sources in large environmental data sets acquired in monitoring studies.

INTRODUCCIÓN

Esta tesis nace de la necesidad de utilizar nuevas técnicas de análisis de datos, sobre todo cuando se trata de datos obtenidos instrumentalmente o cuando se cuenta con un gran número de datos producto del monitoreo de diferentes puntos y épocas de muestreo. Como miembro de la familia universitaria de la UNI, me sentí en la obligación de aportar a ésta, a través de la investigación de nuevas técnicas. La Quimiometría es esta nueva técnica que permitirá visualizar rápida y efectivamente un gran conjunto de datos, de modo que se obtenga información valiosa de los datos obtenidos.

Esta técnica se empieza a utilizar con análisis de aceites comestibles. Lamentablemente para experimentar con la quimiometría son necesarios cientos de datos, los cuales resultan muy costosos de obtener. Felizmente, el Ing. Víctor Cataño Cauti, Director General de la Empresa UNIPETRO ABC S.A.C, tuvo la gentileza de proporcionar todos los datos correspondientes a los análisis de aguas, que forman parte de su Programa de Adecuación y Manejo Ambiental.

Esta es la primera experiencia en la UNI que hace uso de la Quimiometría, a pesar que esta técnica ya es ampliamente utilizada en otros países. Inclusive en muchos de ellos ya forma parte de los currículos de antegrado. Es por ello que esta tesis no trata de abarcar todos los aspectos de la quimiometría, si no dar las bases iniciales para el uso más continuo de esta disciplina.

CONTENIDO

OBJETIVOS DE LA TESIS

Objetivos Generales de la Tesis:

- Aplicar algunas herramientas quimiométricas como el Análisis de Componentes Principales y el Análisis Discriminante para un estudio más profundo de los análisis de aguas de la Empresa UNIPETRO ABC S.A.C.
- Proporcionar los procedimientos necesarios para usar estas herramientas quimiométricas en el tratamiento de datos químicos y en la obtención de información relevante y clasificada.

Objetivos Específicos de la Tesis:

- Utilizar el Análisis de Componentes Principales para obtener información relevante de los análisis de aguas en la Empresa UNIPETRO ABC S.A.C., así como encontrar o descartar una posible contaminación en las aguas del río Pariñas, el cual es aledaño a esta empresa.
- Realizar un Análisis de Clasificación (Análisis Discriminante) para diferenciar a partir de todos sus datos, las aguas de producción con las aguas del río Pariñas.

CAPÍTULO I
ESTUDIO BIBLIOGRÁFICO

1. CAMPO DE LA QUIMIOMETRÍA

1.1. DEFINICIÓN

La quimiometría fue definida como "la disciplina que utiliza métodos matemáticos y estadísticos para diseñar o seleccionar procedimientos o experimentos óptimos, o extraer, la máxima información relevante de los análisis químicos", por The Chemometrics Society en 1975. Sin embargo, D. L. Massart en 1997, amplia esta definición afirmando que la quimiometría es "la disciplina química que utiliza la matemática, la estadística y la lógica formal para diseñar o seleccionar procedimientos experimentales óptimos, proporcionar la máxima información química relevante a partir del análisis de datos químicos y obtener conocimiento a partir de los sistemas químicos"*[1,9]*.

Con el rápido desarrollo de las técnicas instrumentales, en el campo de la química se ha pasado en pocos años de utilizar técnicas en las cuales la concentración de un solo analito se encuentra asociada a una sola señal (por ejemplo, colorimetrías), a utilizar cada vez más, técnicas en las cuales se analizan simultáneamente las concentraciones de diversos analitos asociados a múltiples

señales (por ejemplo, espectroscopia infrarroja). En estas condiciones, y con el gran caudal de información que se puede obtener de por ejemplo un espectro, es necesario disponer de técnicas que permitan extraer la información útil de la que no es.

Así, la automatización y computarización de los laboratorios ha traído consigo diversas consecuencias. Una de las ellas es la rápida adquisición de gran cantidad de datos. Ahora bien, sabemos que la posesión de dichos datos dista, muchas veces, de proporcionar respuestas adecuadas. Obtener los datos no es sinónimo de poseer información; debemos interpretarlos y colocarlos en el contexto adecuado para convertirlos en información útil para el usuario. La quimiometría es la disciplina que tiene esta finalidad *[1]*.

La palabra Quimiometría, introducida hace aproximadamente treinta años, quiere resumir el concepto que engloba la medida en química. Se podría argumentar que, ciertamente, la medida en química siempre ha sido el campo de actuación de la química analítica. La Quimiometría trata, específicamente, de todos aquellos procesos que transforman señales analíticas y datos más o menos complejos en información útil.

La Quimiometría utiliza métodos de origen matemático, estadístico y otros procedentes del campo de la lógica formal para conseguir sus fines. Por todo ello, la quimiometría se sitúa en un campo interdisciplinario. Aunque sus métodos y herramientas provienen de otras disciplinas (como, de hecho, ocurre habitualmente en la química analítica), claramente los fines de la quimiometría están ligados a la química y su éxito depende de los problemas químicos que sea capaz de resolver *[2]*.

Una vez obtenidos los datos analíticos, la quimiometría permite mejorar el rendimiento del proceso analítico en todas sus etapas y asegura la calidad de los resultados (Figura 1) *[1]*.

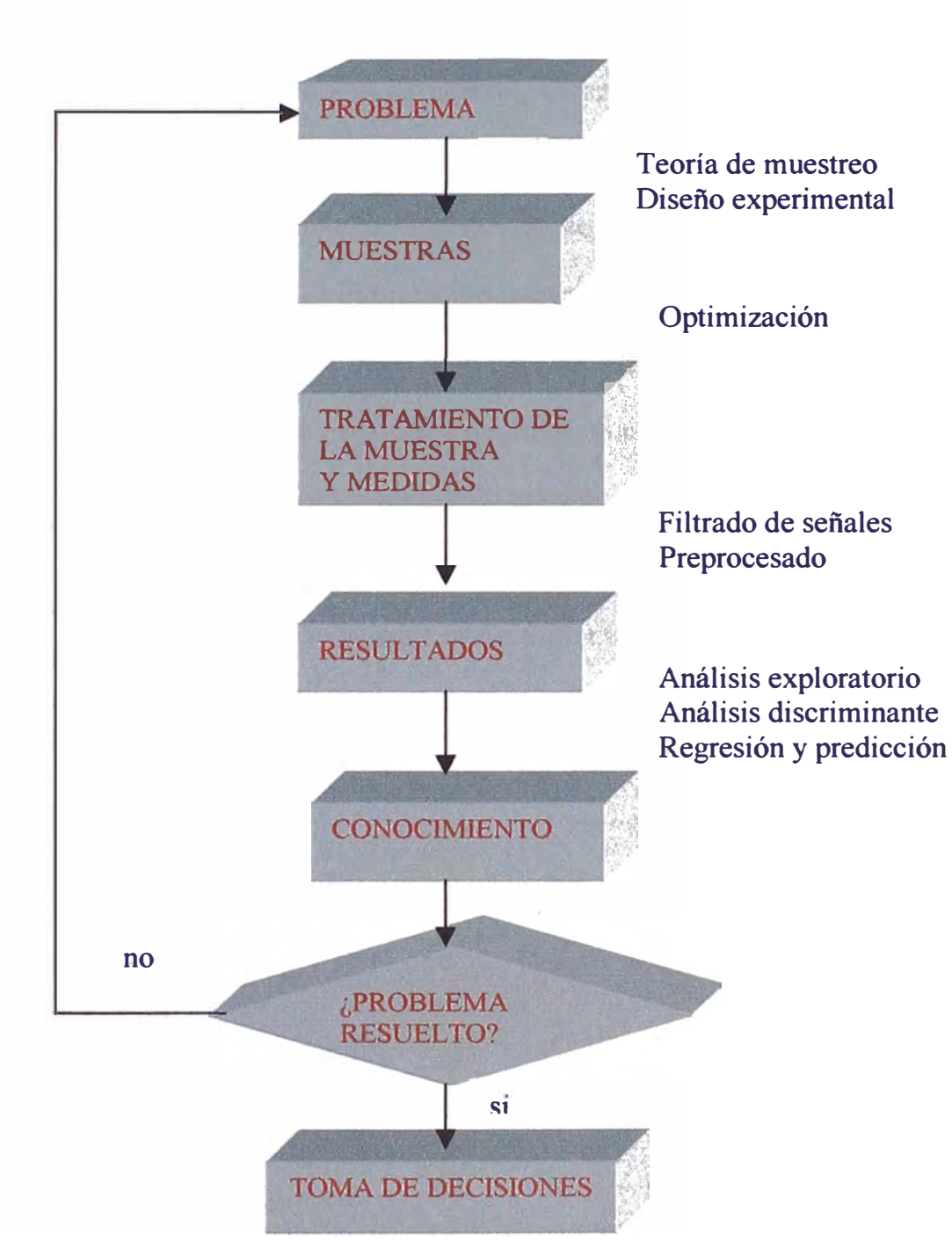


Figura 1. La Quimiometría como herramienta auxiliar o esencial en todas las etapas del análisis

La Quimiometría genera valor añadido a la Química Analítica en dos sentidos:

- Diseña y optimiza los experimentos, de modo que permite obtener más información a partir de los datos.
- Incorpora al químico, y en general, al profesional de laboratorio, a la cadena de control del proceso productivo. Así dicho profesional deja de ser un mero productor de datos, y se convierte en alguien que genera, interpreta y comunica información relevante para la toma de decisiones.

La Quimiometría se puede considerar como una rama aplicada y especializada de la Química Analítica y de la Estadística Aplicada, que en unión con la Informática (Figura 2) permite extraer la información más relevante de los datos químicos. Debe agregarse que el desarrollo de la quimiometría se debe a la evolución de los sistemas informáticos, sin embargo, los resultados que éstos emitan solo serán útiles cuando el analista químico los interprete adecuadamente *[1,9]*.

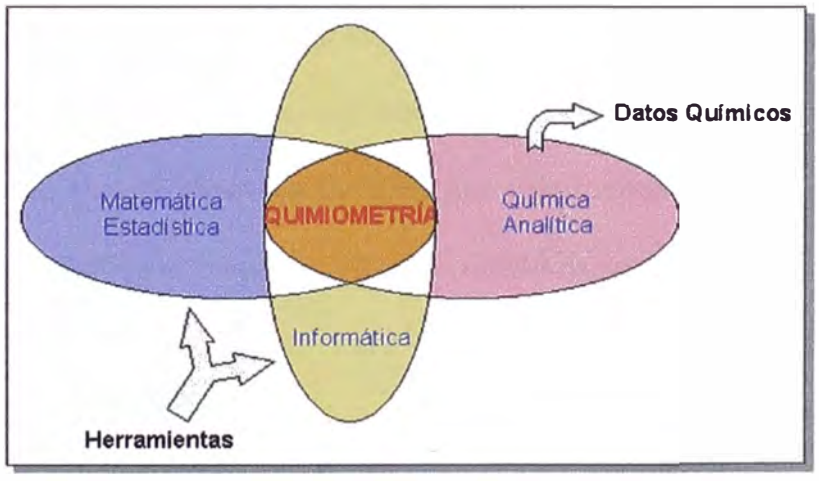


Figura 2. Quimiometría y su interrelación con otras disciplinas [3].

1.2. EL ANÁLISIS DE DATOS

Los científicos con cierta experiencia en el análisis de datos reales saben que, frecuentemente, éstos no contienen la información suficiente. A menudo, los datos registrados no son representativos del fenómeno que quiere estudiarse, no contienen suficiente variabilidad; la parte aleatoria es más relevante que la parte correspondiente a la variación sistemática o existen otras razones por las que las técnicas quimiométricas no puedan extraer información útil sobre el conjunto de datos *[2]*.

Si no existe información es evidente que es imposible extraerla, aún con las técnicas quimiométricas más potentes. Quizás motivado por el hecho de que la información raramente se manifiesta de forma explícita a través de los datos registrados, este principio, que parece tan obvio, se ignora a menudo.

Las consecuencias son desastrosas para la quimiometría, dado que puede atribuirse el fracaso del análisis a las técnicas utilizadas cuando en realidad es debido a la deficiencia de los datos registrados. Debemos asegurarnos, por tanto, que la información se encuentra efectivamente en los datos a analizar. La única forma de hacerlo es programando con detalle las experiencias que conducen a ellos *[8]*.

Normalmente existen diversos factores que intervienen en las experiencias tales como la composición y concentración de reactivos, presencia y ausencia de catalizador, presión, temperatura, etc. Para llevar a cabo el número mínimo de experimentos (que suele ser un aspecto muy parecido en ámbitos industriales, dado el costo que representan) deben variarse simultáneamente los valores que asignamos a

las variables experimentales. Esta estrategia, opuesta a la variación de un solo factor en cada experiencia mientras el resto permanece constante, demanda un cierto conocimiento de las técnicas adecuadas, a cambio, asegura la presencia de información relevante en los datos y una forma mucho más rápida y fiable de obtenerla.

El diseño de experiencias o experimentos es la parte de la quimiometría que estudia dónde, cómo y cuándo deben realizarse las experiencias para que contengan la información necesaria. El empleo de los distintos tipos de diseños experimentales (factoriales completos o fraccionales, de Hadamard, etc.), de técnicas de optimización de resultados o de superficies de respuesta, constituyen algunos de los aspectos estudiados por esta extensa área de la quimiometría *[8]*.

El éxito de la quimiometría depende en buena parte, tal como ocurre en otras disciplinas que se centran en el estudio de la medida (econometría, biometría, infometría, ...), de la resolución de nuevos métodos, y que aquellos que están consolidados se apliquen con éxito a la resolución de problemas reales. En el caso de la quimiometría, los problemas prácticos a solucionar deben estar relacionados con la química, de esta forma, la quimiometría no corre el riesgo de convertirse en una rama de la ciencia alejada de la realidad. Más aún, los practicantes de la quimiometría deben estar atentos al desarrollo constante de la química y generar métodos que se adapten a las nuevas problemáticas que van surgiendo *[2]*.

En consecuencia, por una parte, se necesitan investigadores bien formados y atentos a la realidad cambiante de la química. Por otra parte, se requieren expertos,

conocedores de las distintas técnicas, que sepan aplicar con éxito las herramientas disponibles. Sin embargo, lo más importante es que exista un buen número de usuarios. Usuarios formados e informados, que conozcan y puedan definir con precisión los problemas químicos y a la vez conozcan el potencial de la quimiometría. Ellos deberían ser los interlocutores obligados de los expertos.

Para que todo ello sea una realidad, es necesario que se consolide e incremente los niveles de información, de formación y la accesibilidad al conocimiento y a las herramientas que hacen posible su aplicación *[2]*.

1.3. QUIMIOMETRÍA Y ESTADÍSTICA MULTIVARIANTE

1.3.1. Fases de la Quimiometría

Las técnicas clásicas de análisis químicos generan a lo más una variable (generalmente concentración) a la cual se le puede aplicar las técnicas clásicas de la Estadística Descriptiva Univariante.

Sin embargo, las nuevas técnicas instrumentales, y lo que venga en el futuro en Química Analítica, generan tres o más variables, y en algunos casos cientos de ellas (cromatografía, espectroscopia, etc). En este caso debe de aplicarse la Estadística Multivariante, es decir, la Quimiometría *[3,7]*.

La quimiometría, así, se ha convertido en la herramienta que logra el Aseguramiento de la Calidad en el trabajo de los laboratorios modernos (lo que ya se ha convertido en una nueva disciplina química: la Cualimetría). La Quimiometría, y en relación a la Figura 1, comprende las siguientes fases:

i. Selección y Evaluación del Método a emplear

Comprende la aplicación adecuada de la Teoría del Muestreo, un diseño experimental adecuado que permita extraer la información requerida y aplicación de la Teoría de la Información, es decir, como manejar los datos obtenidos.

ii. Optimización

Se logra con un diseño experimental adecuado, logrando optimizar el rendimiento experimental a partir de los recursos a los que estamos sujetos.

iii. Transformación de las señales en Información Química

Es necesario un tratamiento matemático de las señales y/o valores obtenidos para lograr extraer la información química.

iv. Transformación de la Información Química en diagnóstico

Si la información química extraída es útil, entonces estaremos en condiciones de establecer y/o reconocer ciertos modelos o patrones en nuestra estructura de datos, la cual nos permitirá hacer un diagnostico del sistema químico estudiado *[3,7]*.

De esta manera, la Quimiometría, a partir de un experimento en el cual se extraen cientos de datos, los transforma en información. Esta información, con el uso de sistemas expertos (inteligencia artificial) se transforma en conocimiento, a partir del cual logramos con creatividad establecer una hipótesis. Esta hipótesis permitirá, si se resuelve el problema, hacer un diagnóstico. Y si la hipótesis no es adecuada, nos servirá como base del diseño de nuevos experimentos que permitan finalmente dar solución al problema (Figura 3) *[3,7]*.

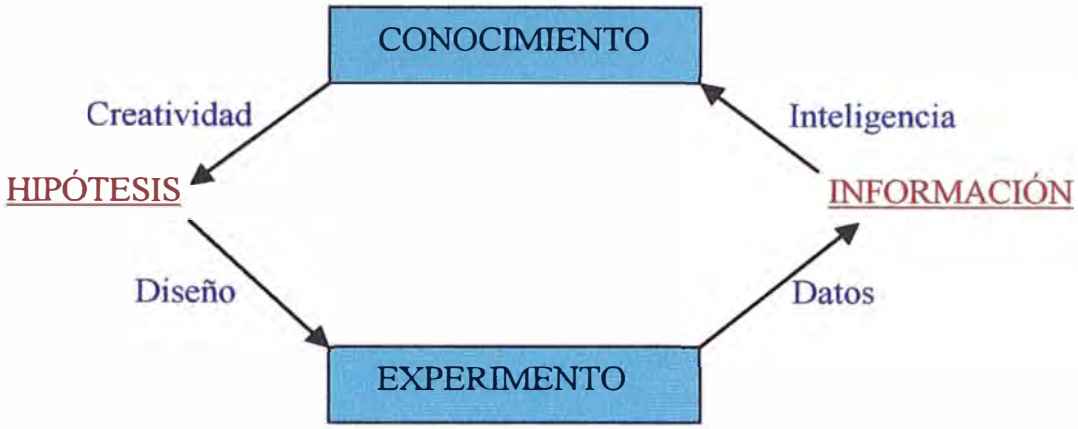


Figura 3. El ciclo de la Quimiometría

1.3.2. Fases del Análisis Multivariante

El Análisis Multivariante es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariantes en el sentido de que hay varias variables medidas para cada individuo u objeto estudiado. Su razón de ser radica en un mejor entendimiento del fenómeno objeto de estudio, obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir. Las técnicas que comprende se pueden clasificar en tres grandes grupos *[4,7]* (ver Figura 4):

i. Métodos de dependencia

Suponen que las variables analizadas están divididas en dos grupos: *las variables dependientes y las variables independientes.* El objetivo de los métodos de dependencia consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma.

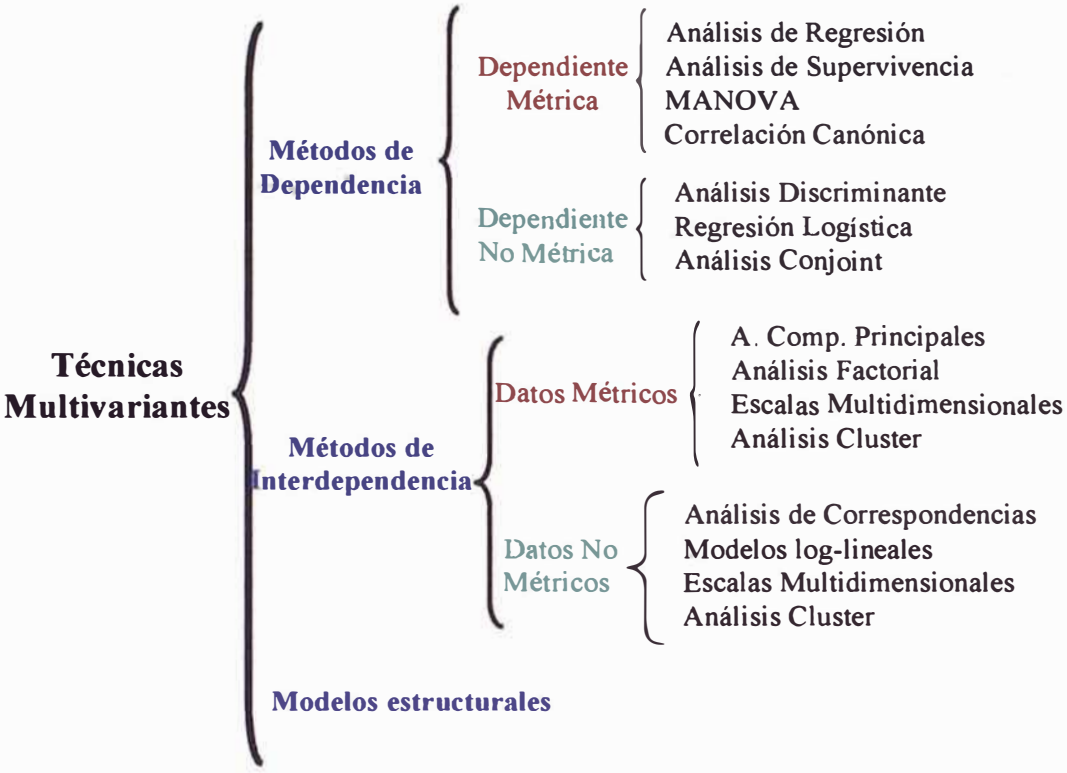


Figura 4. Análisis Multivariante y sus diversas técnicas

ii. Métodos de interdependencia

Estos métodos no distinguen entre variables dependientes e independientes y su objetivo consiste en identificar qué variables están relacionadas, cómo lo están y por qué *[7]*.

iii. Métodos estructurales

Suponen que las variables están divididas en dos grupos: el de las variables dependientes y el de las independientes. El objetivo de estos métodos es analizar, no sólo cómo las variables independientes afectan a las variables dependientes, si no también cómo están relacionadas las variables de los dos grupos entre sí *[7]*.

Así por ejemplo, un cierto número de objetos, a los que se les ha determinado las variables x_1 y x_2, pueden representarse en el plano, y podremos observar algunos agrupamientos de los objetos; y dado un nuevo objeto será posible asignarle una clase o grupo (Figura 5).

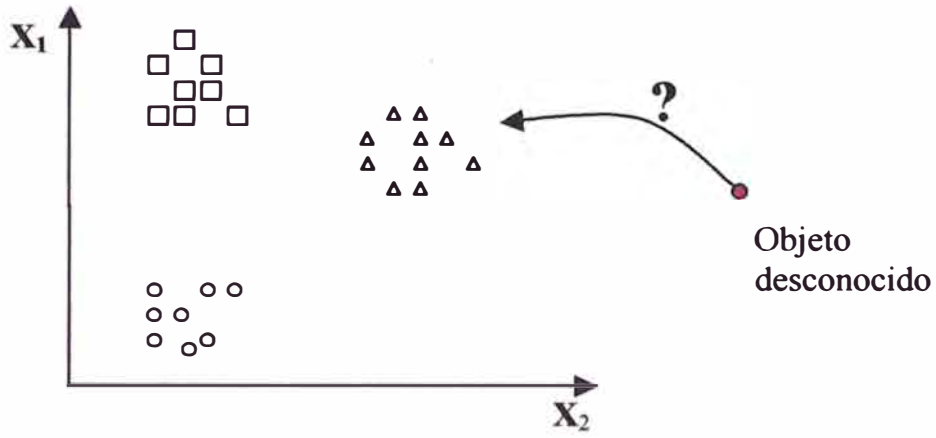


Figura 5. Agrupación de objetos en Estadística Multivariante

Esta tesis no pretende hacer uso de todas las herramientas quimiométricas, si no, más bien, como primera aproximación a éstas, hará sólo uso del análisis exploratorio de datos a través del Análisis de Componentes Principales *[1,15]*.

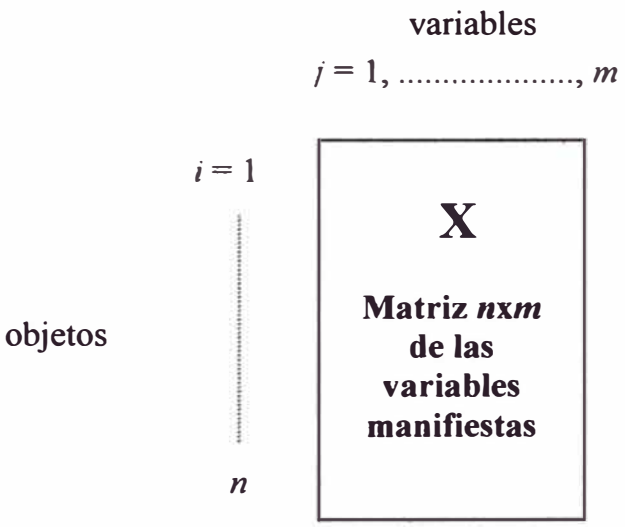
2. EXPLORACIÓN DE DATOS

El Análisis Multivariado surge cuando a un mismo individuo se le pide más de una característica de interés. Los métodos del análisis multivariado son un conjunto de técnicas que permiten al investigador interpretar y visualizar conjuntos grandes de datos; además nos permitirán encontrar relaciones entre las variables,

entre los objetos y entre ambos, pero en general tendrán un carácter exploratorio y no tanto inferencial *[5]*.

2.1. LA MATRIZ DE DATOS

Supongamos que determinamos m variables manifiestas (parámetros) correspondientes a n objetos (muestras), los resultados pueden presentarse mediante una matriz de datos, en la que cada fila representa un objeto y cada columna una variable medida.



Uno de los problemas centrales en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de las m variables, correspondientes a n objetos, mediante un pequeño subconjunto de variables $r < m$, se habrá producido la disminución del problema a costa de una pequeña pérdida de información *[5]*.

2.2. EXAMEN PRELIMINAR DE LA MATRIZ DE DATOS

Una vez obtenida la matriz de los datos la quimiometría recomienda la exploración de datos, que permitirán poner de manifiesto y resaltar la información contenida en dicha matriz de datos multidimensional, **X**, la que ya hemos dicho, está constituida por *n* objetos y *m* variables. Las variables son *m* vectores, $\mathbf{x_1}$, $\mathbf{x_2}$,....., $\mathbf{x_j}$, ...,$\mathbf{x_m}$, que definen un espacio *m*-dimensional, y los objetos constituyen las filas de la matriz, por tanto, se pueden considerar como un conjunto de *n* vectores transpuestos: $\mathbf{o}^T_1$, $\mathbf{o}^T_2$,..., $\mathbf{o}^T_i$,..., $\mathbf{o}^T_n$. En el espacio constituido por las variables, los objetos forman estructuras de dos tipos, grupos y correlaciones. Si se transpone la matriz de datos ($\mathbf{X} \rightarrow \mathbf{X}^T$), son las *m* variables las que forman estructuras en el espacio definido por los *n* objetos *[1]*.

Se consigue una visión intuitiva y completa de las estructuras presentes en la matriz de datos mediante el uso de representaciones gráficas: el histograma de un vector (una variable o un objeto), el gráfico de dispersión para dos vectores (una variable frente a otra; o un objeto frente a otro), los gráficos de dispersión sobre los componentes principales, para tres o más vectores y otras herramientas gráficas *[1]*.

2.2.1. Reconocimiento de estructuras entre objetos.

Existen muchas formas para poner de manifiesto las estructuras ocultas en la matriz de datos. Sin embargo, la mejor visión de las estructuras presentes se logra proyectando los objetos sobre planos oblicuos definidos por vectores que son combinaciones lineales de las variables de la matriz de datos. Existen varios criterios

para construir tales vectores, sin embargo, cuando se explora una matriz de datos por primera vez, es recomendable utilizar la rotación propia, que da lugar a un conjunto de vectores denominados "componentes principales" *[7,15]*.

La observación directa de los datos y los gráficos de dispersión sobre pares de variables son herramientas útiles, pero insuficientes, para poner de manifiesto y describir las estructuras presentes en el caso general, en el que tres o más variables contienen información en parte coincidente y en parte distinta, y cada una de ellas puede dividir los objetos en dos o más categorías *[1]*.

Cuando se tienen tres variables el uso de las computadoras personales resuelve el problema de las proyecciones sobre, sin embargo, este procedimiento no soluciona el caso general para más de tres variables. Precisamente por ello las técnicas de exploración de datos se hacen necesarias puesto que nos permiten la reducción de dimensiones, de modo que la información relevante contenida en la matriz multidimensional pueda quedar reflejada, del mejor modo posible, sobre dos o tres dimensiones oblicuas obtenidas como combinaciones lineales de las variables originales *[1]*.

La exploración de datos, aplica herramientas que analizan las estructuras formadas por los objetos en el espacio de las variables, pero también otras que ponen de manifiesto las relaciones de las variables entre sí. Si se trabaja sobre la matriz original, se obtienen representaciones de objetos (Figura 6), pero si se transpone la matriz, las mismas herramientas de cálculo proporcionan representaciones de las variables (Figura 7). La transposición de la matriz evita duplicar las herramientas

gráficas y de cálculo programadas en el paquete estadístico *[1]*.

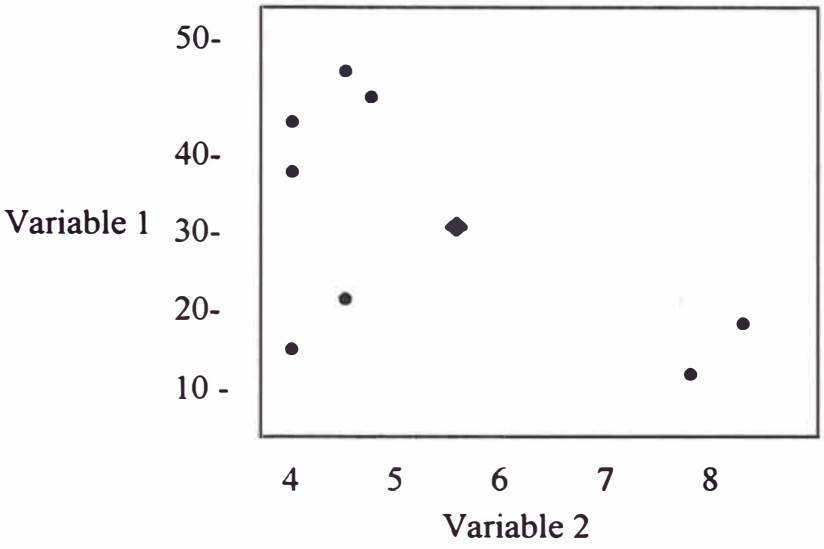


Figura 6. Proyección de los objetos (puntos) y su centroide (rombo) sobre un plano formado por las variables

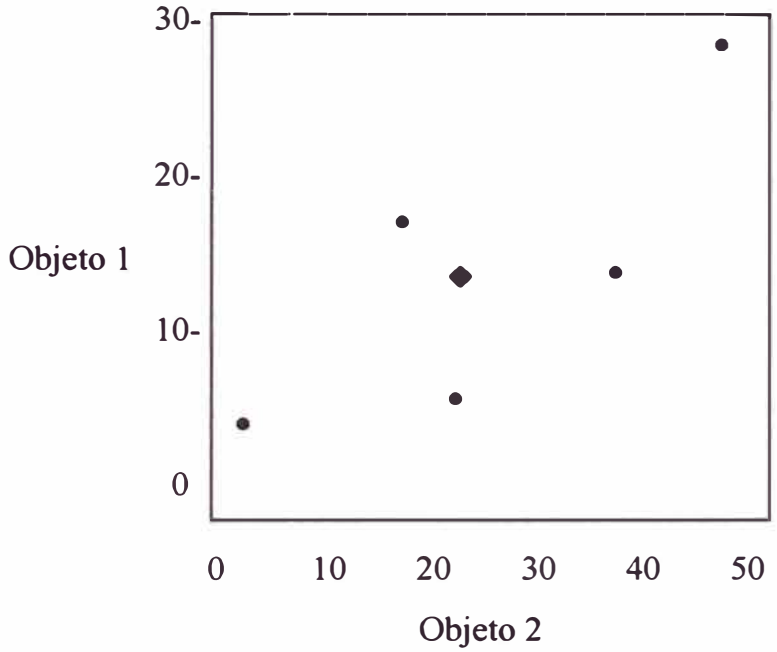


Figura 7. Proyección de variables y su centroide sobre un plano formado por los objetos

2.2.2. Variables Manifiestas, Latentes y Fundamentales

Uno de los objetivos de la exploración de datos es poner de manifiesto las tendencias ocultas presentes en los datos, y estudiar las clases o las fuentes de varianza a la que obedecen. El estudio de las fuentes de varianza o "análisis factorial" comienza por la selección de las "variables manifiestas" que van a constituir la matriz de datos, es decir aquellas variables que pueden medirse experimentalmente, y que se sabe o se supone que varían con los factores que se están buscando, esto es, factores subyacentes (ambientales, clínicos, industriales, geográficos, históricos, etc.), que expliquen la configuración de los grupos y las correlaciones observadas. Se intenta identificar un pequeño número de factores que explique la mayor parte de la varianza aportada por un número mayor de variables manifiestas. Y luego, se diseña el conjunto de objetos sobre los que se va a realizar la medida de las variables manifiestas. Se obtiene así la matriz de datos *[1]*.

Una vez construida la matriz se aplican una serie de herramientas estadísticas que revelan las estructuras presentes. Una de las operaciones de procesado es la rotación propia y el análisis de los vectores obtenidos o "componentes principales". Los vectores que se obtienen siguiendo algún criterio racional, con la finalidad de describir tendencias ocultas en los datos se denominan "variables latentes". En el procesado se obtienen también proyecciones de los puntos sobre planos constituidos por pares de variables latentes.

El análisis exploratorio de datos es útil para todo tipo de actividades científicas y tecnológicas (Tabla 1). En concreto, el análisis factorial intenta

responder a preguntas tales como el origen geográfico de un alimento o de un vertido contaminante, la época de fabricación de un objeto prehistórico, las causas de un problema de contaminación, o las características que hacen que un producto sea preferido por los consumidores frente a otros productos similares *[1]*.

Tabla 1. Aplicaciones del Análisis Factorial

Área científica o tecnológica	Tipo de objetos	Variables manifiestas	Factores buscados
Geología	Aguas subterráneas	Trazas metálicas	Formaciones rocosas, desplazamiento de fallas
Fisiología	Leche	Ácidos grasos	Vías metabólicas
Ciencias ambientales	Suelos, aguas	Contaminantes, nutrientes	Agentes naturales, actividades humanas
Arqueología	Restos cerámicos	Trazas metálicas	Lugar y fecha de fabricación
Arte	Pigmentos	Compuestos orgánicos e inorgánicos	Época y autor
Policía científica	Billetes falsos	Trazas metálicas	Fabricante del papel
Alimentación	Quesos, harinas	Proteínas, polisacáridos	Especies animales o vegetales
Alimentación	Vinos, aceites	Trazas metálicas, aromas	Origen geográfico

Resumiendo, el tratamiento de datos parte de las variables manifiestas y las combina con el fin de revelar tendencias ocultas. Las combinaciones lineales de las variables manifiestas obtenidas en función de criterios racionales que modelan la "nube de datos" se denominan "variables latentes". Las variables latentes se utilizan para descubrir en el espacio aquellas direcciones que indican correlación entre las causas objetivas de varianza. Dichas direcciones son los “factores subyacentes” o “variables fundamentales” (ver Figura 8) *[1]*.

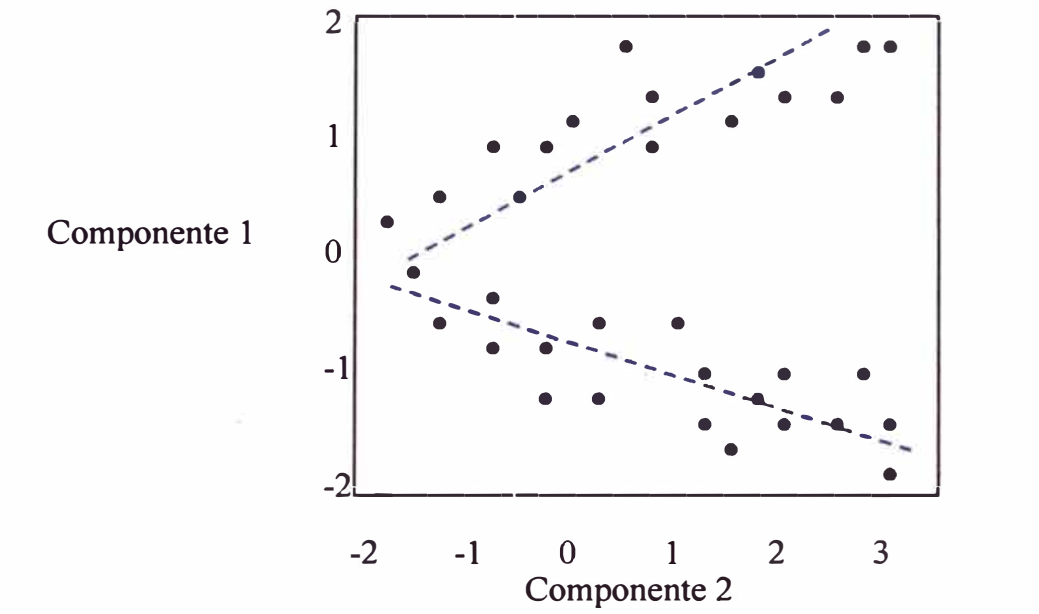


Figura 8 .Proyección de 53 objetos (nube de datos) sobre el plano de 2 componentes principales (variables latentes); a su vez 2 variables fundamentales (líneas de trazas)

2.2.3. Preprocesado de la matriz de datos

Generalmente antes del procesado de los datos originales, éstos deben de transformarse de modo que permitan aplicar otras herramientas estadísticas, ya sea para hacer posible su aplicación, mejorar los resultados, o evitar que puedan alcanzarse conclusiones incorrectas o incluso absurdas *[1,15]*. Las técnicas de preprocesado recomendados son:

- El relleno de huecos.
- El cambio de escala o escalado.
- El centrado, o traslación del origen de coordenadas al centroide.
- La eliminación de valores anómalos.

Si se tienen muchos objetos (por ejemplo, más de 50), o muchas variables

(por ejemplo, monitoreo de muchas variables en un periodo de tiempo muy grande), la observación directa de los datos es de escasa utilidad. En estos casos, es útil comenzar el preprocesado mediante la obtención de algunos descriptores muestrales a lo largo de las filas o de las columnas, tales como los valores máximo y mínimo, la media y la desviación estándar, y el número de datos perdidos o "huecos" .

2.2.4. Relleno de huecos

En una matriz de datos se producen huecos por diversos motivos:

- Algunos datos no se han podido obtener a causa de un error accidental durante la experimentación, y ya no es posible repetir las medidas en las mismas condiciones.
- Se han eliminado unos pocos datos por ser anómalos.
- En algunas muestras el analito se halla por debajo del limite de detección (LD) y, por tanto, su valor no se puede conocer sin cambiar de método.
- Los datos proceden de fuentes bibliográficas (por ejemplo, constantes termodinámicas) y no están completos.

Puesto que no es posible operar con vectores y matrices incompletos, es necesario eliminar o rellenar los huecos de algún modo. Una mala solución sería eliminar las filas o las columnas que contienen huecos, ya que de este modo se desecha la información útil aportada por los datos disponibles de las filas o columnas eliminadas. La forma correcta de trabajar es aplicar los criterios conocidos como "relleno medio" y "relleno al azar", que atribuyen valores supuestos a los datos no

disponibles. La asignación se hace procurando distorsionar mínimamente las medias y varianzas de filas y columnas *[1]*.

Cuando el hueco es debido a que la concentración de analito está por debajo del LD del método, y no se tiene información alguna sobre su posible valor, el relleno medio puede hacerse con la media entre el LD y cero que implica asignar al hueco el valor LD/2.

En el relleno al azar, el hueco se sustituye por un valor al azar pero acotado dentro de los límites de la variable.

Debe tenerse en cuenta que el relleno siempre introduce distorsiones en los datos, cualquiera que sea el criterio aplicado. El relleno medio tiende a reducir la dispersión de las variables, mientras que el relleno al azar tiende a aumentar la dispersión, por lo que en ambos casos se falsea la realidad. En principio, aumentar o reducir la dispersión a lo largo de una columna o de una fila es igualmente perjudicial *[1,15]*. Sin embargo, cuando existen huecos, no se puede actuar de otra forma que no sea rellenarlos de algún modo, y proceder luego a estudiar el efecto del relleno. Si las conclusiones del estudio estadístico varían ampliamente, con el valor de los rellenos, se deduce que la matriz no tiene información suficiente relacionada con el objetivo que se pretende. En este caso, se hace necesario mejorar la matriz, ya sea midiendo más objetos, o buscando nuevas y mejores variables. En cambio, si las conclusiones son estables, no habrá importado cómo se hayan rellenado los huecos.

2.2.5. Escalado

La mejor representación gráfica de un conjunto de datos se tiene cuando éstos llenan todo el espacio disponible para cada variable. Esto se logra mediante el escalado. Esta técnica es especialmente importante en la exploración de datos multivariantes. Si no se realiza de un modo adecuado para todas las variables, las que contienen datos de mayor valor numérico impiden observar la información aportada por variables con datos de bajo valor numérico. Igualmente, variables muy dispersas impiden apreciar diferencias sutiles entre variables menos dispersas.

Si no se escalan adecuadamente las variables, alguna información decisiva puede pasar desapercibida. El escalado es absolutamente necesario cuando una misma matriz contiene datos de trazas y de componentes mayoritarios, o señales débiles frente a otras mayores.

Además del criterio intuitivo de "llenar todo el espacio", son habituales otros dos criterios: el escalado por el intervalo y el autoescalado o transformación z multidimensional.

El escalado por el intervalo consiste en colocar el menor valor de cada variable en el origen, y dividir todos los valores por el intervalo de la variable. Todas las variables quedan acotadas entre cero y la unidad. Para la variable x_j la transformación es:

$$x_{ij}' = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}$$

Una variable preprocesada, x', se denomina "característica" para distinguirla de la variable original, x, que contiene los datos sin modificar. Así mismo, los valores a lo largo de una característica se denominan "puntuaciones", para distinguirlos así de los valores originales o "coordenadas". Una *puntuación* es, por tanto, una coordenada sobre una variable que ha experimentado algún tipo de transformación *[1]*.

2.2.6. Autoescalado

El autoescalado, tipificación o transformación z multidimensional, consiste en un centrado por columnas junto con un división de cada columna por su desviación estándar. Para centrar una columna se resta la media de la columna a todos sus valores: $x_{ij}' = x_{ij} - \bar{x}_j$. Las "puntuaciones z" vienen dadas por:

$$z_{ij} = x_{ij}' = \frac{x_{ij} - \bar{x}_j}{s_j}$$

En notación vectorial se tiene

$$z_j = \frac{x_j - \bar{x}_j}{s_j} = \frac{x_j^*}{s_j}$$

donde $\bar{x}_j$ es un vector cuyos elementos son todos iguales a la media de la columna j, $\bar{x}_j$. Al restar $\bar{x}_j$ a todos los elementos de $\mathbf{x_j}$ se obtiene el vector centrado: $\mathbf{x_j^*} = (\mathbf{x_j} - \bar{x}_j)$. Dividiendo los elementos de $\mathbf{x_j^*}$ por la desviación estándar, s_j, se obtiene el vector autoescalado, $\mathbf{z_j}$, lo que se aplica a todas las variables *[15]*.

En un autoescalado, y como consecuencia del centrado, el origen de

coordenadas queda situado en el *centroide* de la nube de datos. Por otra parte, al dividir por s_j las características (nuevos ejes de coordenadas) quedan expresadas en unidades de su propia desviación estándar: comprimidas las de s_j grande, y expandidas las de s_j pequeña. De este modo, la nube de datos aparece distribuida simétricamente en torno a su centro de gravedad, y además, la expansión de la nube es la misma en cualquier dirección del espacio, independientemente de que los valores numéricos de las variables manifiestas sean grandes o pequeños *[1,15]*.

El autoescalado es una operación frecuente e importante en análisis multivariante, puesto que, al igual que otras operaciones de escalado, ofrece las mismas oportunidades a todas las variables para influir en las conclusiones, tanto si poseen valores pequeños como si éstos fueran grandes *[1]*.

2.2.7. Correlaciones como medidas de similitud

En la matriz de datos no transpuesta, es decir, el espacio definido por las variables, los coeficientes de correlación lineal de Pearson, o correlaciones bivariadas (tomadas de dos en dos), r_{kl}, son útiles para agrupar variables porque indican asociación, semejanza o similitud entre ellas. Así, valores de r^2_{kl} próximos a la unidad señalan "perfiles similares"(en el mismo sentido), o bien "perfiles opuestos"(sentidos opuestos), entre las variables $\mathbf{x}_k$ y $\mathbf{x}_l$, indicando que ambas no son independientes, sino que se asocian de algún modo.

$$Coeficiente\ de\ correlación\ de\ Pearson = r_{xy} = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

En ese mismo espacio, los objetos se agrupan según las distancias que los separan. Objetos próximos entre sí, y a la vez alejados de otros objetos, tienen coordenadas parecidas y, por tanto, son similares y forman grupo.

Al transponer la matriz de datos, son entonces los objetos los que se pueden agrupar en función de sus correlaciones mutuas, mientras que las variables se pueden agrupar según las distancias que las separan en el espacio definido por los objetos.

Las correlaciones se presentan en forma de matriz. La matriz de las correlaciones, **R,** es una matriz cuadrada, de dimensiones *m*x*m*, formada por las correlaciones lineales bivariadas de las variables.

Por definición, dos vectores son linealmente dependientes o colineales tan sólo si $r^2_{kl}=1$, o lo que es igual, si una de las dos expresiones, $\mathbf{x_1} = c_o + c_1 \mathbf{x_k}$ ó $\mathbf{x_1} = c_o - c_1 \mathbf{x_k}$, se cumple exactamente. Sin embargo, en estadística multivariante es frecuente utilizar conceptos tales como "casi" colineales o "fuertemente" correlacionados, indicando que los vectores tienen direcciones próximas entre sí, o próximas a ser opuestas, si bien no exactamente coincidentes u opuestas del todo.

Finalmente, valores alejados de +1 y –1 (no necesariamente próximos a cero), indican variables no correlacionadas: la proyección de la nube de puntos sobre el plano formado por $\mathbf{x_l}$ y $\mathbf{x_k}$ tiene un aspecto redondeado.

En general, los grupos de variables fuertemente correlacionadas entre sí obedecen a una misma causa o fuente común de varianza, esto es, a una variable fundamental *[1,15]*.

i. Interpretación geométrica de la correlación

Un vector **x** se puede interpretar como un segmento que une el origen del espacio con un punto cuyas coordenadas son los valores o elementos del vector (Figura 9).

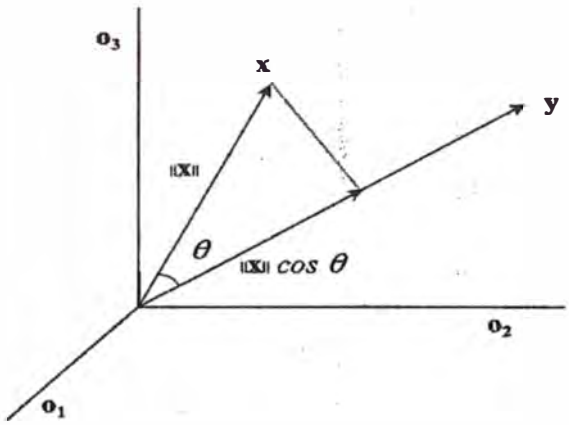


Figura 9. Proyección del vector x sobre el vector y

La distancia de Euclides (euclídea o en línea recta) desde el origen del espacio a ese punto es el módulo, norma o longitud del vector, $\|\mathbf{x}\|$. El producto interno o escalar de dos vectores **x** e **y** se define como el producto de sus módulos por el coseno del ángulo que forman:

$$\mathbf{x}^T\mathbf{y} = \sum x_i y_i = \|\mathbf{x}\| \, \|\mathbf{y}\| \cos\theta$$

que se puede interpretar como el producto de $\|\mathbf{y}\|$ por la proyección de **x** sobre **y**, cuyo valor es $\|\mathbf{x}\|\cos\theta$.

El producto escalar de un vector consigo mismo será:

$$\mathbf{x}^T\mathbf{x} = \sum x_i^2 = \|\mathbf{x}\|^2$$

De acuerdo con la definición de módulo, este producto es igual al cuadrado de la distancia euclídea entre el origen de coordenadas y el extremo del vector.

Un vector normalizado es el cociente entre el vector y su módulo: $\mathbf{u} = \mathbf{x} / \|\mathbf{x}\|$. La longitud de un vector normalizado es la unidad, esto es, $\|\mathbf{u}\| = 1$, o lo que es igual, la suma de los cuadrados de todos sus elementos es igual a 1. Para que se cumpla esta condición, todos los elementos del vector normalizado deben ser inferiores a $|\mathbf{1}|$.

Los elementos de un vector normalizado son los cosenos de los ángulos que forma el vector con los ejes de coordenadas o "cosenos directores" del vector.

El producto escalar de dos vectores normalizados es igual al coseno del ángulo entre ambos:

$$\mathbf{u}^T\mathbf{v} = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \cos\theta$$

Entonces se deduce que r_{xy} es el coseno del ángulo que forman los dos vectores **x** e **y** centrados y normalizados *[1]*, esto es:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} =$$

$$= \frac{(\mathbf{x} - \bar{\mathbf{x}})^T(\mathbf{y} - \bar{\mathbf{y}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|\,\|\mathbf{y} - \bar{\mathbf{y}}\|} = \frac{(\mathbf{x}^*)^T\mathbf{y}^*}{\|\mathbf{x}^*\|\,\|\mathbf{y}^*\|} = \cos\theta$$

ii. Correlación entre vectores y rango de la matriz

Al encontrar correlación entre los *m* vectores columna que definen el espacio, o de la correlación entre los *n* vectores fila representados en el mismo, se puede concluir que ambas, columnas y filas, se pueden representar sin error, o en

todo caso, con un error pequeño, sobre un número reducido de vectores.

El rango de una matriz es el número mínimo de vectores independientes, p, que pueden construirse utilizando las columnas o las filas de la misma. El rango de una matriz también es igual al orden del mayor determinante no nulo que puede encontrarse en la misma. A este respecto, debe tenerse en cuenta que si una matriz cuadrada contiene una columna o una fila linealmente dependiente de otra, su determinante es cero, y por tanto, el rango de la matriz es menor que el orden del determinante. El rango es también el número de vectores necesarios para representar sin error todos los vectores columna, o todos los vectores fila de la matriz. Sin error significa que la representación preserva las longitudes de los vectores y los ángulos entre ellos.

Por tanto, si una matriz contiene vectores linealmente dependientes (esto es, perfectamente correlacionados), es posible representar los m vectores columna sin error en un subespacio de dimensión p tal que $p < m$, y también, los n vectores fila sin error, siendo igualmente $p < n$. Más aún, vectores "casi" colineales con los que forman el subespacio de dimensión p se pueden representar en el mismo con un error pequeño. Es así porque los vectores "casi" colineales con otros del subespacio apuntan en direcciones próximas al mismo *[1]*.

iii. Interpretación de las medidas de similitud

Las correlaciones estiman similitudes asociadas al "perfil" de los vectores que se comparan, esto es, a su posición angular. Sin embargo, la interpretación de una u

otra medida de similitud depende del significado de las posiciones angulares y de las longitudes de los vectores. Si se normalizan los vectores que se comparan (por ejemplo, mediante un autoescalado), las correlaciones estiman similitudes asociadas exclusivamente a la posición angular de los vectores *[1]*.

3. ANÁLISIS DE COMPONENTES PRINCIPALES

3.1. INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Las estructuras o patrones presentes en una nube de datos multidimensional pueden visualizarse al proyectarse éstos puntos sobre un plano. Para ello, es necesario seleccionar cuidadosamente, mediante la aplicación de criterios racionales, el plano de observación. La expresión algebraica de estos criterios da lugar a una serie de poderosas herramientas de análisis de datos, seguida del análisis de los resultados obtenidos o "Análisis de los Componentes Principales"*[1,15]*.

Se construyen modelos de la nube de puntos, buscando secuencialmente las direcciones del espacio que ofrecen la mejor visión posible de la misma. Los nuevos ejes de coordenadas, denominados vectores propios (eigenvectors), autovectores o componentes principales, se construyen como combinaciones lineales de las variables manifiestas. Los componentes principales son variables latentes, esto es, modelan las principales tendencias presentes en la nube de datos.

Además, el estudio de algunas matrices ofrecen información de gran valor sobre las estructuras formadas por objetos y variables, y sobre las relaciones entre ambos. Finalmente, la información disponible es de gran ayuda para identificar y

modelar, mediante rotaciones adicionales, las variables fundamentales o factores subyacentes que obedecen a causas o fuentes de varianza con sentido químico-físico, ambiental, etc.

3.1.1. Análisis supervisado versus el no supervisado

En la Figura 10 se presenta el diagrama de dispersión de 36 muestras de café en grano sobre dos variables: GRASA (contenido en grasa) y CMIN (contenido mineral). Los objetos pertenecen a dos categorías, que corresponden a dos variedades de café: robusta (grupo más compacto, con 28 objetos) y arábiga (grupo más disperso, con 8 objetos).

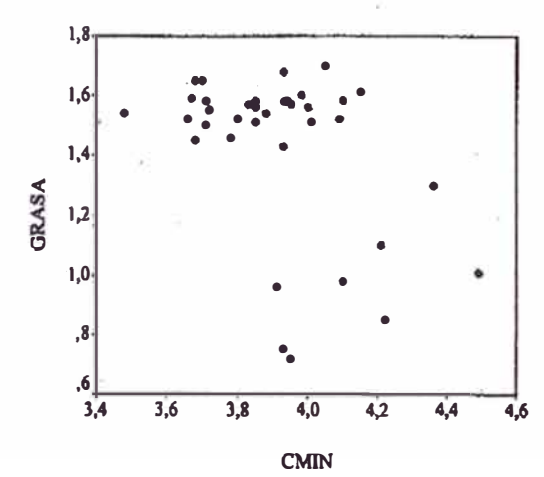


Figura 10. Diagrama de dispersión sobre el plano CMIN-GRASA para el ejemplo del café en grano.

El histograma de la Figura 11 muestra una distribución de los puntos prácticamente aleatoria, cercana a la distribución normal, a lo largo de la variable CMIN. En cambio, como se observa en el otro histograma de la Figura 12, la

distribución a lo largo de la variable GRASA se aleja mucho de la normalidad. El histograma sobre GRASA muestra una incipiente separación de los dos grupos: la variedad arábiga con contenidos bajos en grasa, y la variedad robusta, rica en grasa. Si sólo se observara una variable, GRASA daría una mejor visión de la nube multidimensional que CMIN.

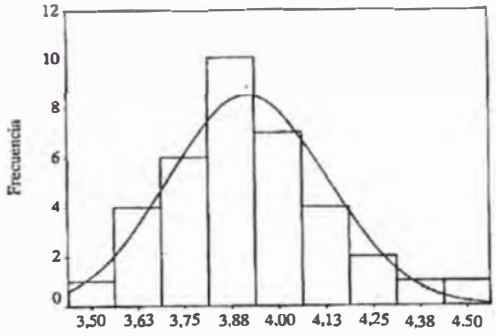


Figura 11. Histograma de la variable CMIN para el ejemplo del café en grano.

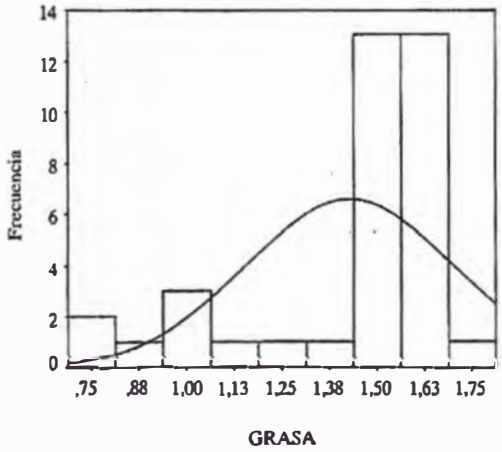


Figura 12. Histograma de la variable GRASA para el ejemplo del café en grano.

Sin embargo, la observación de la Figura 11 sugiere la posibilidad de obtener un histograma más ilustrativo representando las proyecciones de los puntos sobre una dirección oblicua. Los mejores vectores son los que apuntan en direcciones del

espacio caracterizadas por distribuciones estructuradas, esto es, bien diferenciadas respecto a la distribución meramente aleatoria. El vector que apunte en esa dirección oblicua "preferente" se puede encontrar atendiendo a distintos criterios. Para establecer tales criterios es necesario elegir previamente entre las dos opciones siguientes.

- No se conoce, o se ignora deliberadamente, la existencia de categorías.
- Se tiene en cuenta la pertenencia de los objetos a distintas categorías, que pueden ser conocidas o supuestas.

El primer caso constituye un problema de exploración de datos o "análisis no supervisado", mientras que si se asignan los objetos a distintas categorías se tiene un problema de análisis clasificatorio o "supervisado". *El análisis de componentes principales es una técnica de exploración de datos*, y por ello, los componentes se buscan ignorando la posible presencia de categorías.

3.1.2. División de la varianza total en explicada y residual

La varianza total de los datos, $\mathbf{s}_T^2$, es la suma de cuadrados de las distancias de los puntos a su centroide, dividida por el número de puntos, n. En exploración de datos, las varianzas se calculan dividiendo siempre por el número de puntos, n, al contrario de lo que es habitual en Estadística descriptiva, en la que se divide por $(n-1)$. Esta diferencia es debida al distinto enfoque de los objetivos que se persiguen: en estadística descriptiva interesa describir las propiedades estadísticas de la muestra, mientras que en exploración de datos lo importante es revelar las estructuras

presentes. Cuando se construye un vector que modela la nube de puntos, la varianza total queda dividida en dos, la explicada por el vector y la residual que se puede expresar como sigue:

$$s_T^2 = s_{\exp}^2 + s_{res}^2$$

Si el vector es un componente principal, su varianza explicada $s_{\exp}^2$ se denomina "autovalor".

Un vector puede representar una nube de puntos cuando pasa a través de su centroide. En la Figura 13 se indican cuatro puntos y un vector cualquiera que pasa por su centroide, $\mathbf{x}'_1$. En esta gráfica, la varianza explicada es la varianza del vector, o varianza de las proyecciones de los puntos sobre el vector, que se calcula como la suma de cuadrados de las distancias de los puntos al centroide en la dirección del vector (líneas de trazos de la figura), dividida por *n*. Por su parte, la varianza residual es la suma de cuadrados de distancias de los puntos al modelo (líneas de puntos, perpendiculares al vector), dividida también por *n*.

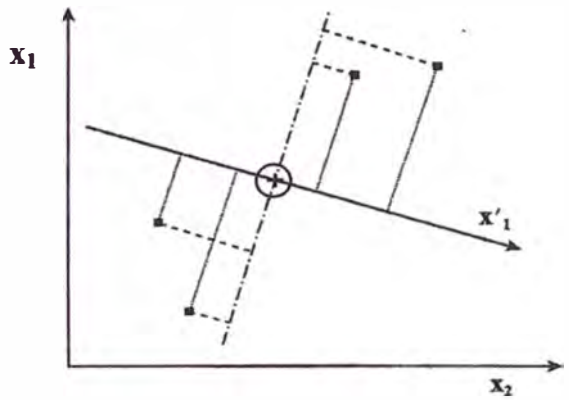


Figura 13. Significado de la varianza explicada y residual respecto a un vector cualquiera que pasa por el centroide de cuatro puntos.

Cuando los datos están fuertemente expandidos en la dirección del vector, éste describe adecuadamente la tendencia de la nube de datos. En tal caso, el vector constituye un buen modelo de la nube multidimensional, la varianza explicada por el mismo es alta y, en consecuencia, la varianza residual es pequeña. Por ejemplo, en la Figura 14 se ha construido con los puntos de la Figura 13 un nuevo modelo, $\mathbf{u}_1$, que es mejor que $\mathbf{x}'_1$. Las distancias en la dirección de $\mathbf{u}_1$ (líneas de trazos) son más largas que las distancias de los puntos al modelo (líneas de puntos), lo que indica que la varianza explicada por $\mathbf{u}_1$ es considerablemente mayor que la residual.

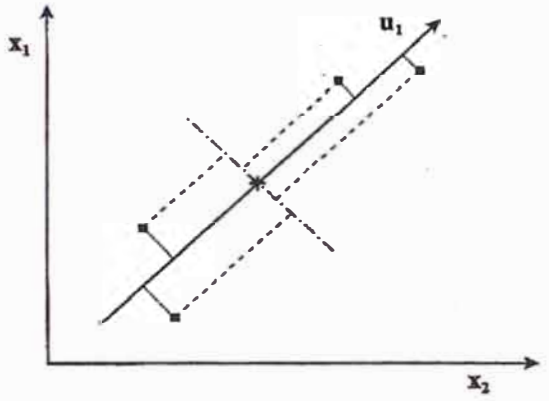


Figura 14. El primer componente principal de la nube de cuatro puntos, u_1, hace máxima la varianza explicada (el autovalor) y hace mínima la residual.

La determinación de los componentes principales correspondientes a una nube de dato se hace, precisamente, *siguiendo el criterio de hacer máxima la varianza explicada*, o lo que es igual, hacer mínima la varianza residual. Así, $\mathbf{u}_1$ es el primer componente principal de los cuatro puntos de la Figura 14, puesto que coincide con la dirección del espacio definido por $\mathbf{x}_1$ y $\mathbf{x}_2$ que hace máxima la varianza explicada y mínima la residual.

Es posible, dada una matriz **X** con *m* variables manifiestas linealmente independientes, construir tantos componentes principales como dimensiones tiene el espacio definido por las variables manifiestas. Cuando se han construido todos los componentes principales, la varianza residual se reduce a cero.

La matriz **U** de los componentes principales se obtiene a partir de la matriz **X** de las variables manifiestas, teniendo ambas matrices las mismas dimensiones, *n* x *m*. Los vectores que definen el nuevo espacio, $\mathbf{u}_1$, $\mathbf{u}_2$ ..., $\mathbf{u}_m$, se han formado a partir de los vectores originales, $\mathbf{x}_1$, $\mathbf{x}_2$ ..., $\mathbf{x}_m$ si bien, se impone la condición de que los componentes principales deben ser ortogonales entre ellos. Ortogonalidad quiere decir que el producto escalar de dos componentes principales cualesquiera es siempre cero ($\mathbf{u}_p\mathbf{u}_q$,= 0), o lo que es igual, sus correlaciones bivariadas son cero. El primer componente, $\mathbf{u}_1$, sigue la dirección de máxima varianza explicada, mientras que la posición del segundo componente en el plano, $\mathbf{u}_2$, queda fijada por la condición de ortogonalidad con el primero.

La dirección de $\mathbf{u}_1$ se ha optimizado para una máxima varianza explicada "entre los puntos de la nube", ignorando la existencia de categorías. Si se tuviesen en cuenta las categorías presentes, se estaría fuera del ámbito del análisis exploratorio, y estaríamos en el campo que corresponde al análisis supervisado. Se dice que los componentes principales "modelan" pero no "discriminan".

El objetivo del análisis de componentes principales (PCA) es reducir las dimensiones del espacio definido por las variables manifiestas a un espacio de dos o tres dimensiones, definido por los dos o tres primeros componentes principales.

Cuando son tres las variables manifiestas, y ya se ha construido el primer componente, quedan aún dos dimensiones para fijar la posición del segundo componente. Puesto que los componentes deben ser ortogonales entre sí, $\mathbf{u_2}$ y $\mathbf{u_3}$ forman un plano perpendicular a la dirección de $\mathbf{u_1}$. Como se esquematiza en la Figura 15, la posición de $\mathbf{u_2}$ en dicho plano queda fijada atendiendo a una regla análoga a la utilizada para construir el primer componente: la varianza no explicada por $\mathbf{u_1}$ se divide en dos, la explicada por $\mathbf{u_2}$, que tiene que ser la máxima posible, y la residual. La varianza explicada por $\mathbf{u_2}$ es la suma de cuadrados de las distancias de los puntos al centroide en la dirección marcada por $\mathbf{u_2}$ o varianza de las proyecciones de los puntos sobre $\mathbf{u_2}$. Como antes, esta varianza será grande si $\mathbf{u_2}$ describe adecuadamente una tendencia en los datos.

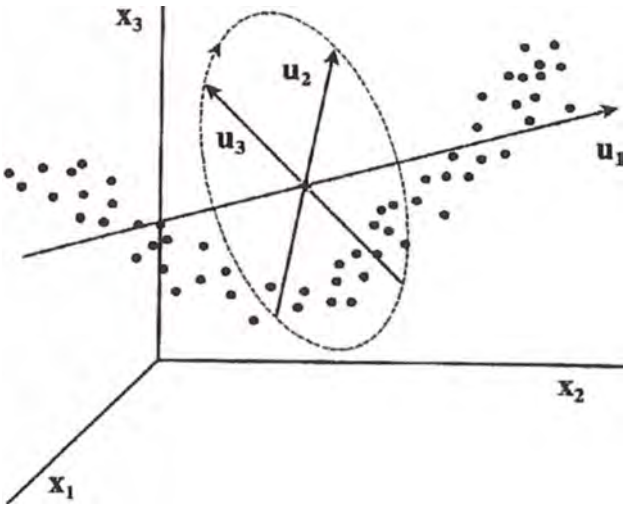


Figura 15. Optimización de la dirección del segundo y tercer componentes en un espacio de tres dimensiones

Una vez que se han construido dos o más vectores, se debe considerar la *varianza explicada acumulada*, que se calcula como la suma de sus varianzas

explicadas. Así, la varianza acumulada por $\mathbf{u}_1$ y $\mathbf{u}_2$ es la explicada por el modelo constituido por esos dos vectores, o suma de cuadrados de distancias de los puntos al centroide en las dos direcciones que indican los vectores, dividida por *n*. Por su parte, la varianza residual es la suma de los cuadrados de las distancias de los puntos al plano formado por esos dos vectores, dividida por *n [1,15]*.

Si los dos componentes principales describen bien las tendencias de los datos, y lo que queda son pequeñas desviaciones aleatorias por encima y por debajo del plano, entonces, la varianza explicada acumulada será grande, y la residual será muy pequeña. Finalmente, la posición de $\mathbf{u}_3$ queda fijada por $\mathbf{u}_1$ y $\mathbf{u}_2$ puesto que tiene que ser ortogonal a ambos y, en este caso, el espacio sólo tiene tres dimensiones .

3.1.3. Varianzas explicada y residual en función del número de vectores del modelo

Cuando se tienen *m* variables manifiestas ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_m$) los *m* posibles componentes principales ($\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_m$) se construyen de uno en uno, haciendo que la varianza explicada (Figura 16) por el nuevo componente sea la máxima posible, si bien, manteniendo la ortogonalidad respecto a los componentes ya construidos *[15]*.

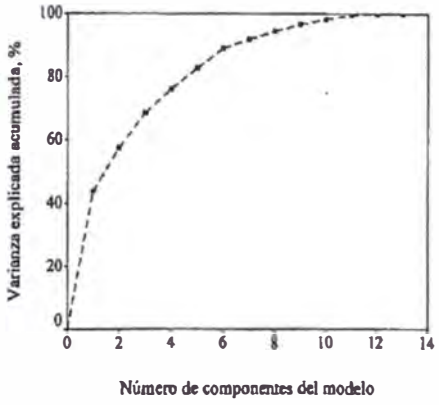


Figura 16. Varianza explicada acumulada(%) en función del número de componentes

Cuando se agrega un nuevo vector al nuevo modelo de p vectores ($p \leq m$) se incrementa la varianza explicada acumulada y se reduce la varianza residual (Figura 17). Cuando el modelo contiene $p = m$ vectores, la varianza explicada acumulada es el 100% de la total y la varianza residual es cero. Las varianzas explicadas por los vectores individuales, así como la residual, se pueden expresar como porcentajes de la varianza total *[1,15]*.

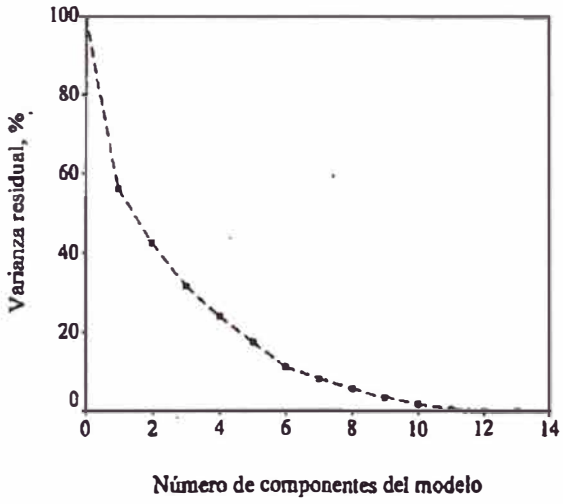


Figura 17. Varianza residual(%) en función del número de componentes.

Para obtener la *varianza explicada acumulada* por los k primeros componentes se suman sus varianzas explicadas (a las que se llaman *autovalores)*, que suelen representarse como λ_p, y se divide por la suma de todas ellas, que es la varianza total.

$$\text{Varianza explicada acumulada (\%)} = \frac{\sum_{p=1}^{k} s_p^2}{\sum_{p=1}^{m} s_p^2} \times 100\% = \frac{\sum_{p=1}^{k} \lambda_p}{\sum_{p=1}^{m} \lambda_p} \times 100\%$$

La varianza explicada acumulada por un modelo constituido por k vectores ($k \leq m$) indica el porcentaje de información contenido en el modelo respecto a la información total contenida en la matriz de datos.

Es también útil el uso del "gráfico de sedimentación" (Figura18), donde se representan directamente los autovalores frente a su número de orden. Debido a la forma en la que se han construido los componentes principales, sus varianzas explicadas están ordenadas de mayor a menor, esto es: $\lambda_1 > \lambda_2 > \ldots\ldots\ldots > \lambda_m$ *[1,5]*.

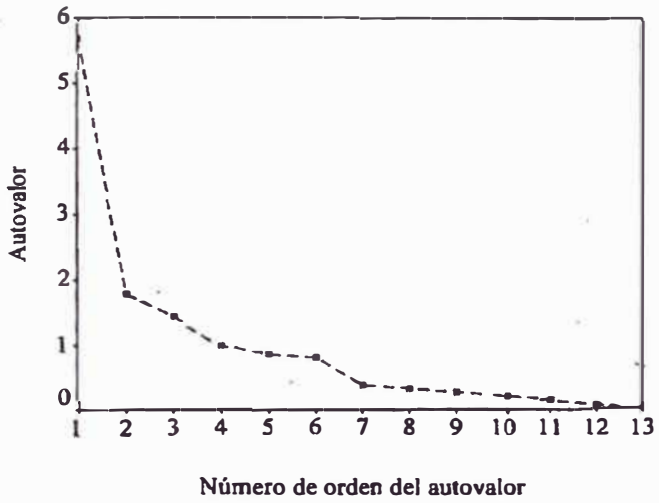


Figura 18. Gráfico de sedimentación

En este gráfico de sedimentación se observarán saltos bruscos. Estos saltos bruscos indican el rango aproximado de la matriz de datos y, por tanto, permiten establecer el número de dimensiones necesarias para modelar, con un error pequeño o aceptable, las estructuras presentes.

3.2. DESCOMPOSICIÓN DE UNA MATRIZ EN PUNTUACIONES, CARGAS Y AUTOVALORES

3.2.1. Descomposición de una matriz de rango *k*.

La construcción de los componentes principales implica hacer uso de la rotación propia, lo que significa obtener un conjunto de nuevas coordenadas de los objetos, $\mathbf{u}_{ip}$, sobre unos nuevos ejes, los componentes principales $\mathbf{u}_1$, $\mathbf{u}_2$,...$\mathbf{u}_p$,...$\mathbf{u}_m$, que son sumas ponderadas de las variables manifiestas, $\mathbf{x}_1$, $\mathbf{x}_2$,...$\mathbf{x}_j$, ..., $\mathbf{x}_m$.. Además, en general, se utilizan componentes normalizados, esto es, divididos por sus correspondientes varianzas, λ_p. Para trabajar adecuadamente, suele ser necesario escalar la matriz de las variables manifiestas, **X**, realizando al menos un centrado por columnas, y con frecuencia también un autoescalado (centrado más división por s_j). En lo que sigue, se supondrá que **X** es una matriz al menos centrada. Así, en un espacio bidimensional, definido por las variables $\mathbf{x}_1$ y $\mathbf{x}_2$, conteniendo *n* objetos, las coordenadas de un objeto cualquiera *i*, $\mathbf{x}_{i1}$ y $\mathbf{x}_{i2}$, se transforman de acuerdo con el sistema de ecuaciones *[1,5]*:

$$u_{i1}\lambda_1 = v_{11}x_{i1} + v_{12}x_{i2}$$

$$u_{i2}\lambda_2 = v_{21}x_{i1} + v_{22}x_{i2}$$

donde $\mathbf{u}_{i1}$ y $\mathbf{u}_{i2}$ (i = 1,2, ..., n) son las puntuaciones o proyecciones de los objetos sobre los componentes $\mathbf{u}_1$ y $\mathbf{u}_2$. Los coeficientes, $\mathbf{v}_{11}$, $\mathbf{v}_{12}$, $\mathbf{v}_{21}$ y $\mathbf{v}_{22}$, se llaman "cargas" o "saturaciones", e indican la importancia que tienen cada variable manifiesta en cada uno de los componentes principales. En general, para *m* variables

manifiestas, las coordenadas de un punto cualquiera *i* en el espacio *m*-dimensional de los componentes se calculan de acuerdo con:

$$u_{i1}\lambda_1 = v_{11}x_{i1} + v_{12}x_{i2} + \ldots + v_{1j}x_{ij} + \ldots + v_{1m}x_{im}$$

$$u_{i2}\lambda_2 = v_{21}x_{i1} + v_{22}x_{i2} + \ldots + v_{2j}x_{ij} + \ldots + v_{2m}x_{im}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$u_{ip}\lambda_p = v_{p1}x_{i1} + v_{p2}x_{i2} + \ldots + v_{pj}x_{ij} + \ldots + v_{pm}x_{im}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$u_{1m}\lambda_m = v_{m1}x_{i1} + v_{m2}x_{i2} + \ldots + v_{mj}x_{ij} + \ldots + v_{mm}x_{im}$$

Este sistema de ecuaciones se puede expresar en forma matricial como sigue:

$$\mathbf{U\Lambda = XV}$$

donde **U** es la matriz *n* x *m* de las puntuaciones de los *m* componentes principales, que son vectores ortonormales, esto es, a la vez ortogonales entre sí y normalizados a la unidad, **X** es la matriz *n* x *m* de las variables manifiestas, y **V** es la matriz cuadrada *m* x *m* de las cargas, constituida por *m* vectores, uno por cada variable manifiesta: $\mathbf{v}_1$, $\mathbf{v}_2$..., $\mathbf{v}_j$ ..., $\mathbf{v}_m$ (columnas de **V**), y también por *m* vectores transpuestos ortonormales, uno por cada componente principal (filas de **V**). *La ortonormalidad de* **U** *y* **V** *son condiciones impuestas que permiten obtener una solución única cuando se resuelve la ecuación matricial [1,7].*

Como se dijo antes, una columna *j* de la matriz de las cargas, $\mathbf{v}_j$, contiene los *m* coeficientes por los que se multiplica una variable manifiesta dada, $\mathbf{x}_j$, en el cálculo de las puntuaciones (nuevas coordenadas de los objetos) sobre los *m* componentes principales, esto es, contiene las "contribuciones" o pesos de la variable manifiesta $\mathbf{x}_j$

a cada uno de los componentes.

Por su parte, una fila *p* de la matriz **V** (o vector $\mathbf{v_P}^T$) contiene los *m* coeficientes por los que se multiplican cada una de las variables manifiestas para hallar las puntuaciones sobre el componente *p*, esto es, contiene las "contribuciones" de las *m* variables manifiestas a la construcción de ese componente. Por su parte, **Λ** es una matriz diagonal cuadrada (todos los elementos distintos de la diagonal son cero) de dimensiones *m* x *m*, que contiene los autovalores, varianzas explicadas o "pesos" de los componentes.

Por ejemplo para tres variables manifiestas y cuatro objetos, la ecuación quedaría como:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \end{bmatrix} x \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} x \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}$$

De acuerdo con lo explicado más arriba la suma de elementos de la diagonal o "traza" de la matriz **Λ** es la varianza total de los datos:

$$tr(\wedge) = \sum_{p=1}^{m} \lambda_p$$

y la varianza explicada por un componente principal $\mathbf{u_p}$ se expresaría del modo siguiente:

$$\text{Varianza explicada por } \mathbf{u_p} \% = \frac{\lambda_p}{tr(\wedge)} x100\%$$

La ecuación **UΛ = XV** se puede reordenar como sigue :

$$\mathbf{X = U\Lambda V^T}$$

que indica cómo se ha descompuesto **X** en las matrices normalizadas de puntuaciones y de cargas, **U** y **V**, y en el conjunto de autovalores recogidos en Λ (ver Figura 19). Cualquier matriz de rango *k* se puede descomponer de este modo, siendo la solución única, salvo en lo que respecta a los signos de las filas de **U** y **V** que pueden aparecer invertidos. Al resolver la ecuación se obtienen *k* autovalores reales, positivos y decrecientes *[1,5]*.

Puesto que **U** y $\mathbf{V}^T$ son ortonormales, los productos escalares de todos sus vectores tomados de dos en dos son cero. Así, para todo par de vectores de **U**, $\mathbf{u}_p$ y $\mathbf{u}_q$ resulta $\mathbf{u}_p\mathbf{u}_q = 0$ y lo mismo sucede para todo par de vectores de $\mathbf{V}^T$, $\mathbf{v}^T_p\,\mathbf{v}^T_q = 0$.

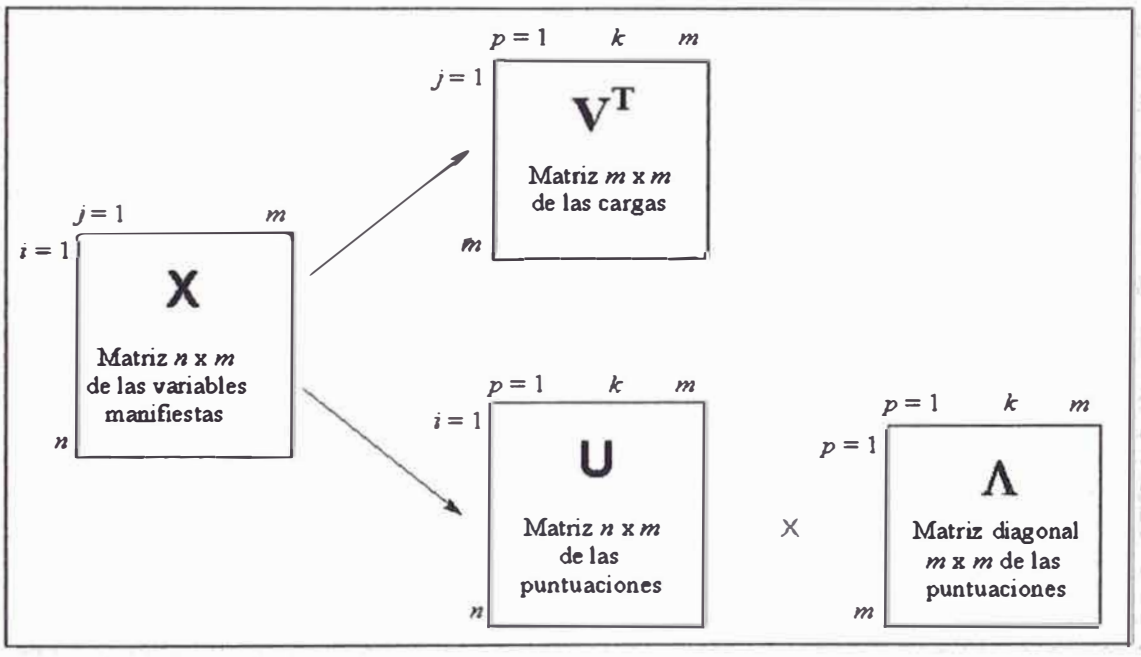


Figura 19. Descomposición de la matriz X en la matriz de cargas, V, y en el producto de la matriz de las puntuaciones normalizadas, U, por la matriz diagonal de los autovectores, Λ. Los índices son: objeto i, variable manifiesta j, y el componente principal p.

Esto último significa que: todos los elementos de la matriz de correlaciones, **R,** construida con los vectores de una matriz ortogonal son cero, excepto los elementos de la diagonal que son todos la unidad (esto es, $\mathbf{R} = \mathbf{I}$, donde **I** es la matriz identidad). Los componentes principales son, por tanto, variables "incorreladas" entre ellas *[5,7]*.

La normalización de **U** por columnas implica que las longitudes o módulos de todos los componentes principales son la unidad (los cuadrados de los elementos de cada columna de **U** suman la unidad). Además, la información relacionada con las posiciones relativas de los puntos queda contenida en **U**, mientras que la información relacionada con la longitud de cada vector se encuentra en Λ. La normalización de **U** permite trabajar con puntuaciones normalizadas, que dan lugar a gráficos más ilustrativos y más fáciles de interpretar que los obtenidos con puntuaciones no normalizadas.

El gráfico de puntuaciones sobre componentes normalizados contiene información relativa tan sólo a la disposición de unos puntos respecto a otros, en un espacio cuyos ejes tienen todos la misma longitud. La información relacionada con la distinta varianza explicada, o "estiramiento" de la nube de puntos a lo largo de los componentes ha quedado separada del gráfico de puntuaciones, y está recogida en los autovalores (Figura 20) *[1]*.

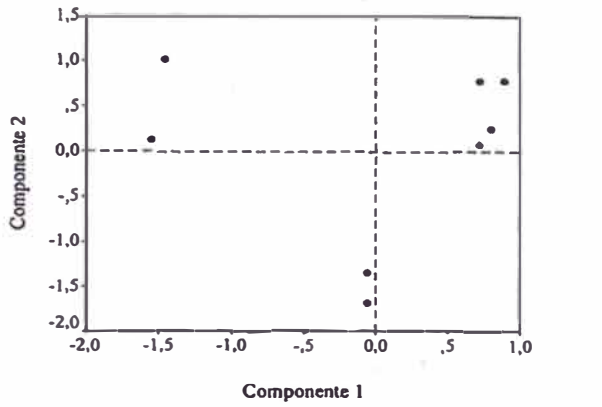


Figura 20. Diagrama de puntuaciones sobre el plano u_1u_2 para dos variables

3.2.2. Reducción de dimensiones

La más importante de las aplicaciones de la rotación propia es la reducción de la dimensionalidad del espacio en el que se representan los objetos, o en su caso las variables. En la rotación propia se ha modelado la nube de datos, obteniendo vectores ordenados de acuerdo a su varianza explicada o capacidad descriptiva. El siguiente paso es depurar la información, reteniendo tan sólo los mejores vectores. Si se toma un número reducido de componentes, de 1 a *k* (líneas de trazos de la Figura 20), las matrices **U**, **Λ** y **V** reducidas contienen sólo información estructural, y la información irrelevante más el ruido aleatorio quedan eliminados. La información relevante queda representada, con cierto error, en un espacio reducido a *k* dimensiones. El espacio reducido queda descrito por:

- Las *k* primeras columnas de **U**, que contienen las puntuaciones normalizadas.
- Las *k* primeras filas de **V**, que dan información sobre las variables manifiestas que construyen los componentes retenidos.
- Los *k* primeros autovalores, que indican el peso o longitud de cada

componente retenido, y que permiten calcular la cantidad de varianza explicada o porcentaje de información retenida por cada componente, y por el modelo formado por los k primeros componentes, o por cualquier otra combinación de vectores.

La proyección de los puntos sobre el plano $\mathbf{u_1u_2}$ ofrece el mayor porcentaje de varianza explicada que puede obtenerse en dos dimensiones. Análogamente, las proyecciones de los puntos sobre los planos $\mathbf{u_1u_3}$ y $\mathbf{u_2u_3}$ ofrecen los mayores porcentajes de varianza explicada que pueden obtenerse en dos dimensiones ortogonales a $\mathbf{u_1u_2}$ y ortogonales entre sí, y así sucesivamente. Sin embargo, con frecuencia, la observación del plano $\mathbf{u_1u_2}$ es suficiente para tener una idea satisfactoria de la forma de la nube de datos en el espacio *m*-dimensional, y se recurre a las proyecciones sobre $\mathbf{u_1u_3}$ y $\mathbf{u_2u_3}$ tan sólo cuando es necesario (Figura 21).

El análisis matemático correspondiente a la construcción de los componentes principales y el algoritmo que permite su transformación en un programa informático, se encuentra en el Anexo 1.

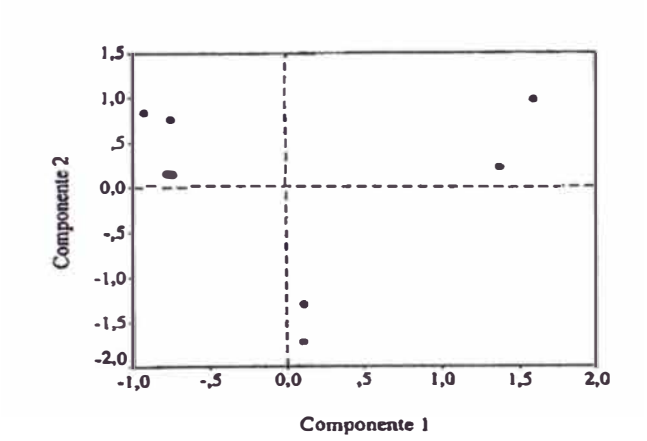


Figura 21 . Diagrama de puntuaciones sobre el plano u_1u_2 para cuatro variables

3.2.3. Selección del número óptimo de componentes

Una vez obtenido los componentes principales debemos disponer de ciertos criterios que indiquen cuántos vectores son realmente necesarios, esto es, cuál es el valor de k para el que se puede considerar que las estructuras presentes han sido convenientemente modeladas. La selección del valor óptimo de k puede hacerse en función de los siguientes criterios *[5,7]*:

- Por el número de fuentes significativas de varianza. El número de vectores que es necesario retener en el modelo no es mayor que el número de fuentes significativas de varianza presentes en los datos. El conocimiento previo del sistema en estudio permite tener una idea aproximada del valor óptimo de k.
- Por el porcentaje satisfactorio de varianza explicada acumulada, tal como, por ejemplo, el 60% o el 70%. El porcentaje más adecuado varía ampliamente dependiendo del problema, pero este procedimiento puede ser útil cuando se realizan estudios sobre problemas de un mismo tipo, para los cuales se ha establecido previamente un porcentaje adecuado de varianza explicada acumulada.
- Mediante el estudio gráfico de los autovalores, o de las varianzas explicada, acumulada y residual. Con frecuencia, el gráfico de sedimentación suele mostrar una zona inicial de caída brusca, seguida de otra donde la caída es gradual. Se observan también cambios análogos en los gráficos de las varianzas explicada acumulada y residual frente al número de orden de los

componentes. Se toma como valor óptimo de *k* el orden del autovalor o del componente donde se produce el cambio de caída brusca a gradual.

- Mediante el estudio gráfico de cocientes de autovalores sucesivos. Este gráfico de "cocientes de autovalores sucesivos" (λ_1/λ_2, λ_2/λ_3, λ_3/λ_4, etc) convierte en "picos" los cambios bruscos de pendiente del gráfico de sedimentación. El valor buscado de *k* es igual al número de orden del autovalor que figura como denominador cuando se produce el primer pico.

3.3. ANÁLISIS DE LA MATRIZ DE LAS CARGAS

Se ha visto que las puntuaciones normalizadas o elementos de **U** ofrecen información sobre las relaciones entre los objetos, y los autovalores o elementos de la diagonal de Λ indican la varianza o información retenida por cada componente principal y, por tanto, su importancia relativa. Del mismo modo, y de acuerdo a Ramis *[1]*, la matriz **V** de las cargas contiene información sobre las variables, es decir, indica cómo están correlacionadas las variables manifiestas entre sí y con los componentes principales. Sólo tienen interés las *k* primeras filas de **V**, puesto que los elementos de estas filas son los coeficientes que construyen los *k* primeros componentes principales. El análisis de la matriz **V** se realiza de dos modos:

- Por filas, mediante la observación directa o gráfica (diagrama de barras) de los elementos de la fila o cargas correspondientes a un componente determinado.
- Por columnas (desde su primer elemento hasta el *k),* mediante diagramas de dispersión en los que las coordenadas de los puntos son parejas de cargas

tomadas de la misma columna de **V** (primer y segundo elementos de cada columna, primer y tercero, segundo y tercero, etc.), por lo que cada punto representa a una determinada variable manifiesta.

3.3.1. Estudio de las cargas de un componente

El estudio de las primeras filas de la matriz **V** permite saber qué variables manifiestas contribuyen en mayor medida a la varianza de cada componente, y también indica si el componente tiene carácter de cantidad o de contraste. Las cargas son las correlaciones entre las características, $\mathbf{x}_j$, y los componentes, $\mathbf{u}_p$, esto es, son los cosenos *de* los ángulos que forman las características con los componentes, ambos centrados y normalizados. Una carga alta, v_{pj}, a lo largo de una fila cualquiera *p* indica que $\mathbf{x}_j$ contribuye en gran medida a construir $\mathbf{u}_p$. También se puede decir que $\mathbf{x}_j$ cede gran parte de su varianza a $\mathbf{u}_p$, o bien, que $\mathbf{u}_p$ explica gran parte de la varianza aportada por $\mathbf{x}_j$.

Las cargas altas sobre un determinado componente indican correlación entre las variables manifiestas implicadas entre sí y con el componente. Por el contrario, las cargas bajas indican independencia entre las variables manifiestas y el componente principal *[1,5]*.

Un diagrama de barras resulta muy útil para comparar las magnitudes de las cargas sobre un componente. Para distinguir mejor las cargas altas de las bajas, es conveniente ordenarlas de mayor a menor, y prescindir de su signo. Las Figuras 22, 23 y 24 muestran los diagrama de barra de los tres primeros componentes de una muestra de agua de lluvia.

El primer componente muestra cargas altas sobre Na^+ y Cl^-, indicando que la primera fuente de varianza es la influencia marina. Las mayores cargas sobre el segundo componente tienen también un significado claro desde el punto de vista químico-físico, puesto que se trata de metales de la mena del cobre. Por tanto, el segundo componente explica varianza debida a una determinada industria (una fundición de cobre). Las cargas mayores sobre el tercer componente, nitrato y H^+, obedecen también a una causa común, la liberación de óxidos de nitrógeno por parte de los motores de combustión y su posterior oxidación aérea a ácido nítrico. Por consiguiente, las cargas sobre $\mathbf{u_3}$ apuntan al tráfico rodado como la tercera fuente de la varianza de los datos *[1]*.

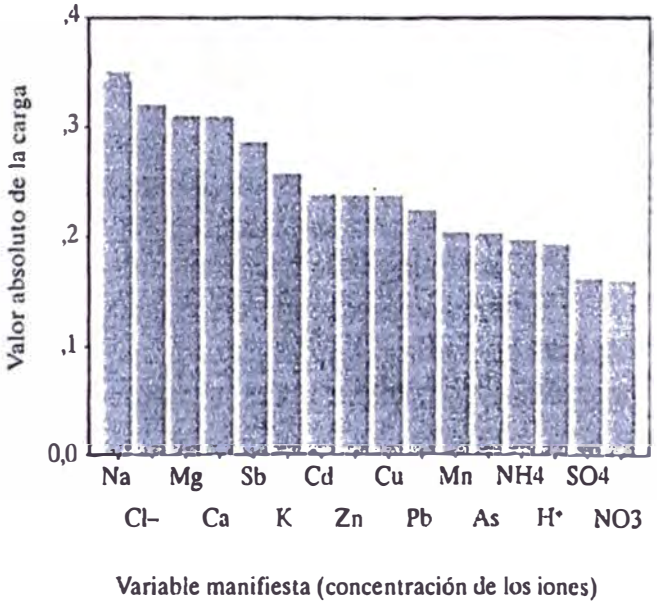


Figura 22. Diagrama de barras de las cargas sobre u_1 para un muestra de agua de lluvia

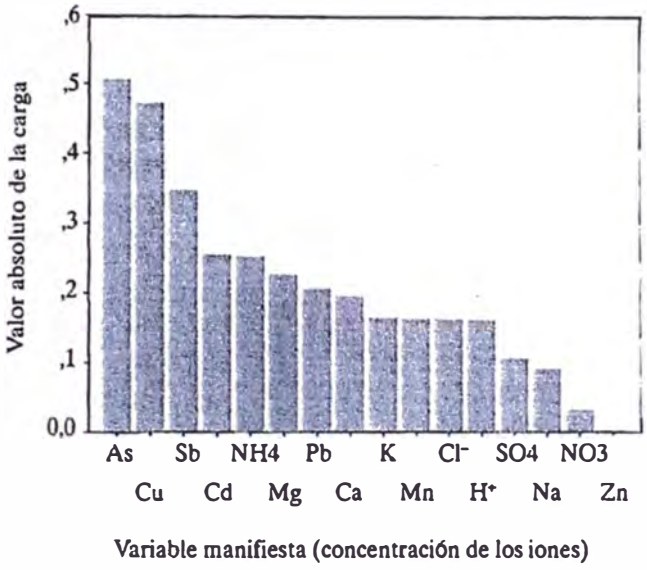


Figura 23. Diagrama de barras de las cargas sobre u_2 para una muestra de agua de lluvia

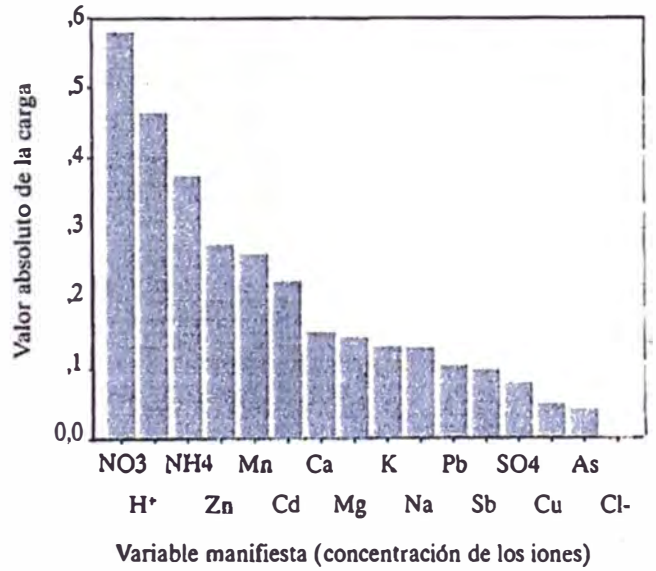


Figura 24. Diagrama de barras de las cargas sobre u_3 para una muestra de agua de lluvia

La interpretación de los demás componentes, con porcentajes menores de la varianza total, no es fácil ni segura, por lo que en general no se interpretan.

También los signos de las cargas nos traen información. Estos signos indican si $\mathbf{x}_j$ y $\mathbf{u}_p$ están directa o inversamente correlacionadas. Una carga alta y a la vez positiva indica un ángulo pequeño entre ambos vectores en el espacio *m*-dimensional, mientras que una carga alta pero negativa indica un ángulo próximo a 180°. Más aún, si las cargas altas sobre un componente tienen todas ellas signos positivos, el componente indica cantidad, en cambio si las cargas altas tienen signos positivos y negativos, el componente indica contraste entre propiedades o entre variables fundamentales *[1]*.

3.3.2. El diagrama doble

El llamado *diagrama doble* o *biplot* se obtiene superponiendo los diagramas de puntuaciones y cargas para un mismo plano. El diagrama doble informa sobre las relaciones entre los objetos y los grupos formados por las variables manifiestas, y por ello permite extraer conclusiones acerca de las relaciones entre los objetos y las variables fundamentales.

Estos diagramas interpretan preferentemente las relaciones entre objetos y variables que se encuentran a grandes distancias respecto al origen de coordenadas, esto es, los que son "casi" coplanares entre sí y con el plano observado. Si un grupo de objetos se encuentra alejado del origen en la dirección marcada por una o más variables manifiestas con cargas altas sobre el plano observado, es porque dichos

objetos tienen valores altos de dichas variables. Finalmente, si un grupo de objetos muestra una relación de "ortogonalidad" respecto a algunas variables, es porque dichas variables contienen aleatoriamente valores altos y bajos para esos objetos, esto es, los objetos no están asociados con esas variables.

Una interesante interpretación del biplot es que si las variables se encuentran cerca del origen de coordenadas son aproximadamente ortogonales respecto al plano de observación, por lo que no puede afirmarse nada más sobre ellas. Debe tenerse en cuenta que los objetos cercanos al origen, sobre el plano de observación, pueden estar en realidad lejos del mismo en el espacio multidimensional. Por ello, una variable que aparece cerca del origen puede estar correlacionada o no con los objetos que también se encuentran cerca del origen *[1]*.

3.4. PRUEBA PARA VALORES ATÍPICOS O ANÓMALOS

Una importante aplicación de las pruebas Estadísticas es el reconocimiento de los valores atípicos o anómalos (outliers). Los valores atípicos en una serie de medidas son aquellos que evidentemente están alejados de la nube de datos. Por ejemplo al aplicar PCA a un conjunto de datos podríamos obtener la gráfica de la Figura 25.

En la figura 25 el valor titulado 95 está evidentemente muy alejado de la nube de datos por lo que es considerado un valor atípico. Esta apreciación visual evidente, sin embargo, puede sustentarse aplicando algunas pruebas estadísticas, ya que la eliminación sin criterios de estos outliers puede llevar a conclusiones erróneas.

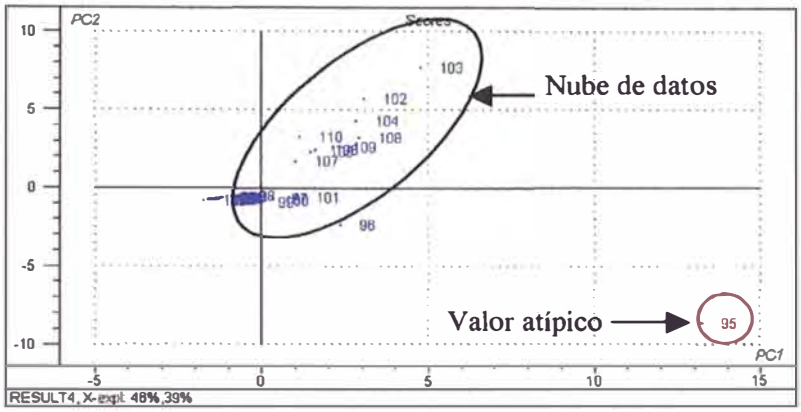


Figura 25. Nube de datos y valor atípico

El criterio estadístico que se emplea para la eliminación de valores atípicos, cuando la oblación de medidas esta entre 3 y 150 es la prueba de Grubbs. Esta prueba asume que la población esta normalmente distribuida. Un valor x^* no es considerado un outlier dentro de una serie de n medidas, con un nivel de aceptación α, si el valor de prueba T es :

$$T = \frac{|\bar{x} - x^*|}{s} < T(1-\alpha; n)$$

donde T (1 - α; n) son los valores críticos de la prueba de Grubbs que están dados en la Tabla 2 *[15]*.

Así por ejemplo, para un conjunto de 7 valores (n = 7) de media $\bar{x}$ = 5,29 y desviación estándar s = 0,411, el valor x = 6,00 puede ser considerado un outlier puesto que:

$$T_7 = \frac{|5{,}29 - 6{,}00|}{0{,}411} = 2{,}21$$

es mayor que T (1- α = 0,99; n=7) = 2,10 a un nivel de significancia α = 0,01.

Tabla 2. Valores críticos para la prueba de Grubbs
Para dos niveles de significancia

n	T (0,95; *n*)	T (0,99; *n*)
3	1,15	1,16
4	1,46	1,49
5	1,67	1,75
6	1,82	1,94
7	1,94	2,10
8	2,03	2,22
9	2,11	2,32
10	2,18	2,41
12	2,29	2,55
15	2,41	2,71
20	2,56	2,88
30	2,75	3,10
40	2,87	3,24
50	2,96	3,34

4. ANÁLISIS CLASIFICATORIO

Las técnicas quimiométricas han sido ampliamente aplicadas a la resolución de los distintos problemas de clasificación de sustancias mediante una serie de características o magnitudes físicas de nuestras medidas o determinadas previamente. Las técnicas quimiométricas utilizadas en problemas de análisis cualitativo se conocen de forma general con el nombre de Métodos de Reconocimiento de Pautas (PRM, Pattern Recognition Methods) *[6, 16]*.

En el análisis clasificatorio se construyen modelos capaces de pronosticar la pertenencia de un objeto a una categoría sobre la base de las características del objeto. La matriz de datos contiene al menos una variable categórica, que indica la categoría a la que pertenece cada objeto y que constituye la respuesta o variable que

se quiere predecir, y una o más variables de escala que describen otras tantas características de los objetos y que se utilizan como variables predictoras.

Para construir el modelo es necesario disponer de una muestra de objetos cuya categoría sea conocida y para los que también se conozcan los valores de las variables predictoras. La pertenencia de los objetos a las categorías puede ser supuesta, esto es, puede tratarse una hipótesis a comprobar. La asignación de los objetos a las categorías debe ser exhaustiva (todos objetos pertenecen a una categoría) y mutuamente exclusiva (ningún objeto pertenece a más de una categoría). Estos objetos forman el "conjunto de entrenamiento", con el cual se construye el modelo de clasificación. Una vez construido, el modelo se utiliza para predecir la categoría de nuevos objetos a partir de la medida de las variables predictoras *[1]*.

4.1. MÉTODOS DE RECONOCIMIENTOS DE PAUTAS

Los métodos de reconocimiento de pautas son un conjunto de herramientas quimiométricas que permiten establecer agrupaciones de muestras en función de características comunes o relaciones que existan entre ellas o bien definir métodos de clasificación para muestras desconocidas. Existen una gran variedad de métodos de reconocimiento de pautas y continuamente aparecen nuevas variantes de los ya existentes. La mayoría de métodos de reconocimiento de pautas se basan en la medida de la similitud, parámetro que indica en qué medida un objeto es igual a otro. La manera más común de expresar la similitud es a través de las medidas de correlación o distancias *[6]*.

i. Medidas de correlación: se basan en el cálculo del coeficiente de correlación entre dos muestras:

$$r_{jk} = \frac{\sum_{i=1}^{p}(x_{ij} - x_j)(x_{ik} - x_k)}{\sqrt{\sum_{i=1}^{p}(x_{ij} - x_j)^2 \sum_{i=1}^{p}(x_{ik} - x_k)^2}}$$

Oscila de –1 a +1. El valor de 1 indica coincidencia total entre los dos conjuntos de datos *[16]*.

ii. Medidas de distancia: estas medidas se basan en el cálculo de una distancia que representa cuan diferente es una muestra de otra o bien de un punto en el espacio que represente el modelo de una clase. El cálculo de la distancia D entre una muestra x_i y el centroide de una clase μ se determina mediante la distancia euclídea y la distancia de Mahalanobis:

$$\text{Distancia euclídea} = d_{kl} = \left[\sum_{j=1}^{m}(x_{kj} - x_{lj})^2\right]^{1/2}$$

La distancia de Mahalanobis es una generalización de la distancia euclídea para cuando existe correlación entre las características *[16]*.

Los métodos de reconocimiento de pautas pueden clasificarse según se conozca a priori o no, la pertenencia de los objetos a clases determinadas siendo denominados respectivamente métodos supervisados y métodos no supervisados *[6]*.

4.1.1. Métodos no supervisados

Se basan en descubrir agrupaciones de pautas en el espacio de *n*–dimensiones sin saber a priori a qué clase pertenece cada muestra. Algunos de los más comunes son *[6]*:

- *Análisis de clusters*: bajo esta denominación quedan englobados toda esta serie de métodos diseñados para entender la estructura de una gran matriz de datos, reconociendo similitudes entre objetos (o variables), y así llegar a distinguir algunas clases, que serán conjuntos de objetos similares.
- *Minimal Spanning Tree (MST):* se basa en conectar puntos (objetos) de forma que la longitud total es la mínima de todas las combinaciones posibles. El algoritmo busca todas las distancias entre objetos la máxima y divide los objetos conectados en dos clusters. Sigue haciendo lo mismo en cada uno de los nuevos clusters, hasta cumplir una condición impuesta previamente.
- *Redes neuronales no supervisadas (Kohonen):* Las redes neuronales se definen como un sistema iterativo de cálculo que intenta reproducir, de forma simple y sencilla, el sistema de conexiones que existe entre las neuronas del cerebro humano. Este tipo de red halla la neurona que se parece más a un objeto presentado a ella y modifica sus pesos para que se parezca al ejemplo presentado. Después de un número determinado de entradas de los datos a la red, diversas zonas de la red de Kohonen responden a diferentes tipos de las clases presentes en el conjunto de datos.

4.1.2. Métodos supervisados

En estos métodos, la clasificación se basa en un aprendizaje previo del sistema, con conjuntos de calibración (o entrenamiento) de objetos que definirán cada clase. Estos objetos son de conocida pertenencia a una de las clases. La calidad de los resultados de clasificación vendrá influenciada por la calidad de los conjuntos de entrenamiento. Estos método pueden ser divididos en dos subgrupos: métodos discriminantes y métodos de modelado *[6, 16]*.

i. Métodos discriminantes

Dividen el espacio en tantas regiones como clases halla en el conjunto de entrenamiento creando unos límites compartidos por los espacios. Siempre clasifican una muestra desconocida como perteneciente a una de las clases. Los más comunes son:

- *Análisis discriminante (DA):* Están basados en el concepto de una función discriminante que divide el espacio en regiones características para cada una de las clases ($\mathbf{f} = a_1\mathbf{x_1} + a_2\mathbf{x_2} + \ldots + a_m\mathbf{x_m}$), creando fronteras entre cada una de ellas. Los dos métodos más conocidos son Análisis Discriminante Lineal (LDA) y Análisis Discriminante Cuadrático (QDA).
- *KNN (K-nearest neighbour):* Este método clasifica un objeto test como perteneciente a la clase en la cual la mayoría de objetos pertenecientes a ella son más cercanos al objeto test. Generalmente se utiliza la distancia euclídea como medida de distancia.

- *Potential Function Methods (PFM):* Estos métodos, de los cuales el más conocido en química analítica es el denominado ALLOC[ξ], pueden considerarse como generalizaciones del método KNN. En estos métodos, cada objeto de un conjunto de entrenamiento es considerado como un punto en el espacio rodeado por un campo de potencial. La clasificación de un objeto test dentro de una de las clases está determinada mediante el potencial acumulado de la clase en la posición del objeto desconocido. El potencial acumulado se obtiene sumando los potenciales individuales de los objetos de la clase en la posición del objeto desconocido. El objeto test se clasifica dentro de la clase que da el mayor potencial acumulado.

ii. Métodos de modelado

Los métodos de modelado se basan en la creación de unos volúmenes en el espacio, cada uno de ellos con unos límites distintos para cada uno de las clases. Mediante estos métodos, una muestra puede clasificarse como perteneciente a alguna de las clases o a ninguna de ellas. Algunos de los métodos utilizados son *[6]*:

- *PRIMA (Pattern Recognition by Indepent Multicategory Análisis):* En este método cada clase es modelada mediante su centroide, teniendo el modelo una forma esférica alrededor de éste. La semejanza entre un objeto y el modelo de cada clase se mide como la distancia euclídea calculada con las

[ξ] ALLOC es una Técnica de modelamiento como función de las distancias.

variables autoescaladas. La distancia crítica (d_{crit}) define el radio de las esferas de cada clase.

- *UNEQ*: El modelo se construye mediante el centroide de la clase, teniendo forma de hiper-elipsoide, siendo el centroide la media de la población y definiendo su tamaño mediante el valor de la d_{crit}. Teniendo en cuenta que el modelo de cada clase se construye mediante el cálculo de las distancias de Mahalanobis de forma individual, siendo la matriz de dispersión distinta para cada clase, UNEQ es especialmente adecuado para clases que presenten diferente dispersión (UNEQual dispersed classes)
- *Método de varianza residual*: Estos métodos se basan en PCA de cada uno de los conjuntos de entrenamiento, creando un modelo para cada uno de ellos. Un espectro test se reconstruye según todos los modelos, y los residuales obtenidos son utilizados para calcular la probabilidad de que la muestra desconocida pertenezca o no a alguna de las clases. SIMCA[Ψ] es probablemente el más conocido de todos los métodos de varianza residual.
- *Redes neuronales artificiales supervisadas*: Estos métodos muestran un gran potencial en el campo de la clasificación de sustancias, debido a su gran capacidad de modelado, hecho que las hace especialmente adecuadas para la resolución de una amplia variedad de problemas. Existen diferentes tipos de redes neuronales artificiales según sea su proceso de entrenamiento. De entre todas, la más utilizada en el campo químico es la Multi-Layer Perceptron.

[Ψ] SIMCA es la abreviatura para *Soft Independent Modeling of Class Analogies* (método de reconocimiento de patrones y clasificación)

Cualquiera que sea la técnica usada, para construir un buen modelo se deben cumplir las siguientes condiciones *[6, 16]*:

a) El conjunto debe estar constituido por un mínimo de 4 ó 5 objetos independientes de cada categoría, de modo que la dispersión interna de cada categoría quede bien representada.
b) Las variables predictoras deben contener información discriminante, esto es, deben ser capaces de distinguir entre las categorías.

Como se indica en el esquema de la Figura 26 , existe un claro paralelismo entre los problemas de clasificación (predicción cualitativa) y los de predicción cuantitativa (regresión y calibración). Como se puede notar ambas técnicas comparten conceptos iguales o análogos, tales como los de variable predictora y variable respuesta, conjunto de entrenamiento o de calibración, y conjunto de evaluación. La diferencia está en que en regresión la respuesta es una variable de escala, mientras que en análisis clasificatorio se predice una variable categórica, si bien los modelos construidos en análisis clasificatorio se pueden utilizar también para predecir variables de escala *[1]*.

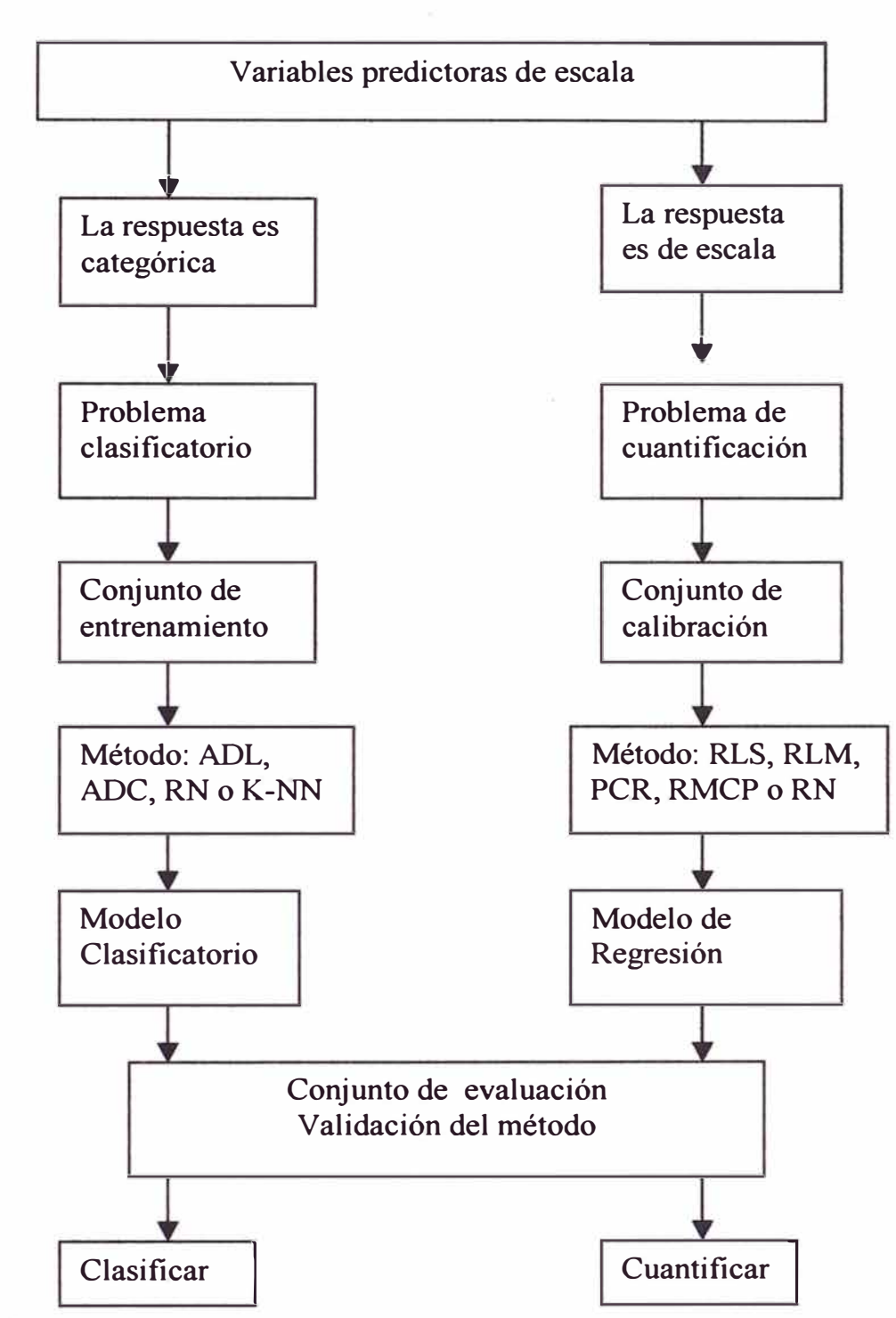


Figura 26: Analogías entre el análisis clasificatorio y las técnicas cuantitativas (RLS =Regresión Lineal Simple, RLM = Regresión Lineal Múltiple, RCP = Regresión de Componentes Principales, RMCP = Regresión de mínimos Cuadrados Parciales)

4.2. VALIDACIÓN DE UN MODELO DE CLASIFICACIÓN

Cuando se construye un modelo que representa una realidad siempre es necesario validarlo o comprobarlo. La validación de un modelo de clasificación consiste en establecer sus capacidades de reconocimiento y de predicción, y su estabilidad o robustez. La capacidad de reconocimiento se evalúa, directamente, como el porcentaje de objetos del conjunto de entrenamiento que son clasificados correctamente *[1, 16]*. Por ejemplo, 48 objetos clasificados correctamente de un total de 50 indica una capacidad de reconocimiento de (48 / 50)x100 = 96%.

Pero, para establecer la capacidad de predicción del modelo, es necesario definir previamente un conjunto de prueba o de evaluación, que debe estar formado obligatoriamente por objetos de los que se conoce sus pertenencias a las categorías, pero que no se incluyeron en el conjunto de entrenamiento con el que se desarrolló el modelo. Así, podemos definir la capacidad de predicción como el porcentaje de objetos del conjunto de evaluación que son clasificados correctamente por el modelo *[1]*. Por ejemplo, si se tienen 25 objetos nuevos (no utilizados para desarrollar el modelo), y de ellos 23 son clasificados correctamente, entonces la capacidad predictiva del modelo es de (23 / 25)x100 = 92%.

Por último, un modelo es estable o robusto si la eliminación de un objeto, o de unos pocos objetos, o la sustitución de unos objetos por otros en el conjunto de entrenamiento y predicción no hace variar sus capacidades de reconocimiento y predicción. Los modelos construidos con un número insuficientes de objetos de entrenamiento son inestables *[1,5]*.

Ambas capacidades, de reconocimiento y de predicción, son necesarias determinarlas, puesto que si sólo se determina la primera es probable que se tenga una idea excesivamente optimista sobre la calidad del modelo. Por otra parte, al formar el conjunto de evaluación, se esta renunciando a utilizar en la construcción del modelo la información contenida en un numeroso grupo de objetos. Se obtendrá así un modelo de peor calidad en relación al que se podría tener si se utilizasen todos los objetos disponibles. Un problema análogo se tiene en la predicción de variables de escala. Tanto en análisis clasificatorio como en calibración, este dilema se resuelve mediante la estrategia conocida como el método del objeto excluido *[5,7]*.

En el método del objeto excluido el modelo se establece con todos los objetos disponibles menos uno. El conjunto de entrenamiento queda constituido por $n - 1$ objetos. El método construye n modelos casi iguales entre sí, y muy similares al que se construirá finalmente con los n objetos. Por ejemplo, si se tienen 50 objetos, se construyen 50 modelos con 49 objetos. Luego se realizan los cálculos siguientes:

a) Se establece la capacidad de reconocimiento de cada uno de los n modelos de $n - 1$ objetos. Si los porcentajes de acierto obtenidos son iguales o casi iguales, se deduce que el modelo construido es robusto o estable.

b) Se establece la capacidad predictiva del modelo de n objetos como el porcentaje de objetos excluidos que han sido clasificados correctamente por los n modelos de $n - 1$ objetos.

Cuando se tienen muchos objetos, es posible excluir varios, cada vez, en lugar de uno solo. Así, si se excluyen p objetos seleccionados siguiendo un plan sistemático, se pueden obtener numerosos modelos con $n - p$ objetos. De este modo, la estabilidad del modelo y su capacidad de predicción se pueden conocer sobre una base estadística más completa y rigurosa. Este procedimiento se ha denominado método de la "navaja", aludiendo a que se "corta" una parte de la matriz de datos cada vez.

Las técnicas del objeto excluido y de la navaja se engloban bajo la expresión "técnicas de validación cruzada".

Para mejorar la capacidades de reconocimiento y predicción, se deben buscar nuevas variables con mayor poder discriminante, o también, utilizar un algoritmo clasificatorio que se adapte mejor a las particularidades del problema. Si las categorías no son linealmente separables, se pueden utilizar el ADC (Análisis Discriminante Cuadrático) o una red neuronal en lugar del ADL (Análisis Discriminante Lineal). Así mismo, para mejorar la estabilidad del modelo se debe ampliar el número de objetos del conjunto de entrenamiento. Especialmente, se deben añadir nuevos objetos a las categorías que tienen pocos, o que los tienen muy dispersos. Además, es importante eliminar los posibles objetos anómalos, que están claramente alejados de cualquiera de las categorías reconocidas *[1,8]*.

CAPÍTULO II
AGUAS DE PRODUCCIÓN EN LA INDUSTRIA DEL PETRÓLEO

La industria petrolera enfrenta muchos problemas ambientales, entre ellos:

· El control de la descarga de contaminantes en la atmósfera,

· El control de la descarga de contaminantes en aguas superficiales,

· La reinyección de agua de formación en aguas subterráneas y

· El manejo de residuos sólidos y peligrosos *[10]*.

1. AGUA DE PRODUCCIÓN

El agua de producción es una mezcla de agua de formación y petróleo que se forma durante la extracción. Debido a los volúmenes producidos (que pueden ir de 2000 a 7000 m^3/día), la forma más práctica de disponer de ella es descargarla al mar, aunque a veces también puede reinyectarse a los pozos. Su composición es altamente variable, de modo que es difícil generalizar acerca de sus impactos ambientales. Sin embargo, se espera que contenga hidrocarburos y sales minerales (de la formación que se está perforando) además de biocidas, inhibidores de corrosión y otros

componentes del agua de inyección y de los desechos de la plataforma. Así, su toxicidad es variable, pero no suele representar una grave amenaza en ambientes dinámicos donde se puede producir dilución *[11]*.

El agua de producción es usualmente tratada en la plataforma para reducir su contenido de hidrocarburos, ya que muchos países limitan la concentración de petróleo que puede contener el agua de producción descargada al mar a 50 mg/L *[11]*.

El principal problema del agua de producción se relaciona con su contenido de petróleo. Suele argumentarse que, aunque los volúmenes de descarga son altos, la concentración de petróleo es baja (por lo general menos de 30 mg/L) y que por ejemplo, en el Mar del Norte, el petróleo descargado en el agua de producción representa sólo el 3,5% de las fuentes de contaminación con petróleo. Sin embargo, en regiones menos industrializadas, estas descargas representarán una fracción mucho mayor del total de hidrocarburos descargados al ambiente. Las descargas de agua de producción continúan a un ritmo estable mientras se extrae petróleo y tienden a aumentar en las etapas finales de producción (ver Figura 27) *[11]*.

2. MONITOREO DE LA CALIDAD DEL AGUA

Con el fin de proteger el medio ambiente de las descargas de agua contaminada proveniente de la extracción y refinación de petróleo, es necesario que las empresas petroleras implementen un programa eficaz de monitoreo de la calidad

del agua. Este programa debe incluir tanto el monitoreo de las descargas de aguas residuales en el medio ambiente como el monitoreo de las aguas receptoras *[11]*.

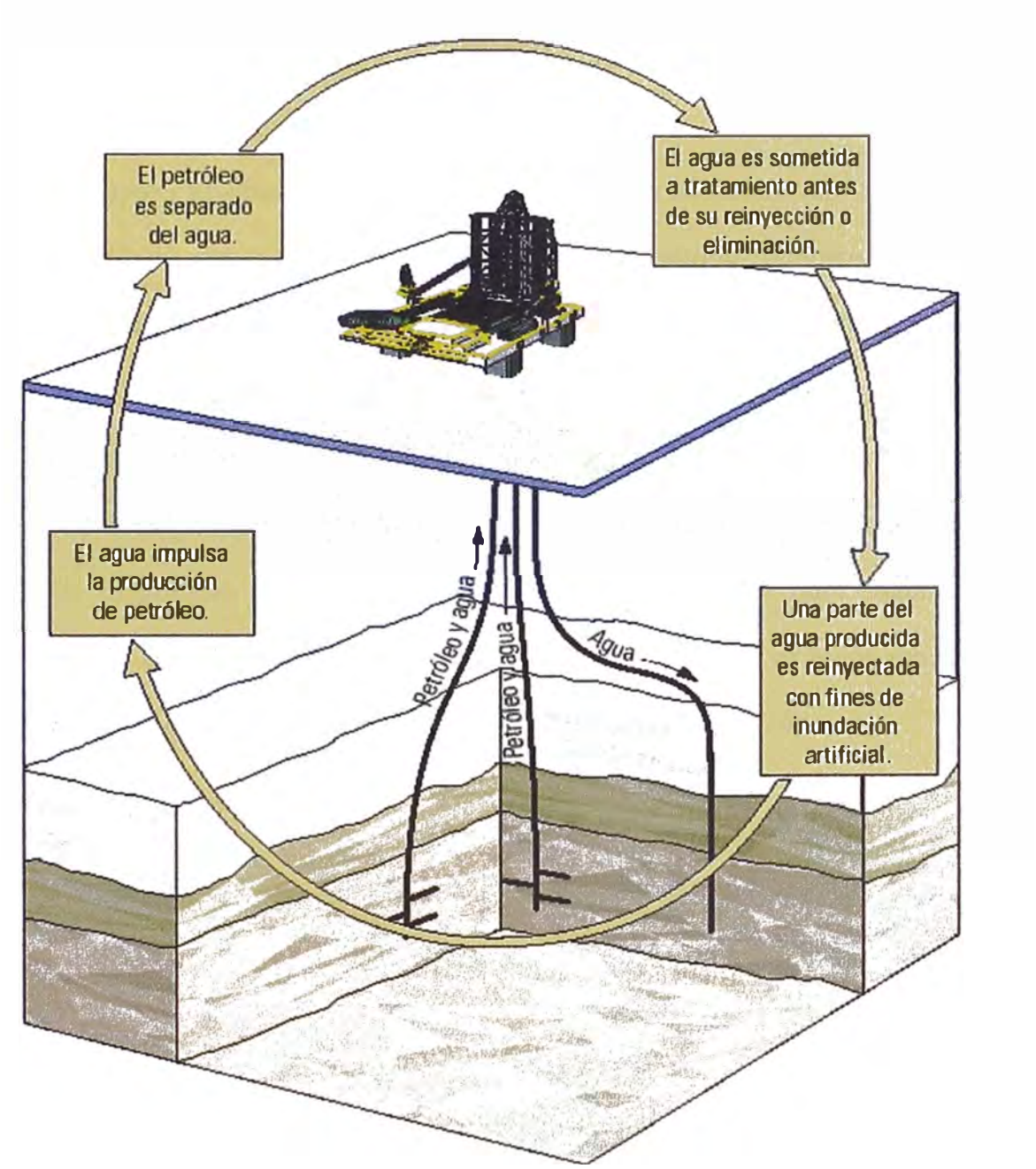


Figura 27. El rol del agua en la producción del petróleo

El Perú tiene considerables reservas petroleras y dichas fuentes están siendo explotadas actualmente tanto en las regiones costeras y en el mar de la parte noroccidental del país como en la selva húmeda amazónica del oriente peruano. En estas regiones también se refina el petróleo y las aguas residuales son descargadas en el ambiente.

Las aguas residuales de la extracción y de la refinación, con frecuencia, son tratadas mediante procesos físicos. También se usan métodos químicos o biológicos, particularmente en el caso de refinerías. El tratamiento físico involucra técnicas de separación por gravedad para remover petróleo y sólidos suspendidos. El separador API (American Petroleum Institute) es el dispositivo más común para la separación por gravedad (Figura 28). Este tipo de separador se utiliza en muchas refinerías del Perú. El dispositivo consiste de un estanque diseñado para maximizar la sedimentación de sólidos y la flotación de petróleo. El petróleo es luego recuperado y bombeado normalmente hacia dispositivos de almacenamiento *[10]*.

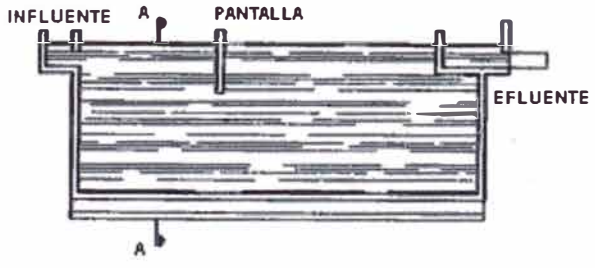


Figura 28. Poza API

3. PARÁMETROS DE MONITOREO

Los parámetros de calidad del agua que deben ser monitoreados en las descargas y aguas receptoras de las instalaciones petroleras tienen que guardar relación con los contaminantes potenciales que pueden estar presentes en las aguas residuales. Los parámetros que se muestran en la siguiente sección deben estar incluidos, como mínimo, en cualquier programa de monitoreo de calidad de aguas producidas por la extracción de petróleo y/o para las aguas residuales de las refinerías de petróleo. Se presenta una descripción breve de las razones para incluir cada parámetro *[10]*.

3.1. TEMPERATURA

El agua extraída de los pozos productivos del Perú tiene temperaturas elevadas en algunos casos (por ejemplo la selva amazónica) y, por lo general, retornan al medio ambiente antes de enfriarse hasta temperatura ambiente. Las descargas de agua a altas temperaturas pueden causar daños a la flora y fauna de las aguas receptoras al interferir con la reproducción de las especies, incrementar el crecimiento de bacterias y otros organismos, acelerar las reacciones químicas, reducir los niveles de oxígeno y acelerar la eutrofización *[12]*.

3.2. pH

El pH es una medida de la concentración de iones de hidrógeno en el agua. Aguas fuera del rango normal de 6 a 9 pueden ser dañinas para la vida acuática (por debajo de 7 son ácidas y por encima de 7 son alcalinas). Estos niveles de pH pueden causar perturbaciones celulares y la eventual destrucción de la flora y fauna acuática. Las aguas residuales de la industria petrolera, particularmente aquellas de las operaciones de refinación, pueden ser muy ácidas o alcalinas por el uso de productos químicos en varios procesos de refinación *[12]*.

3.3. CONDUCTIVIDAD

La conductividad de una muestra de agua es una medida de la capacidad que tiene la solución para transmitir corriente eléctrica. Esta capacidad depende de la presencia, movilidad, valencia y concentración de iones, así como de la temperatura del agua. En el caso de salmueras de campos petroleros y efluentes de refinería, es simplemente un indicador de la salinidad del agua *[12]*.

3.4. SÓLIDOS TOTALES DISUELTOS

Los Sólidos Totales Disueltos (STD) constituyen una medida de la parte de sólidos en una muestra de agua que pasa a través de un poro nominal de 2,0 µm (o menos) en condiciones específicas. Esta medida proporciona otra indicación (como la conductividad) de la salinidad en las descargas de la industria petrolera *[12]*.

3.5. CLORUROS

Los cloruros (Cl^-) son los principales aniones inorgánicos en el agua. A diferencia de los indicadores más generales de la salinidad (la conductividad y los STD), la concentración de cloruros es una medida específica de la salinidad de las descargas de la industria petrolera. Los cloruros son los principales componentes de las salmueras de petróleo. El incremento de cloruro en el agua ocasiona el aumento de la corrosividad del agua. El alto contenido de cloruros impide que el agua sea utilizada para el consumo humano o el ganado. Altos porcentajes de cloruros en los cuerpos de agua también pueden matar a la vegetación circundante *[12]*.

3.6. ACEITES Y GRASAS

Los aceites y grasas se definen en los "Métodos Estándar" como "cualquier material recuperado en la forma de una sustancia soluble en el solvente". El triclorofluoroetano es el solvente recomendado; sin embargo, debido a los problemas ambientales con los clorofluorocarbonos, se incluyen también solventes alternativos. La recolección de muestras y la medición deben realizarse con extremo cuidado. El aceite o petróleo en las salmueras es perjudicial para la vida acuática porque forma películas sobre la superficie del agua, reduce la aeración y disminuye la penetración de la luz solar necesaria para la fotosíntesis (producción primaria) de las plantas acuáticas. El aceite o petróleo en el agua de mar también puede formar "bolitas de alquitrán" en las playas y riberas de los ríos que pueden afectar plantas y animales.

Otro problema que puede causar el petróleo es la eclosión de los huevos de tortugas en los ríos de la selva amazónica. También se ha observado problemas en el desarrollo de cangrejos carreteros, muy-muy y otros organismos que habitan en playas arenosas de la costa *[12]*.

3.7. METALES: BARIO, CADMIO, CROMO, PLOMO, MERCURIO

Estos metales (Ba, Cd, Cr, Pb y Hg) frecuentemente son contaminantes del petróleo crudo y algunas veces están presentes en pequeñas cantidades en las aguas residuales de la industria petrolera. El Bario tiene efectos irreversibles para la salud y es tóxico para los animales. Se puede combinar con sulfatos para formar sulfato de bario insoluble. El Cadmio se acumula en tejidos blandos y puede interferir en el metabolismo. Es conocido que en sistemas acuáticos, el cadmio se acumula fácilmente en las ostras. El Cromo es cancerígeno para el sistema respiratorio y venenoso para los peces. El plomo se acumula en ostras y mariscos. Llega al ser humano a través de la cadena alimenticia y se acumula en los huesos. El plomo es un inhibidor de las enzimas e influye en el metabolismo celular. El mercurio es altamente tóxico a niveles relativamente bajos y se acumula en los peces. Produce "clorosis" en las plantas, es venenoso para los animales y llega al ser humano a través de la cadena alimenticia.

3.8. OTROS ANÁLISIS RECOMENDADOS

Además, de los análisis mencionados también deberían incluirse normalmente datos correspondientes a Demanda Bioquímica de Oxígeno, Demanda Química de Oxígeno, Coliformes Totales, Fenoles, Amoníaco, y los Caudales de las Descargas y los Caudales de las Aguas Receptoras *[12]*.

3.8.1. Demanda Bioquímica de Oxígeno (DBO)

La demanda bioquímica de oxígeno (DBO) es la cantidad de oxígeno usado por las bacterias bajo condiciones aeróbicas en la oxidación de materia orgánica para obtener CO_2 y H_2O. Esta prueba proporciona una medida de la contaminación orgánica del agua, especialmente de la materia orgánica biodegradable.

3.8.2. Coliformes Totales

Los coliformes son bacterias principalmente asociadas con los desechos humanos y animales. Los coliformes totales proporcionan una medida de la contaminación del agua proveniente de la contaminación fecal.

3.8.3. Demanda Química de Oxígeno

La Demanda Química de Oxígeno (DQO) es una medida del equivalente en oxígeno del contenido de materia orgánica en una muestra que es oxidable utilizando un oxidante fuerte. Es diferente a la prueba de la Demanda Bioquímica de Oxígeno

(DBO), pues la DBO mide sólo la fracción orgánica oxidable biológicamente. Es importante obtener una medida de la DQO en aguas residuales de refinería pues estos residuos, con frecuencia, contienen contaminantes orgánicos no biodegradables.

3.8.4. Oxígeno Disuelto

Este parámetro proporciona una medida de la cantidad de oxígeno disuelto en el agua. Mantener una concentración adecuada de oxígeno disuelto en el agua es importante para la supervivencia de los peces y otros organismos de vida acuática. La temperatura, el material orgánico disuelto, los oxidantes inorgánicos, etc. afectan sus niveles. La baja concentración de oxígeno disuelto puede ser un indicador de que el agua tiene una alta carga orgánica provocada por aguas residuales.

3.8.5. Fenoles

Esta medición suministra una indicación de la concentración de la mayoría de compuestos fenólicos (hidróxidos derivados de bencenos y sus núcleos condensados). Los fenoles frecuentemente están presentes en altas concentraciones en las aguas residuales de la industria petrolera. En niveles altos pueden manchar la piel de peces y afectar negativamente la flora, fauna y seres humanos. En niveles relativamente bajos estimulan la producción de cloruros.

3.8.6. Amoníaco

El amoníaco (NH_3) es un compuesto de nitrógeno que con frecuencia está presente en las aguas residuales de las refinerías. También se encuentran niveles altos de amoníaco en aguas servidas. Las concentraciones altas de amoníaco en aguas superficiales son tóxicas para los peces y pueden ser oxidadas y consumir el oxígeno disuelto del agua (nitrificación).

4. AGUAS DE PRODUCCIÓN DE LA EMPRESA UNIPETRO ABC S.A.C.

En este estudio los datos fueron recogidos dentro del campo petrolífero perteneciente a la Empresa Petrolera UNIPETRO ABC S.A.C (Figuras 29 y 30). Esta empresa fue creada el 17 de junio de 1993, día que se suscribió el contrato de servicio para explotar el Lote IX, distrito de Pariñas, provincia de Talara, departamento de Piura, empezando así el único caso en el mundo en que una universidad (Universidad Nacional de Ingeniería, Lima, Perú) esté al frente de las operaciones de un campo petrolífero *[13]*. Comprende un área aproximada de 1554,13 hectáreas, que es atravesada por la Quebrada Pariñas, y comprende los yacimientos Batanes, Algarroba, Cuesta y Leones. La altitud del área promedio es de 90 msnm. La ubicación de la empresa se ve en el mapa del Plano N° 1 y en el mapa del Anexo 2.

Figura 29. Empresa Petrolera UNIPETRO ABC S.A.C.

Figura 30. Empresa Petrolera UNIPETRO ABC S.A.C.

Las localizaciones del muestreo están marcadas en el mapa del Lote IX (Anexo 2) y se aprecian en las Figuras 31, 32, 33, 34, 35, 36 y 37.

Figura 31. Batería 401

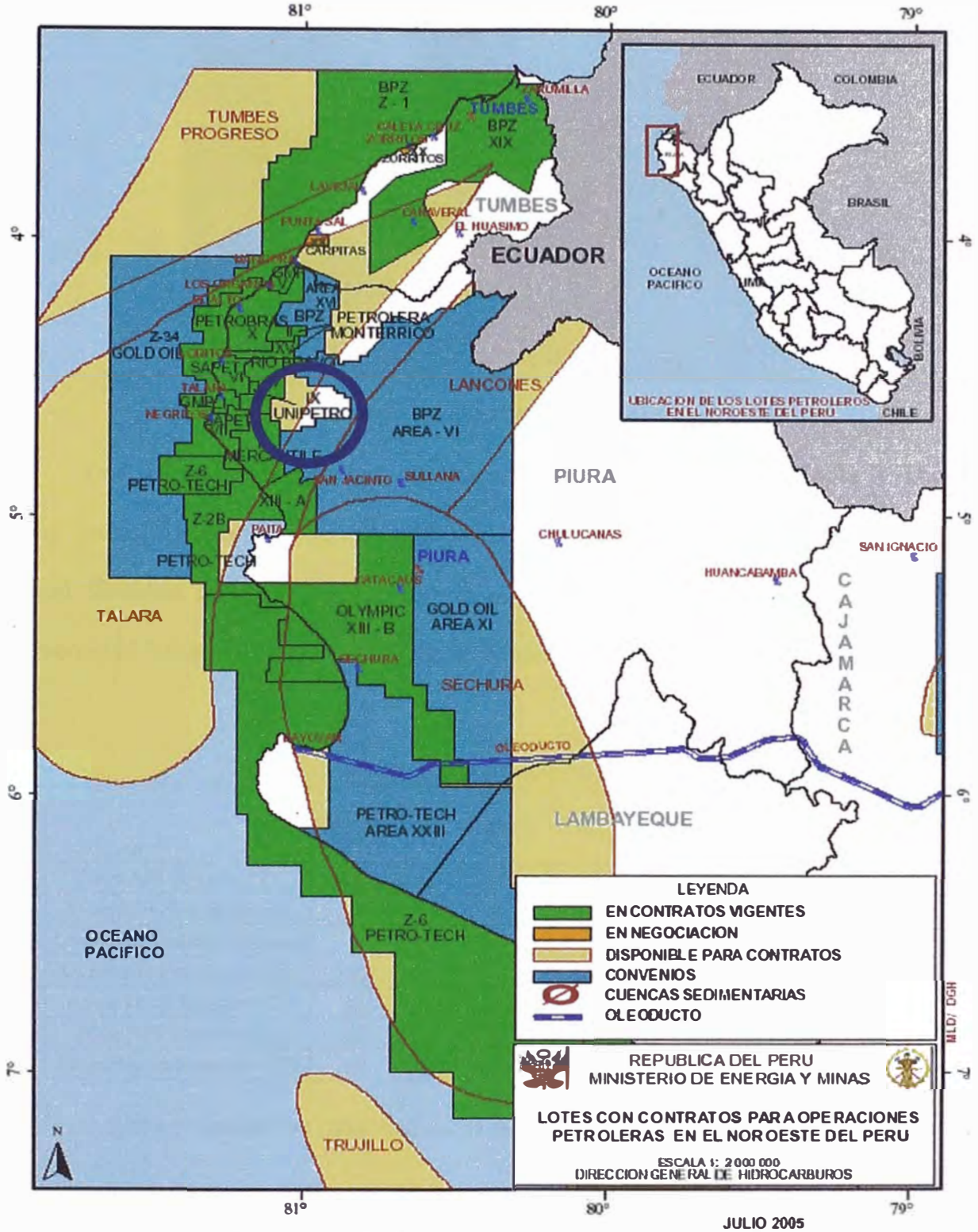
Figura 32. Batería 175

Figura 33. Poza API Batería 401

Figura 34. Poza API Batería 175

Figura 35. Quebrada Pariñas Aguas Abajo

Figura 36. Quebrada Pariñas Aguas Arriba



Plano 1. Ubicación del Lote IX, UNIPETRO ABC S.A.C. [14]

Figura 37. Manifold de Campo

Los diferentes parámetros de muestreo están listados en la Tabla 3. Dichos datos fueron recolectados desde julio de 1998 hasta marzo del 2005. Los análisis fueron llevados a cabo, en su mayoría por el laboratorio CERTIPETRO, de la Universidad Nacional de Ingeniería (Lima, Perú).

Tabla 3. Puntos y Parámetros de muestreo

	Tempe-ratura	pH	Cloruros	STD	Aceites y Grasas	Conduc-tividad	Metales (Ba, Cd, Cr, Hg, Pb)
Poza API Batería 175	Si	Si	Si	Si	Si	Si	Si
Manifold de Campo-2[Φ]	Si	Si	Si	Si	Si	Si	Si
Quebrada Pariñas Entrada	Si	Si	Si	Si	Si	Si	Si
Quebrada Pariñas Salida	Si	Si	Si	Si	Si	Si	Si
Agua Total Batería 175	Si	Si	Si	Si	Si	Si	Si
Poza API Batería 401	Si	Si	Si	Si	Si	Si	Si
Estación de Bombas 172	Si	Si	Si	Si	Si	Si	Si

Los datos obtenidos se muestran en el Anexo 3.

[Φ] Un Manifold de Campo es el lugar donde se reúnen todas las tomas de crudo de parte del lote petrolífero.

5. LÍMITES PERMISIBLES

La industria de refinación de petróleo crudo, sus derivados y petroquímica básica, genera desechos orgánicos e inorgánicos mezclados con aguas excedentes de los procesos de producción, así como aguas de servicio, las cuales, al ser descargadas en los cuerpos de agua modifican las características fisicoquímicas y biológicas naturales de estos cuerpos, disminuyendo en consecuencia su capacidad de autodepuración *[12]*.

El tipo y la cantidad de contaminantes que caracterizan a las aguas residuales de la industria de refinación de petróleo crudo, sus derivados y petroquímica básica, sus descargas a los cuerpos de agua, además de impedir o limitar su uso, produce efectos adversos en los ecosistemas, por lo que es necesario fijar los límites máximos permisibles de contaminantes en estas descargas. Es posible no rebasar los límites máximos permisibles fijados para la industria de refinación de petróleo crudo, sus derivados y petroquímica básica, con diferentes sistemas de tratamiento, que den resultados similares a los que se obtienen con la aplicación de los siguientes procesos: Igualación, separación de grasas y aceites, precipitación química del cromo, oxidación de sulfuros, sedimentación y tratamiento biológico *[12]*.

En mérito a lo anterior se han establecido los siguientes límites permisibles para el agua de producción de las empresas petroleras peruanas (Tabla 4) *[11]*.

Tabla 4. Límites Permisibles en aguas de producción

Parámetro	Límite Permisible	Unidades	Método de Análisis	Referencia Legal
Temperatura	(a)	°F	APHA 2250-B	DS.No 046-93-EM
pH	5,5 - 9		APHA 4500-H+	DS.No 046-93-EM
Cloruros	(b)	mg/L	APHA 4500-Cl	DS.No 046-93-EM
STD	No hay normativa	mg/L	APHA 2540-C	DS.No 046-93-EM
Aceites y Grasas	50	mg/L	EPA 1664	DS.No 046-93-EM
Conductividad a 25°C	No hay normativa	mS/cm	APHA 2510-B	DS.No 046-93-EM
Bario	5	mg/L	APHA 3111-D	DS.No 046-93-EM
Cadmio	0,5	mg/L	APHA 3111-B	DS.No 046-93-EM
Cromo	1	mg/L	APHA 3111-B	DS.No 046-93-EM
Mercurio	0,01	mg/L	APHA 3500Hg-B	DS.No 046-93-EM
Plomo	0,1	mg/L	APHA 3111-B	DS.No 046-93-EM

(a) Las aguas vertidas no deben de alterar en más de 2°F los valores naturales del cuerpo receptor.
(b) Las aguas vertidas no deben de alterar en más de 200 mg/L los valores naturales del cuerpo receptor.

Los métodos de ensayo que se aplicarán para la determinación de los valores correspondientes a los contaminantes en las descargas de aguas residuales de la industria de refinación de petróleo crudo, sus derivados y petroquímica básica, son los contenidos en las normas oficiales peruanas siguientes *[11]*:

EPA 1664: Aguas -Determinación de grasas y aceites- Método de extracción Soxhlet.

APHA 2250-B: Aguas -Determinación de temperatura- Método visual con termómetro.

APHA 4500-H+ : Aguas -Determinación de pH- Método potenciométrico.

APHA 2540-C: Aguas -Determinación de sólidos disueltos totales- Método gravimétrico.

APHA 3111-B: Análisis de agua -Determinación de metales- Método de absorción atómica.

APHA 3500Hg-B: Análisis de agua -Determinación del mercurio- Método colorimétrico.

APHA 4500-Cl: Análisis de agua -Determinación de cloruros- Método argentométrico.

CAPITULO III
ANÁLISIS QUIMIOMÉTRICO DE LOS DATOS DE PARÁMETROS FÍSICO-QUÍMICOS DE LAS AGUAS DE PRODUCCIÓN DE UNIPETRO ABC S.A.C.

1. EXAMEN PRELIMINAR DE LOS DATOS (PRUEBAS ESTADÍSTICAS)

1.1. TRATAMIENTO DE DATOS

La conversión de ficheros, correspondientes a los datos del anexo3, el análisis exploratorio inicial, el preprocesado de datos y los cálculos se realizaron mediante los programas estadísticos de los propios equipos (en el caso de análisis instrumental) y mediante el uso de paquetes estadísticos comerciales:

Software Espectrofotómetro de Absorción Atómica Shimadzu

The Unscrambler (CAMO, Noruega) versión 9.2 (Figura 38)

SPSS (Lead technologies, USA) VERSIÓN 11.0.1 y 13.0.

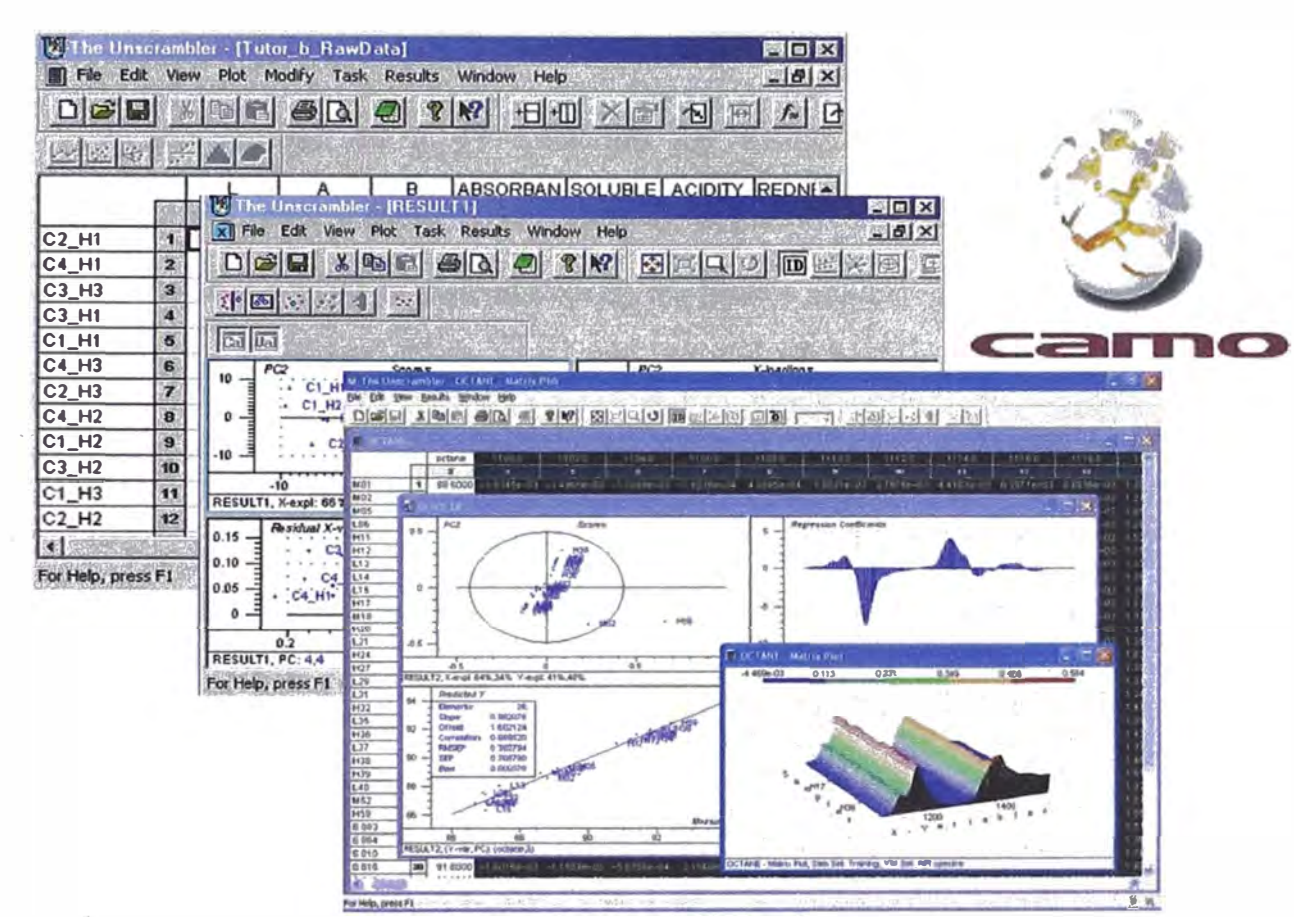


Figura 38.- Software The Unscrambler de Camo

1.2. CORRELACIONES

Usando el programa SPSS se han determinado las correlaciones bivariadas, tomadas de dos en dos, en el espacio definido por las variables. De la gráfica de correlaciones (Figura 39) podemos deducir que hay fuertes correlaciones entre las variables de cloruros, STD y conductividad. Entre las otras siete variables no se observa una correlación apreciable.

1.3. HISTOGRAMAS

Las Figuras 40 a 47 muestran los histogramas correspondientes a las variables: pH, temperatura, STD, cloruros, conductividad, aceites y grasas, bario y cromo. Se observa que la distribución de los puntos son cercanas a la aleatoriedad, es decir, a una distribución normal, en el caso del pH. En cambio en el caso de las otras variables, éstas se alejan mucho de la normalidad. En el caso de la temperatura, éstas se distribuyen alrededor de la media. Para los histogramas de cloruros, STD y conductividad, observamos una distribución similar; en el caso particular de los cloruros se nota una distribución por encima de 8000 mg/L.

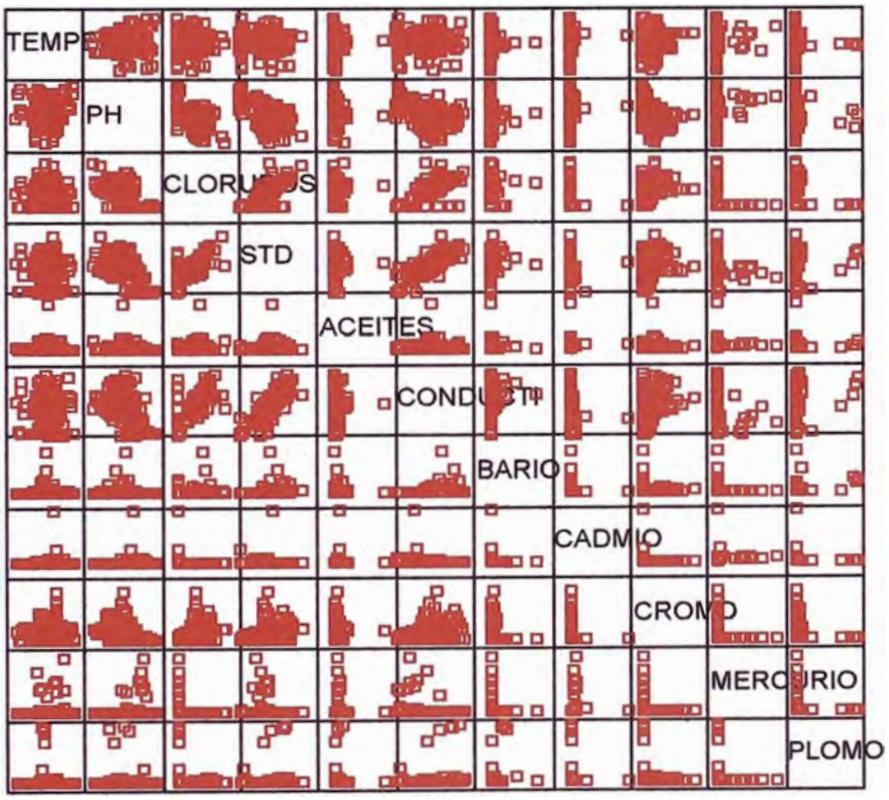


Figura 39. Gráfica de las correlaciones bivariadas

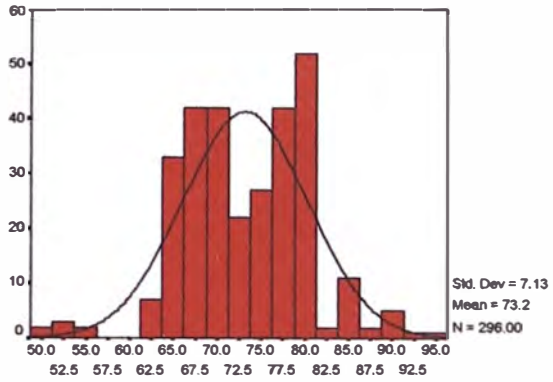


Figura 40. Histograma para la temperatura

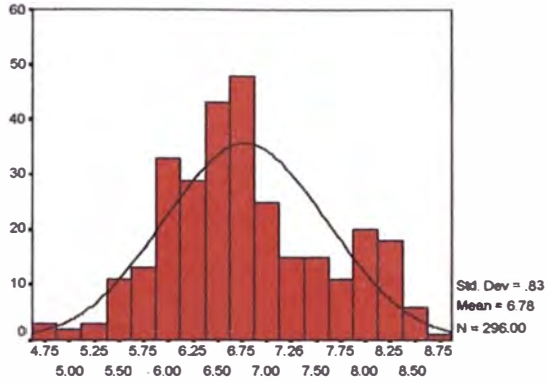


Figura 41. Histograma para el pH

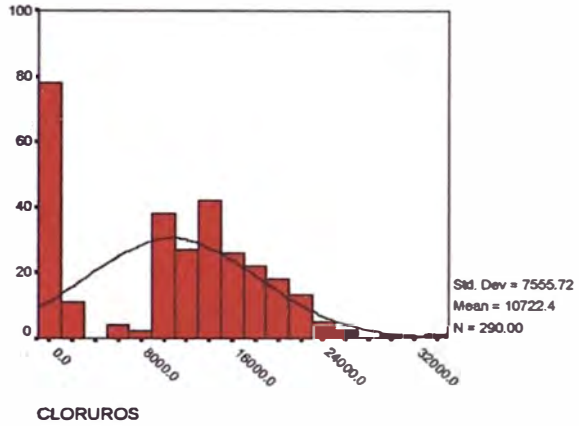


Figura 42. Histograma para los cloruros

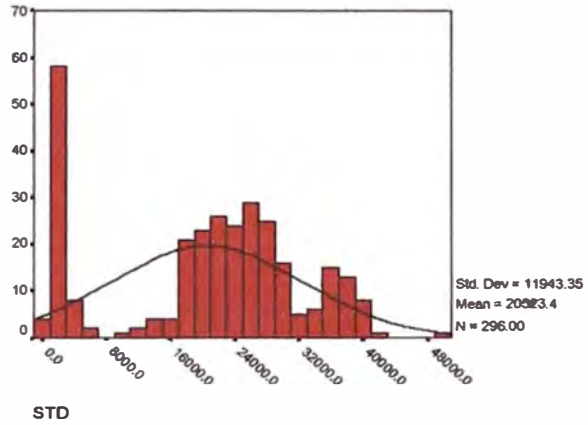


Figura 43. Histograma para STD

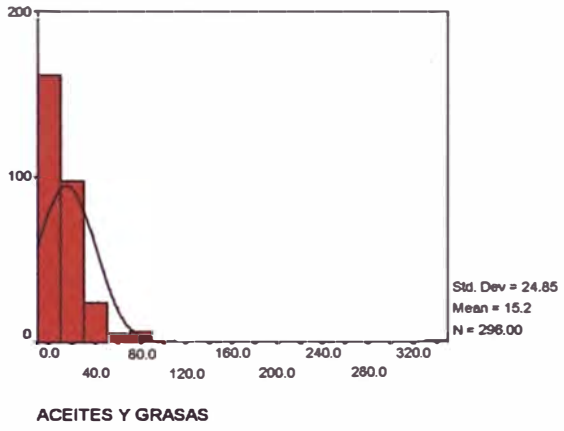


Figura 44. Histograma para aceites y grasas

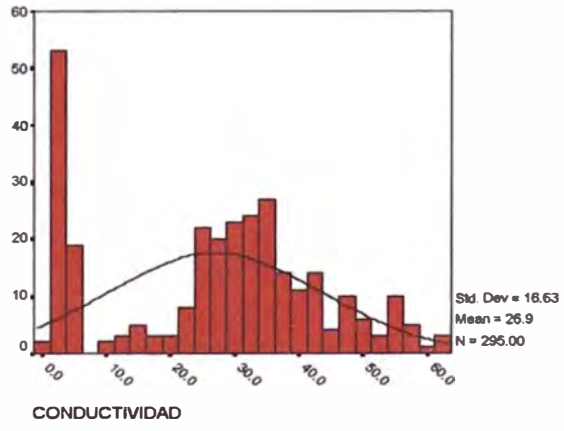


Figura 45. Histograma para la conductividad

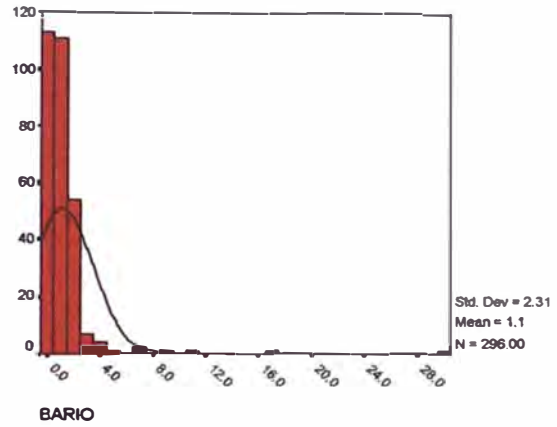


Figura 46. Histograma para el bario

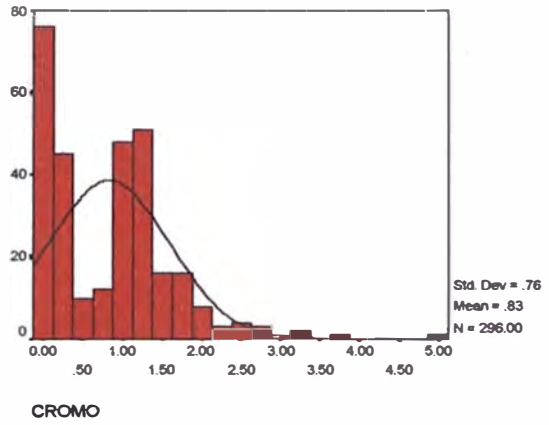


Figura 47. Histograma para el cromo

2. TRATAMIENTO DE DATOS

2.1. RELLENO DE HUECOS

Para aquellos puntos correspondientes a muestras en las cuales no se ha evaluado alguna de las variables, el programa The Unscrambler v.9.2 rellena la matriz de datos con el valor medio, tratando de minimizar distorsiones en las medias y varianzas de los valores correspondientes a dichas variables.

2.2. ESTANDARIZACIÓN DE LOS DATOS

Antes de usar el programa The Unscrambler v.9.2 se deben normalizar los datos, primero centrándolos y luego comparándolos contra su desviación estándar s, de modo que cada variable x se transforma en la variable estandarizada z de acuerdo a la ecuación:

$$z_i = \frac{x_i - \bar{x}}{s}$$

3. ESTUDIO MEDIANTE EL ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

La utilización de las herramientas para Análisis de Componentes Principales (PCA) del programa The Unscrambler v.9.2 nos permite construir una serie de gráficas cuya mejor interpretación pasamos a detallar.

3.1. ANÁLISIS DE COMPONENTES PRINCIPALES PARA TODAS LAS VARIABLES

3.1.1. Gráficos de Sedimentación

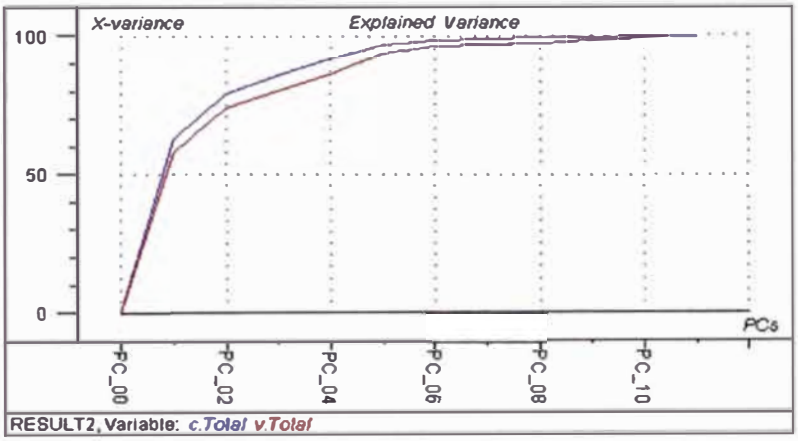


Figura 48. Gráfica de Sedimentación de todas las variables

La gráfica de sedimentación (Figura 48) nos muestra que son dos los componentes principales que explican en forma satisfactoria el global de la información; con ellos se mantiene el 79 % de la información de las variables manifiestas. El 21 % restante, se distribuye entre los demás componentes y es información redundante. Este gráfico también es útil para identificar a los autovalores frente a su número de orden. El primer autovalor es muy elevado con

respecto a los demás, lo que indica que este vector modela bien los datos, describiendo una buena parte de la varianza total. Los demás componentes muestran autovalores gradualmente más bajos.

3.1.2. Diagrama de Cargas (Loadings)

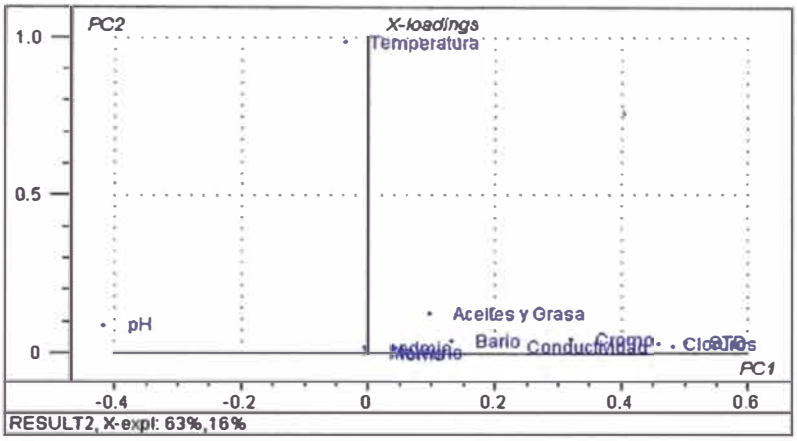


Figura 49. Gráfica de cargas de todas las variables

La gráfica de cargas (loadings, Figura 49) tiene al componente 1 con el 63 % de la información original y al componente 2 con el 16 %, en ambos hay el 79 % de la información total. Del componente 1, podemos decir que, clasifica al agua de acuerdo a su salinidad (cloruros, STD y conductividad) y pH. De hecho las aguas de producción de petróleos se caracterizan por su salinidad y bajo pH con respecto a las aguas dulces de los ríos. El segundo componente clasifica a las muestras de acuerdo a su temperatura. Esta gráfica también muestra que las variables fisicoquímicas que más influyen en la separación de estos grupos de los distintos tipos de agua son cloruros, sólidos totales disueltos (STD) y conductividad (estas tres están muy correlacionadas y prácticamente proporcionan la misma información de las sales

disueltas), el pH y la temperatura. Así, podemos decir que si deseamos hacer una clasificación del tipo de agua, utilizaremos como variables fisicoquímicas adecuadas para la clasificación, las variables STD (uno de los parámetros de las sales disueltas) el pH y la temperatura. El número de variables se podría reducir.

3.1.3. Diagrama de Puntuaciones (Scores)

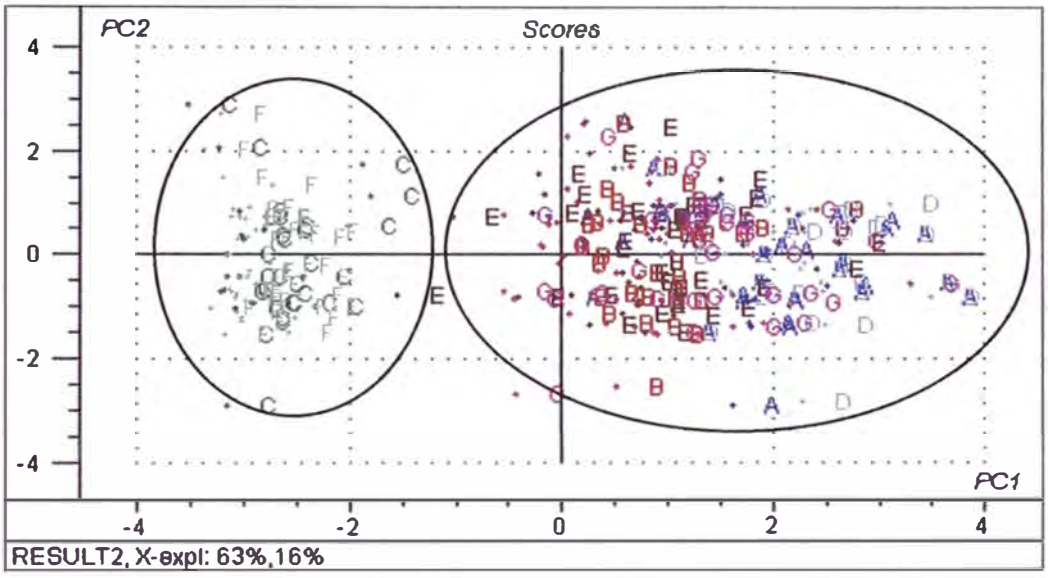


Figura 50. Gráfico de puntuaciones sobre el plano de los componentes 1 y 2 de todos los objetos

El gráfico de puntuaciones sobre el plano de los componentes 1 y 2 (Figura 50) muestra que se pueden distinguir dos grupos claramente definidos: el agua del río Pariñas Aguas Arriba, C, con el agua del río Pariñas Aguas Abajo, F, y las aguas de producción de los pozos petroleros (A, B, D, E, G).

3.1.4. Diagrama Doble (Biplot)

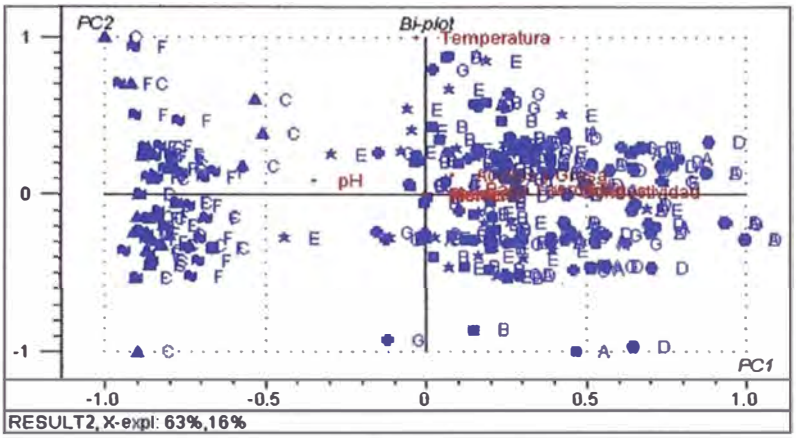


Figura 51. Diagrama doble para todas las variables

El diagrama biplot de la Figura 51 nos muestra como las variables pH, aquellas relacionadas a la salinidad (cloruros, STD y conductividad) y la temperatura, separan a los objetos en dos grupos, los correspondientes al río Pariñas (C y F) y los correspondientes a las aguas de producción (A, B, D, E, G).

3.1.5. Análisis con el Tercer Componente

El tercer componente (contribución del 7 %) clasifica las muestras de acuerdo a su contenido de cromo. Esta es una variable influyente para las aguas de producción que están en contacto con equipos que contaminan las aguas con cromo. (ver Figura 52).

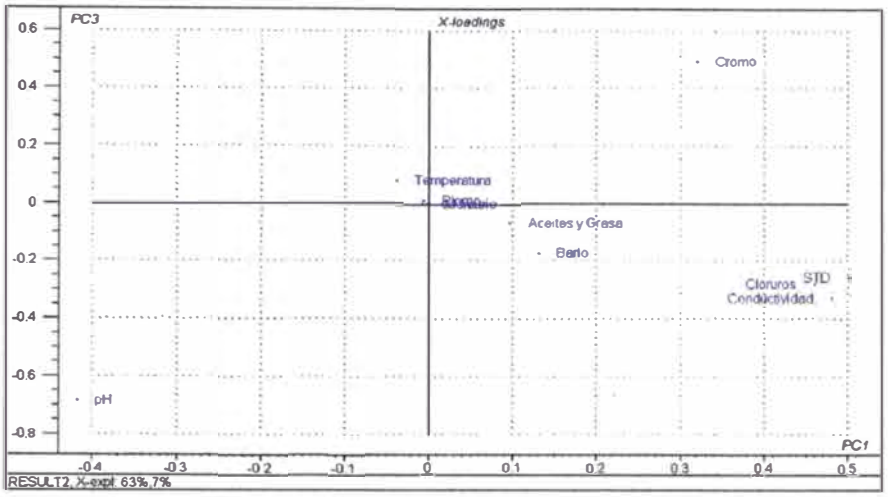


Figura 52. Gráfica de cargas de todas las variables en el plano de los componentes 1 y 3

En el mismo plano de los componentes 1 y 3 (Figura 53) observamos que las aguas del MC-2 (B) y de la Poza API 401 (E) se caracterizan por su contenido de cromo. El punto de muestreo MC-2 (B) tiene mayor contenido de cromo que la Poza API 401, puesto que en esa zona se encuentran un gran número de dispositivos propensos a contaminar las aguas con cromo.

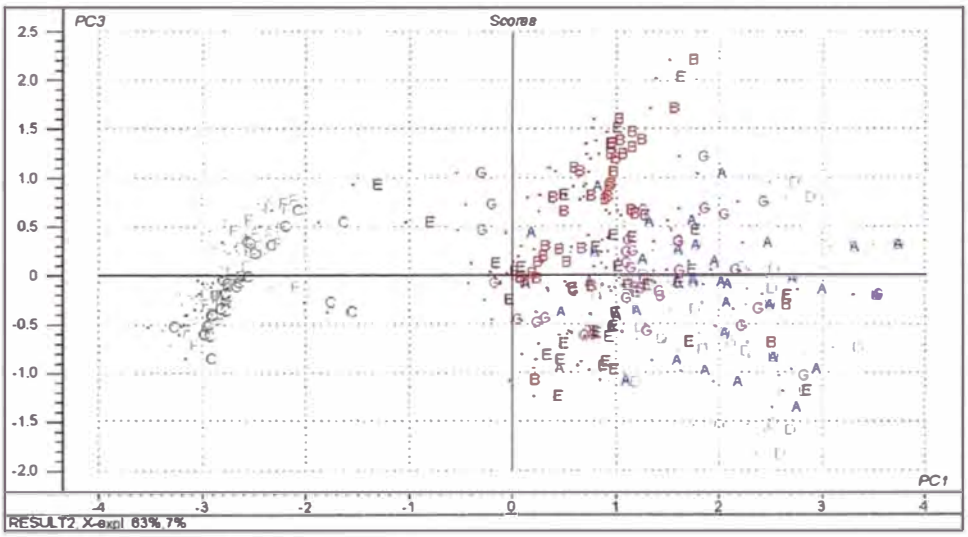


Figura 53. Gráfica de puntuaciones de todas las variables en el plano de los componentes 1 y 3

3.2. ANÁLISIS DE COMPONENTES PRINCIPALES EN LA POZA API 175 (PUNTO A)

Se analizará el comportamiento fisicoquímico de las aguas de esta poza debido a que podría ser fuente de contaminación.

3.2.1. Gráfica de Sedimentación

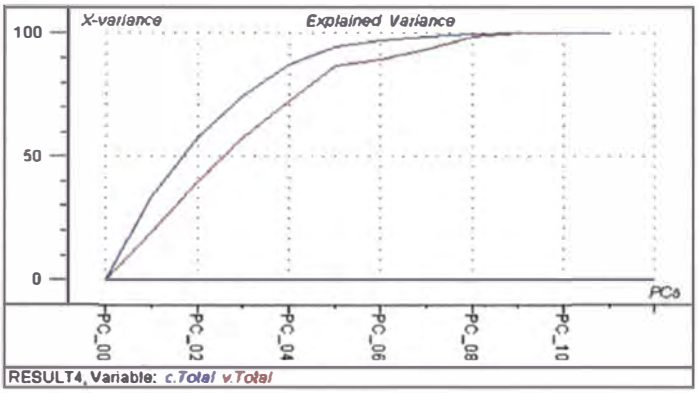


Figura 54. Gráfica de sedimentación para la Poza API 175

La gráfica de sedimentación (Figura 54) nos permite observar que son cuatro los componentes principales que mantienen la información importante (89%), pero analizaremos solo tres componentes por ser más adecuado.

3.2.2. Gráfica de Cargas (Loadings)

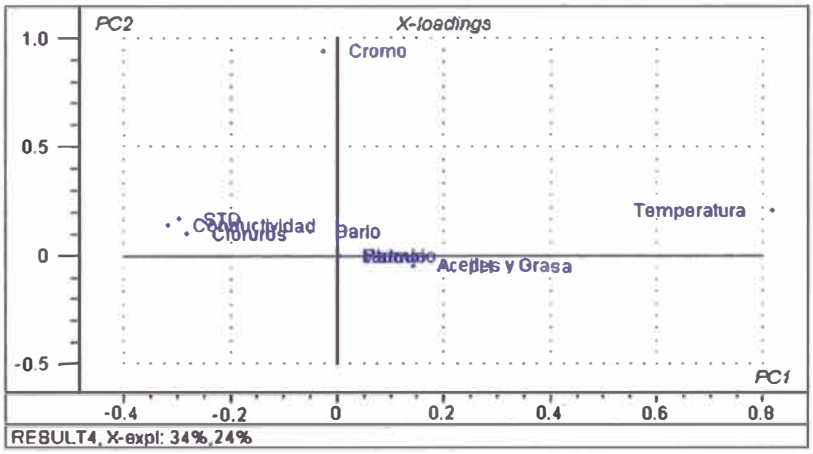


Figura 55. Gráfica de cargas para la Poza API 175

El gráfico de cargas (loadings) de la Figura 55, para la Poza API 175, tiene en el primer componente la temperatura como su variable principal, por lo tanto este componente clasifica los datos de acuerdo a la temperatura. El segundo componente está relacionado con el contenido de cromo.

3.2.3. Gráfica de Puntuaciones (Scores)

La gráfica de puntuaciones (scores) de la Figura 56 clasifica los datos de acuerdo a la temporada del año: los meses más calientes se encuentran en el lado derecho de la primera componente (de diciembre a junio) y los meses más fríos en el lado izquierdo (de julio hasta noviembre).

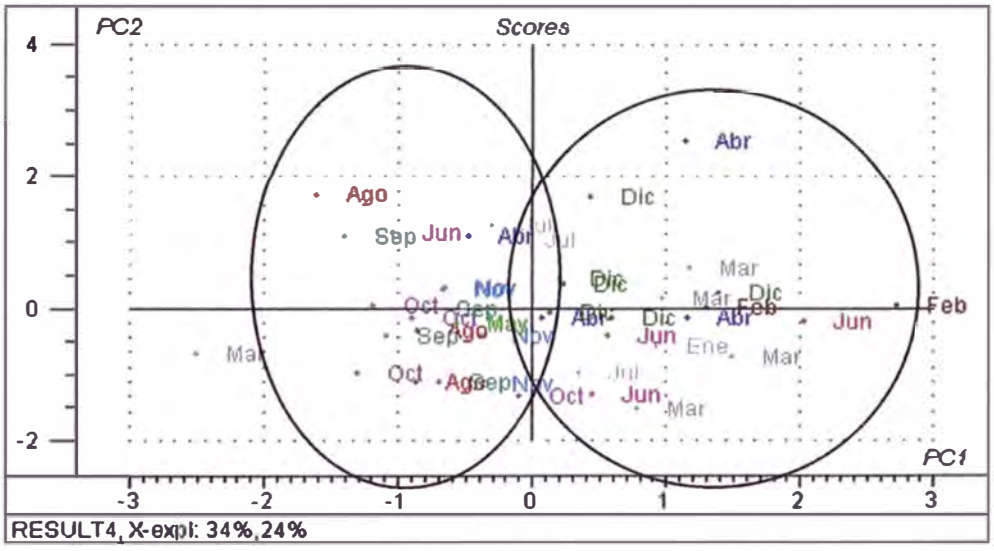


Figura 56. Gráfica de puntuaciones para la Poza API 175

3.2.4. Diagrama Doble (Biplot)

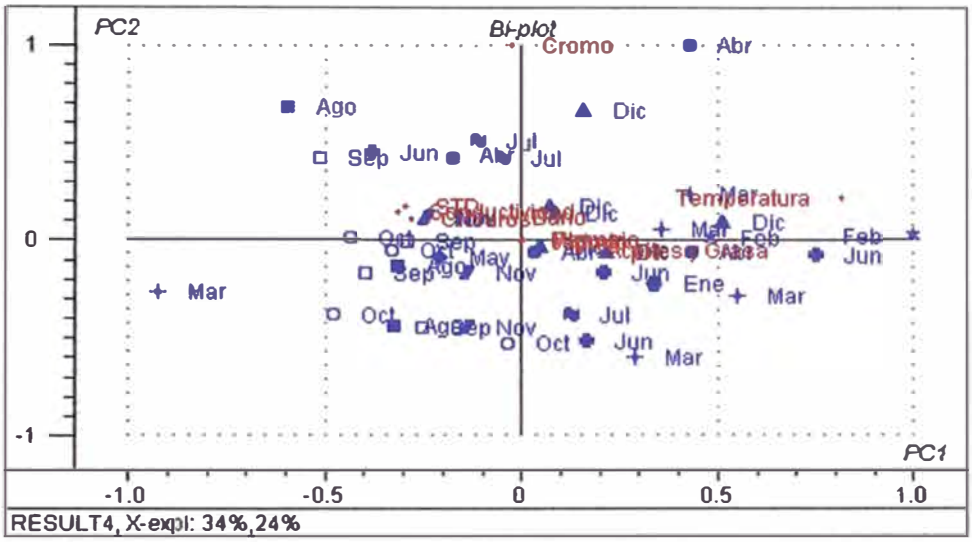


Figura 57. Diagrama doble para la Poza API 175

Según la Figura 57, en los meses de diciembre de 1999, abril del 2000 y agosto del 2004 hubo una mayor concentración de cromo en la Poza API 175.

3.2.5. Análisis con el Tercer Componente

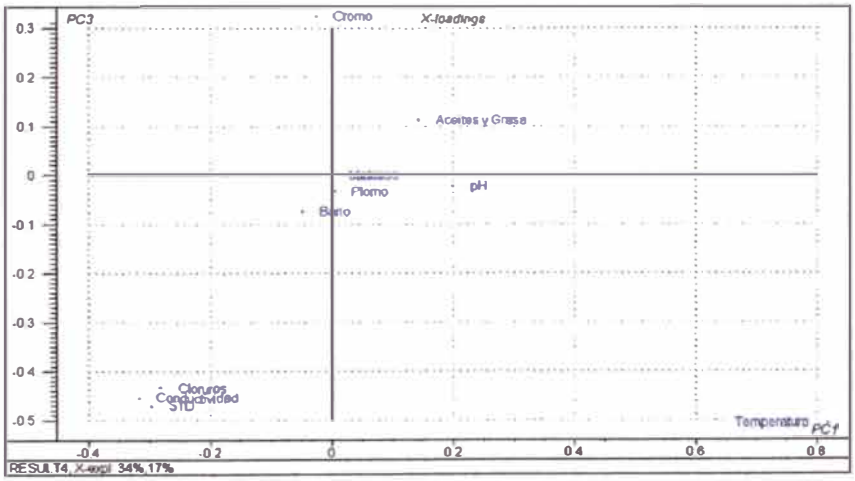


Figura 58. Gráfica de cargas con el tercer componente para la Poza API 175

Podemos observar que, de acuerdo a la Figura 58, el tercer componente no aporta una buena clasificación en particular de acuerdo a estas variables fisicoquímicas.

3.3. ANÁLISIS DE COMPONENTES PRINCIPALES EN EL MANIFOLD DE CAMPO (MC2, PUNTO B)

3.3.1. Gráfica de Sedimentación

La gráfica de la Figura 59 indica que en este punto de muestreo son importantes hasta cuatro componentes (83,6% de varianza acumulada). Solo las dos primeras componentes serán suficientes para indicar el comportamiento en este punto de muestreo.

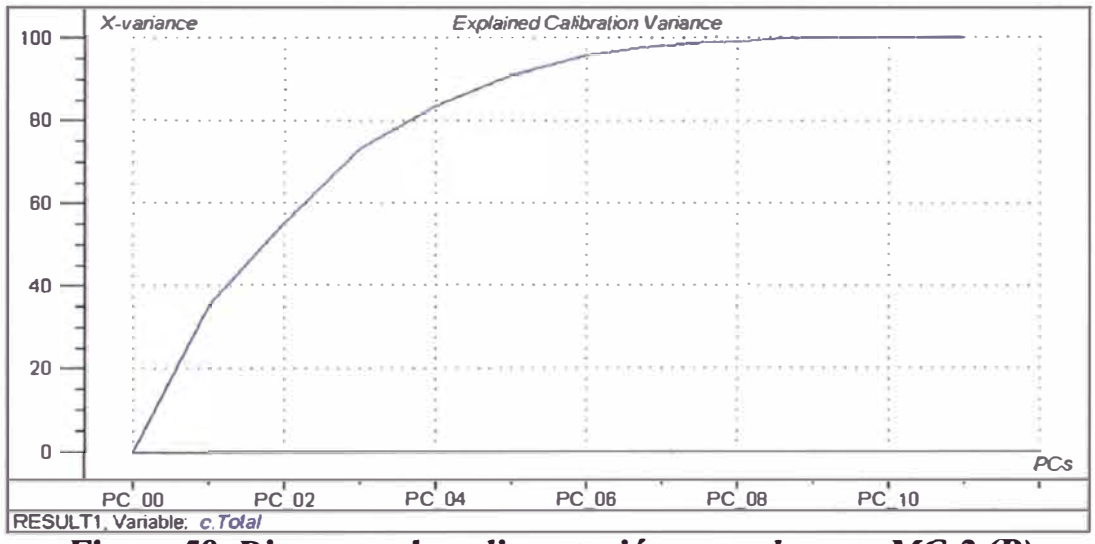


Figura 59. Diagrama de sedimentación para el punto MC-2 (B)

3.3.2. Gráfica de Cargas (Loadings)

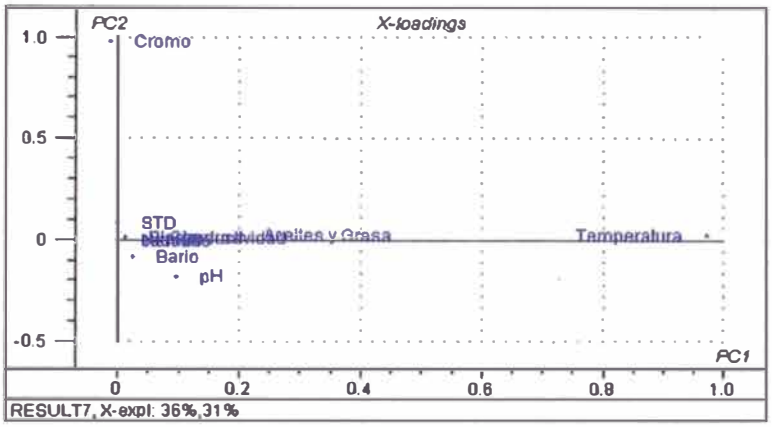


Figura 60. Diagrama de Cargas para el punto MC-2 (B)

En la Figura 60, se puede apreciar que las variables se pueden clasificar según el componente 1 por su temperatura y en el componente 2 por el contenido de cromo, tan igual como el caso de la Poza API 175 (punto A).

Aunque no se presentan las gráficas, todos los puntos del lote, observan un mismo comportamiento fisicoquímico.

3.4. ANÁLISIS DE COMPONENTES PRINCIPALES EN LA QUEBRADA PARIÑAS ENTRADA (C)

3.4.1. Gráfica de Sedimentación

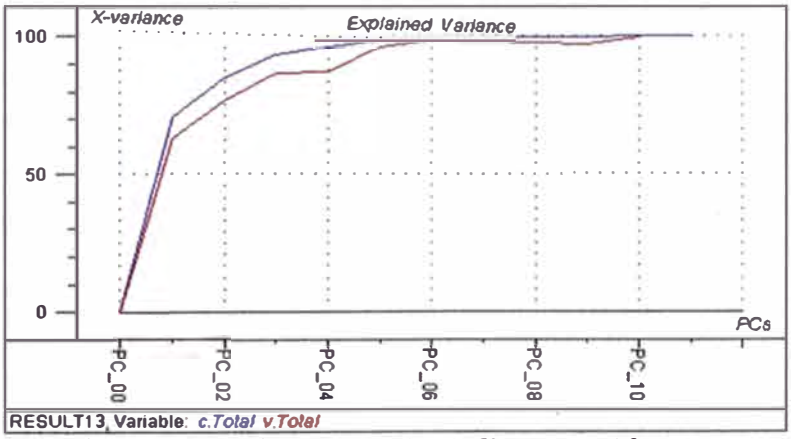


Figura 61. Diagrama de sedimentación para la Quebrada Pariñas Entrada (C)

Se observa de la gráfica de sedimentación (Figura 61) que son dos componentes (1 y 2) los que mantiene la mayor información (70% y 15%, respectivamente), lo que nos da una varianza explicada acumulada del 85%.

3.4.2. Gráfica de Puntuaciones (Scores)

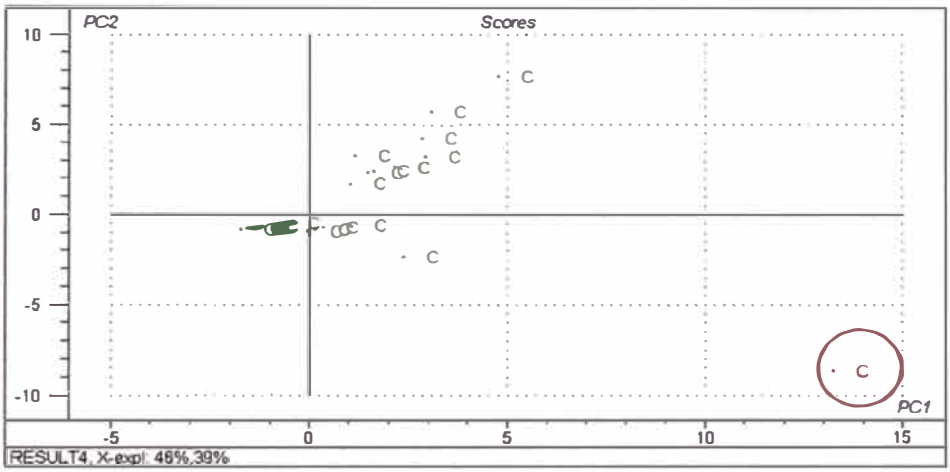


Figura 62. Gráfica de Puntuaciones con todos los datos del punto C

El gráfico de puntuaciones de la Figura 62, correspondientes a todos los datos de la Quebrada Pariñas Entrada, muestra que al menos uno de sus puntos es un valor atípico (outlier) el cual será eliminado para obtener un modelo más acorde a la realidad.

La identificación del valor atípico puede hacerse utilizando el software The Unscrambler, que nos indica que corresponda a la muestra 95 del punto de muestreo Quebrada Pariñas Entrada (Figura 63).

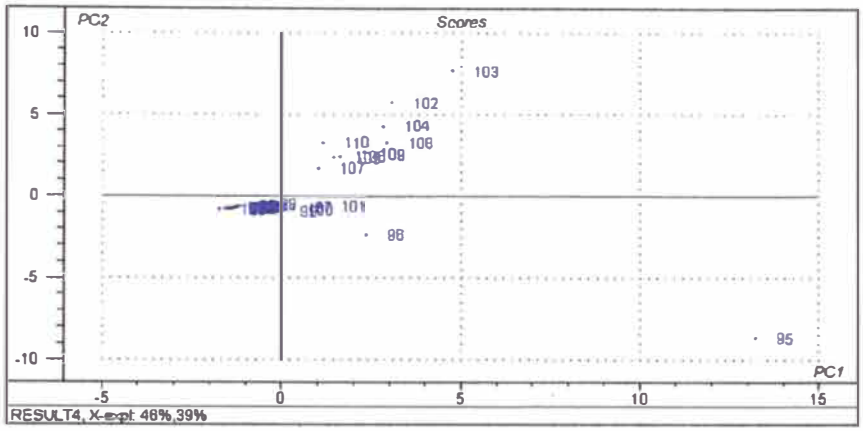


Figura 63. Diagrama de Puntuaciones para la Quebrada Pariñas Entrada mostrando el valor atípico (muestra 95)

Por simple inspección visual de los valores correspondientes a la muestra 95, deducimos que el valor atípico podría corresponder a la variable cadmio:

		Cloruros	STD	tes y Gra	onductivida	Bario	Cromo	Mercurio	cadmio	Plomo
94	8	-1.3030	-1.4865	-0.4905	-1.3814	-0.4166	-1.0130	-0.1704	-0.1132	-0.1863
95	9	-1.3291	0.4217	0.9114	-0.6538	-0.4674	-1.0445	-0.1609	16.3265	-0.2121
96	10	-1.3450	-1.6368	0.7418	-0.4734	-0.4723	-0.9265	-4.3632e-02	3.7080	-0.2121

Aplicaremos la prueba de Grubbs para la variable cadmio. Tenemos 43 datos (tomaremos $n = 40$), la media para el cadmio en este punto de muestreo es $\bar{x} = 0{,}644$ y su desviación estándar s = 2,550, y el punto dudoso corresponde a x = 16,3265 con lo que calculamos T_{40}:

$$T_{40} = \frac{|0{,}644 - 16{,}3265|}{2{,}550} = 6{,}15$$

valor que es mucho mayor que el correspondiente valor crítico T(1- $\alpha = 0{,}99$; n=40) = 3,24 a un nivel de significancia $\alpha = 0{,}01$, por lo que corroboramos que este es un valor atípico.

Eliminando este punto mediante el software se obtiene un nuevo diagrama de puntuaciones (Figura 64). De modo similar se pueden eliminar otros puntos que se consideren atípicos.

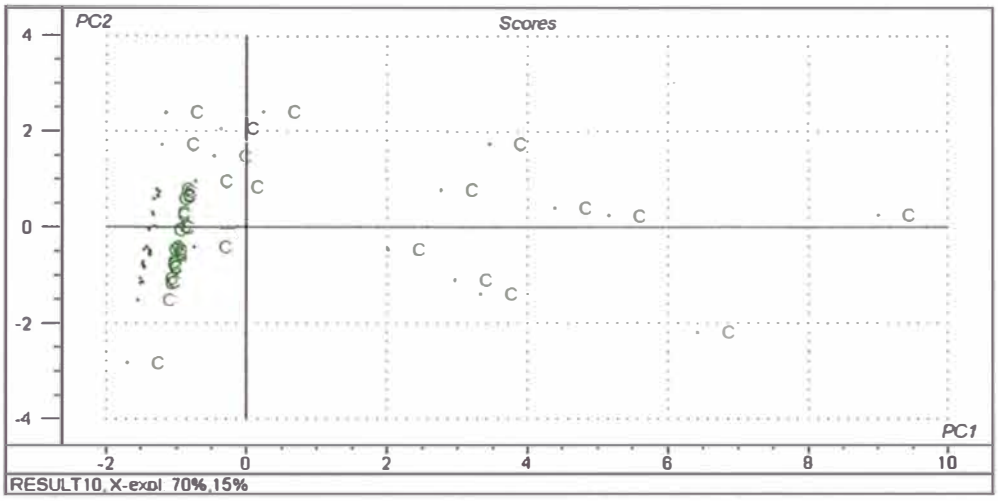


Figura 64. Diagrama de puntuaciones de Quebrada Pariñas Entrada sin considerar el outlier de la Figura 63

3.4.3. Gráfica de Cargas (Loadings)

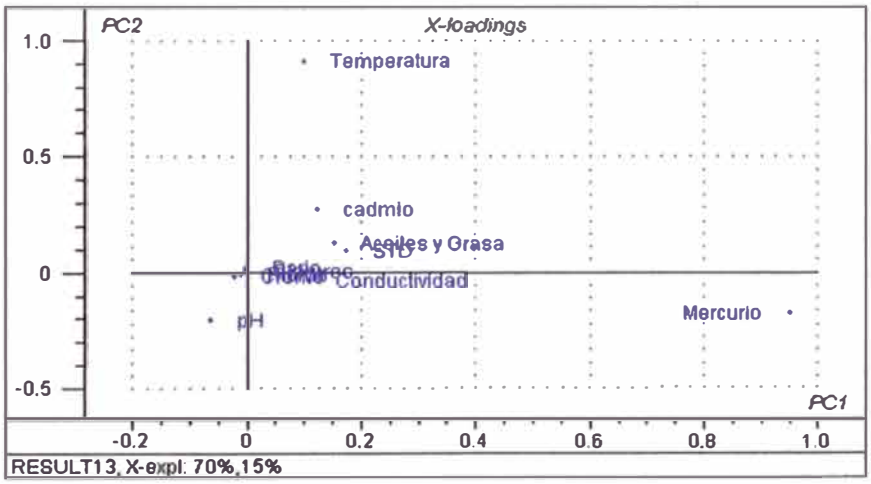


Figura 65. Diagrama de Cargas para el punto C

De la Figura 65, se observa que el primer componente se encuentra relacionado con el mercurio. Esto nos indica que algunos de los datos de muestreo en este punto está influenciado con alto contenido de mercurio.

Haciendo el mismo análisis de componentes principales y eliminando los valores altos de mercurio (valores atípicos de algunas muestras), obtenemos el diagrama de cargas de la Figura 66.

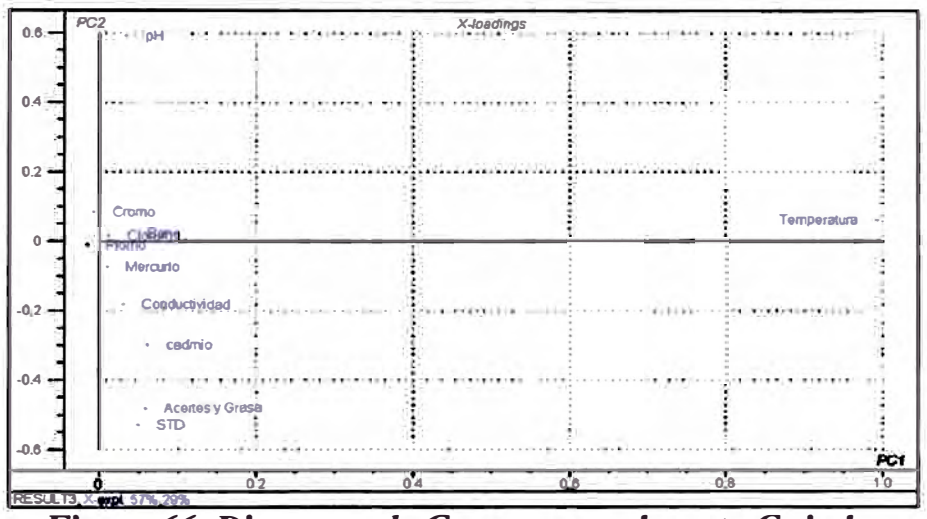


Figura 66. Diagrama de Cargas para el punto C sin los valores atípicos de mercurio

De la Figura 66, podemos observar que la temperatura es la variable que influye en esta primera componente como hemos obtenido en los demás casos. El segundo componente está relacionada con el pH y STD, que vienen a ser las variables relacionadas con las aguas de producción. Aquí si podemos decir que hay una influencia de las variables de las aguas de producción en las aguas río arriba de la Quebrada Pariñas, lo que quizá sea un indicio de contaminación pero mínima, ya que esta componente representa solo el 29% de la información.

3.5. ANÁLISIS DE COMPONENTES PRINCIPALES EN LA QUEBRADA PARIÑAS SALIDA (F)

3.5.1. Gráfica de Sedimentación

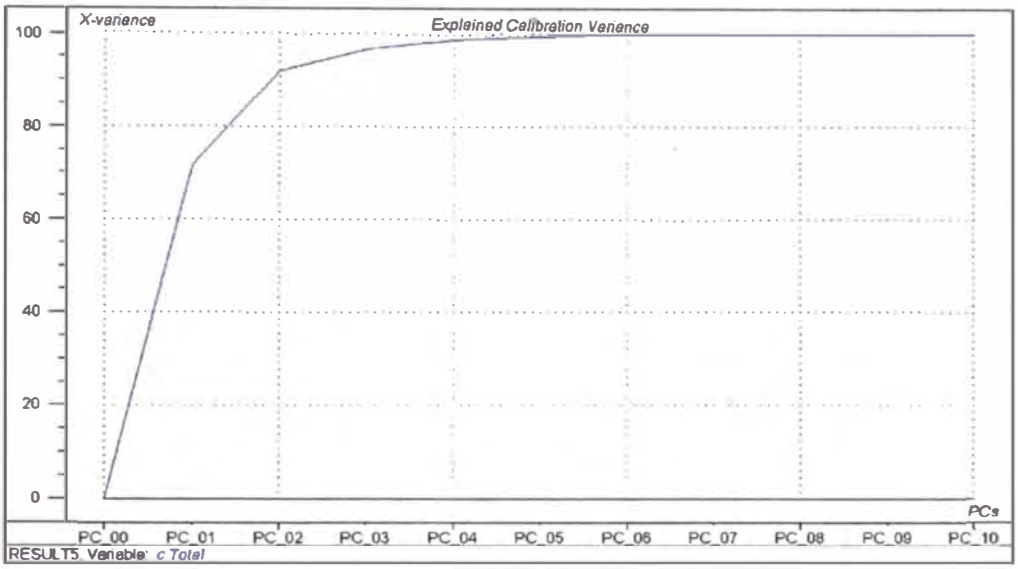


Figura 67. Diagrama de sedimentación para la Quebrada Pariñas Salida (punto F)

Se observa de la gráfica de sedimentación (Figura 67) que son dos componentes (1 y 2) los que mantiene la mayor información (72 y 20%, respectivamente), lo que nos da una varianza explicada acumulada del 92% y una varianza residual acumulada de solo 8%.

3.5.2. Gráfica de Cargas (Loadings)

De la Figura 68 observamos que en su segunda componente existe un aporte diferente de las variables que en el caso de las aguas del río Pariñas Entrada. La posición de ellas en el diagrama significa que no se detecta contaminación, debido quizá al caudal relativamente alto del río Pariñas (20 L/s *[11]*).

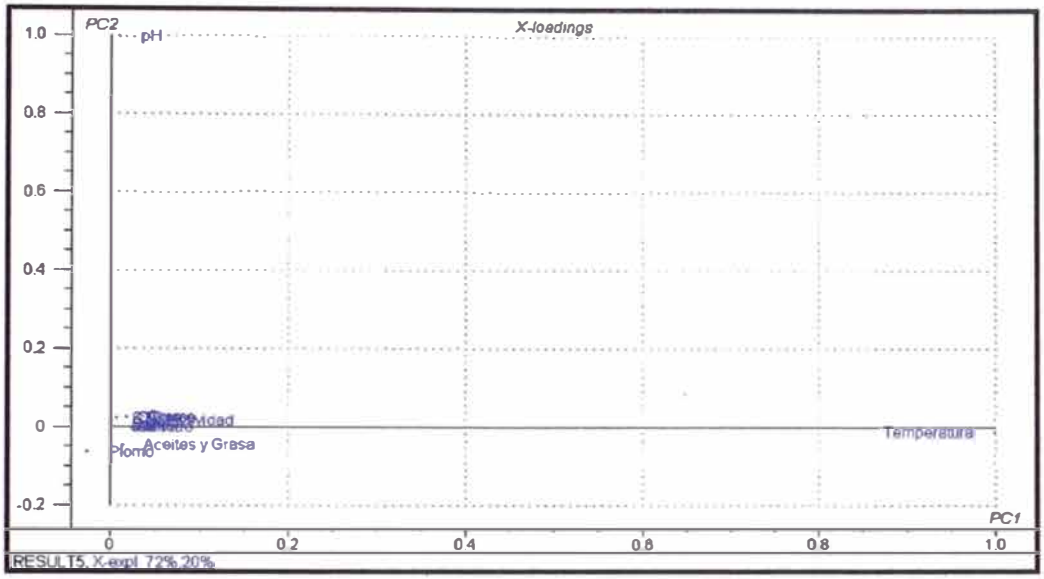


Figura 68. Diagrama de Cargas para el punto F

3.5.3. Gráfica de Puntuaciones (Scores)

Como en el caso de Pariñas Entrada se observa en la Figura 69, que la primera componente (cuya principal variable es la temperatura) clasifica a los meses en fríos y calientes sin notarse ningún efecto notorio del mercurio.

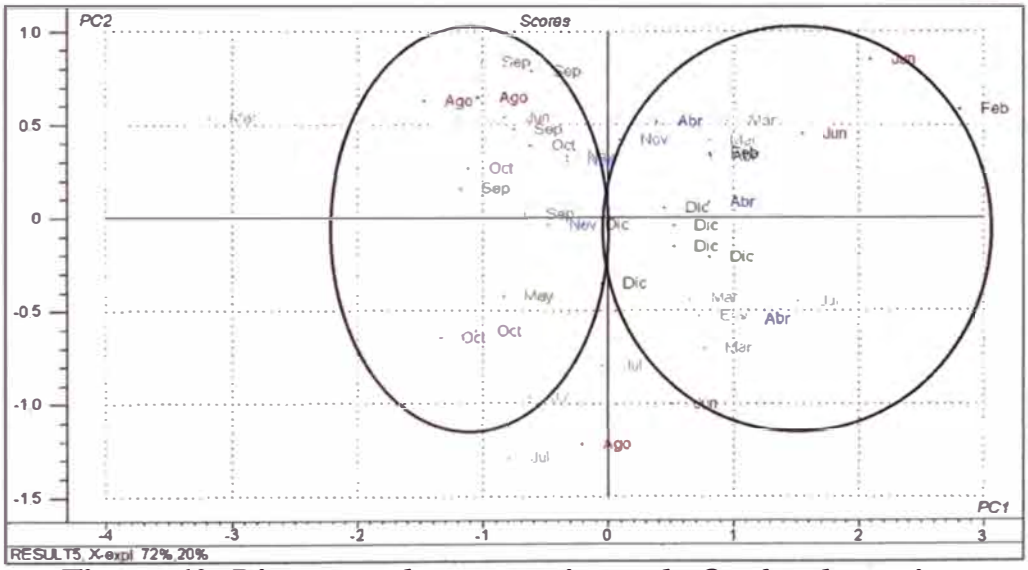


Figura 69. Diagrama de puntuaciones de Quebrada Pariñas Salida (Punto F)

4. ANÁLISIS DE LOS COMPONENTES PRINCIPALES CON REDUCCIÓN DE VARIABLES

En los diferentes diagramas de cargas realizadas anteriormente, se ha observado que las variables que influyen en la construcción de los componentes principales se agrupan en pH, temperatura y aquellas referidas a la salinidad de las aguas (cloruros, STD y conductividad). Además en la gráfica de correlaciones (Figura 39) se observa también una buena correlación entre cloruros, STD y conductividad. Es por ello que se estima que los objetos deberían presentar una distribución similar a la mostrada en la Figura 50 (diagrama de puntuaciones con todas las variables) cuando se representen en el espacio formado por el pH, la temperatura y solo una de las variables correspondientes a la salinidad. Esta reducción de variables se hará a través del software. Efectivamente, en la Figura 70 se observa que reduciendo las variables a tres (pH, temperatura y STD) se logra una misma clasificación de los objetos, lo que podríamos utilizar en un análisis clasificatorio como el análisis discriminante.

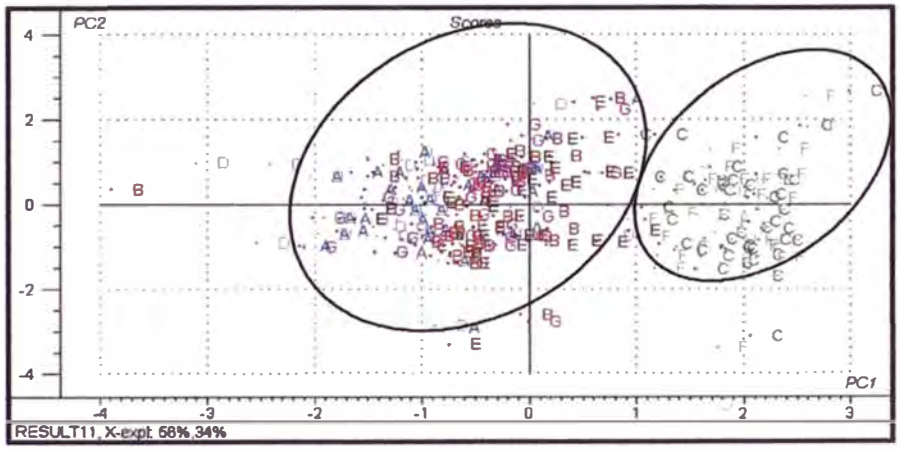


Figura 70. Gráfico de puntuaciones sobre el plano de los componentes 1 y 2 de todos los objetos utilizando una reducción de variables

5. ANÁLISIS PCA PARA LOS PUNTOS DE MUESTREO DEL LOTE IX

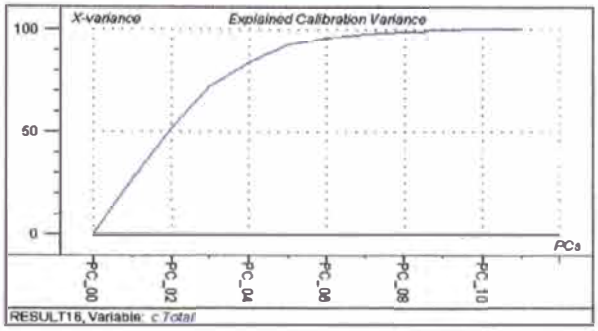


Figura 71. Diagrama de sedimentación para los puntos del Lote IX

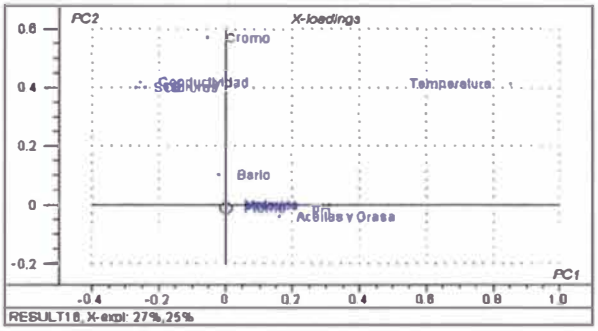


Figura 72. Diagrama de cargas para los puntos del Lote IX

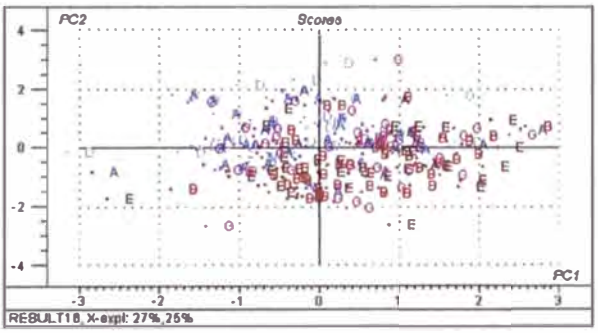


Figura 73. Diagrama de puntuaciones para los puntos del Lote IX

Las Figuras 71, 72 y 73, nos muestran que no hay mayor distinción entre los puntos correspondientes al Lote IX, puesto que hay una distribución sin agrupamiento, lo que indica una distribución normal de los datos.

6. ESTUDIO AMBIENTAL MEDIANTE PCA

Para averiguar la posible contaminación de la Quebrada Pariñas por parte de la empresa, se analizarán las fuentes receptoras en relación a las fuentes contaminantes.

6.1. PCA PARA PARIÑAS ENTRADA (C) Y POZA API 401 (E)

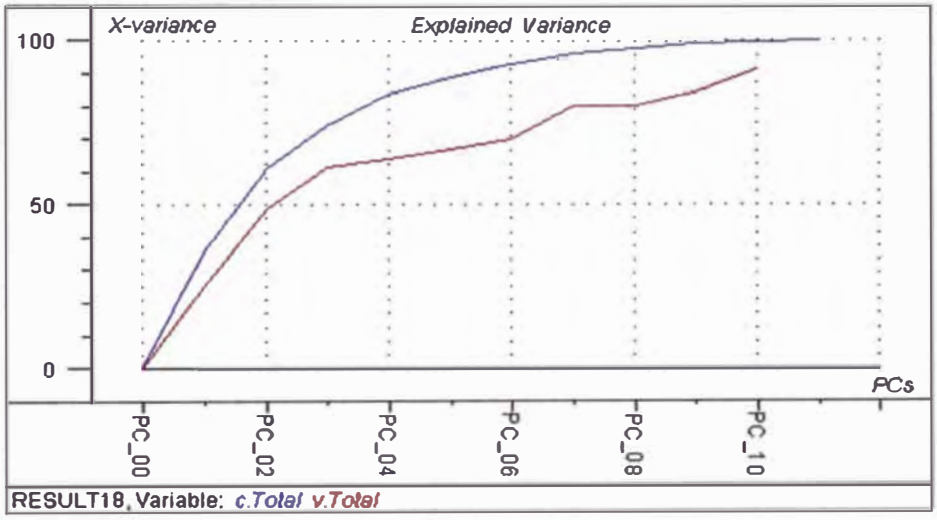


Figura 74. Diagrama de sedimentación para los puntos C y E

En relación a la Figura 74 se muestra, que el número de componentes principales esta dado por 4 componentes (85%), pero analizaremos solo dos componentes debido a que con ellos se obtiene un 61% de varianza explicada acumulada.

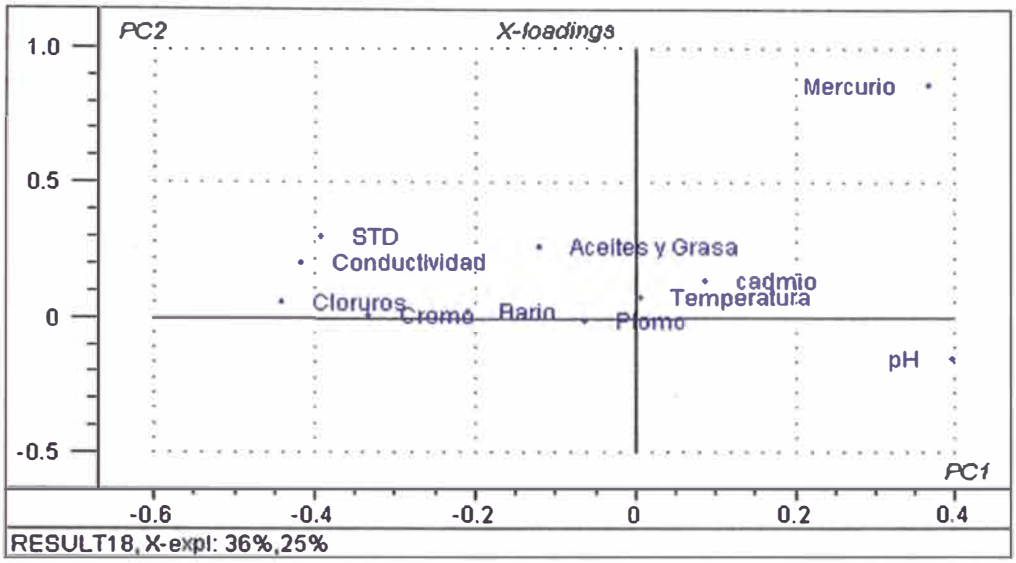


Figura 75. Diagrama de cargas para los puntos C y E

Como se observa en la Figura 75, la primera componente principal está relacionada a la salinidad, incluyendo el pH, y logra clasificar a las aguas en: aguas de producción y aguas correspondientes a la Quebrada Pariñas Entrada, tal como se observa en el diagrama de puntuaciones (Figura 76).

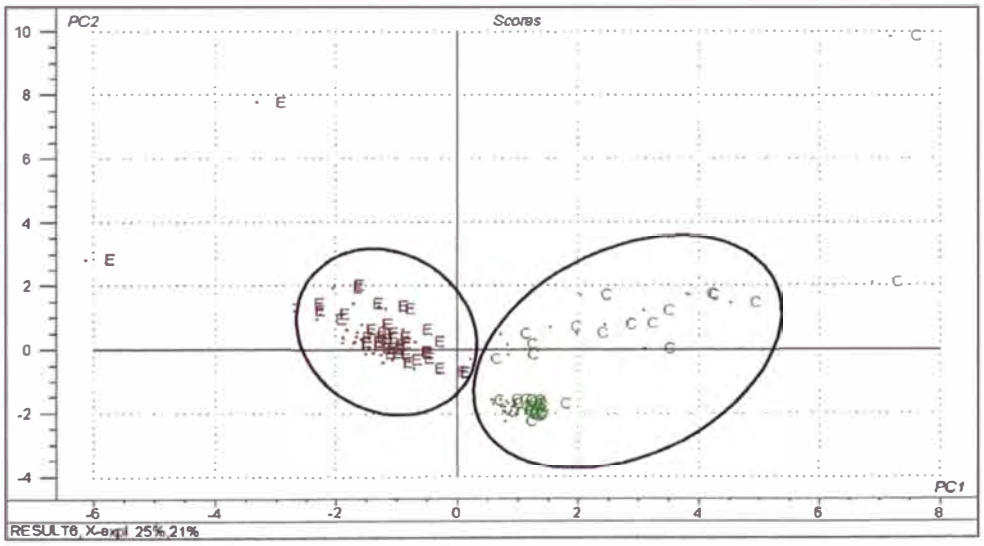


Figura 76. Diagrama de puntuaciones mostrando la separación de los puntos C y E

6.2. PCA PARA PARIÑAS SALIDA (F) Y POZA API 175 (A)

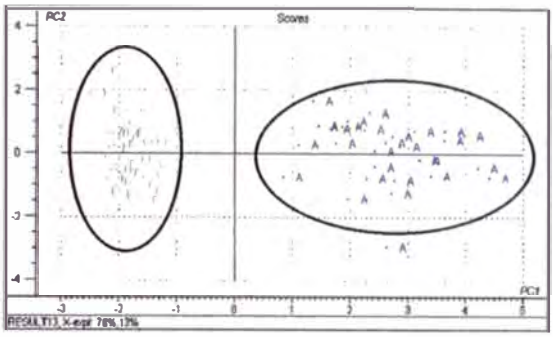


Figura 77a

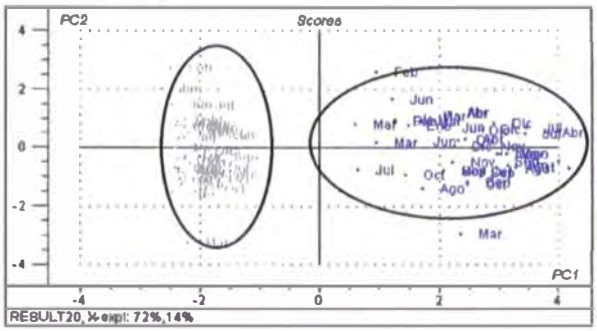


Figura 77b

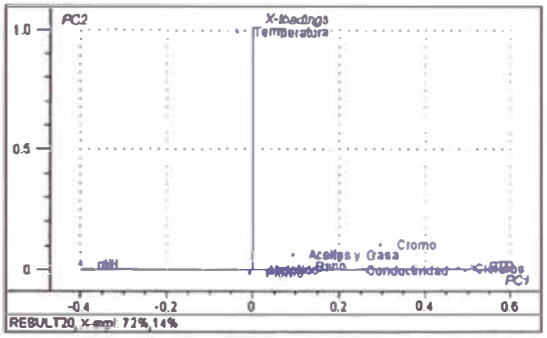


Figura 77c

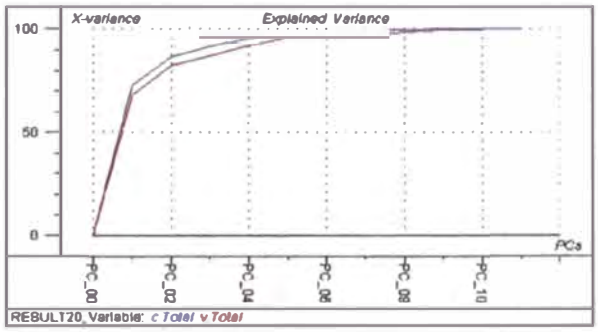


Figura 77d

Figura 77. Gráficos de puntuaciones (Figura 77a y b), de cargas (Figura 77c) y de sedimentación (Figura 77d) para los puntos F y A

La Figura 77a nos indica que no hay ningún tipo de correlación entre las aguas de la Quebrada Pariñas Salida con las aguas de producción procedente de la Poza API 175. La Figura 77c nos nuestra que la temperatura es la que más influye en la segunda componente principal; las variables de contaminación (aceites y grasas y salinidad) caracterizan a los puntos A, mientras que el pH es lo que caracteriza a la Quebrada Pariñas Aguas Abajo.

7. ANÁLISIS CLASIFICATORIO

El Análisis Clasificatorio buscará dividir el espacio en regiones características de clases. En nuestro caso serán considerados dos grupos: Aguas de Producción (1) y Aguas del Río Pariñas (2). Para lograr esta clasificación utilizaremos el Análisis Discriminante tratando de construir, a partir de los datos, una Función Discriminante. Entonces se utilizará como herramienta de análisis el software SPSS v.13.0.

Se analizaron las variables: temperatura, pH, cloruros, STD, aceites y grasas, bario, cadmio, mercurio, plomo y cromo. Se descartaron cloruros y conductividad por estar correlacionadas con STD. Para este procedimiento se requiere que las variables sean linealmente independientes. Todos los cuadros utilizados en esta sección se obtuvieron utilizando el software SPSS v.13.0.

Tabla 4. Resumen del Procedimiento de análisis

Casos		N	Porcentaje
Validos		295	99.0
Excluidos			
	Fuera de rango	0	0.0
	Al menos una de las variables discriminantes perdida	1	0.3
	Tanto Fuera de rango o Al menos una de las variables discriminantes perdida	2	0.7
	Total	3	1.0
Total		298	100.0

De la Tabla 4, se observa que de los 298 datos, 295 han sido utilizados para el análisis discriminante; solo 3 han sido excluido, lo que muestra que los datos pueden ser clasificados en forma satisfactoria.

Tabla 5.Resultado de la prueba[&]

Box's M		5720.762
F	Approx.	121.826
	df1	45
	df2	94377.578
	Sig.	.000

[&]Prueba de hipótesis nula de matrices de covarianza de igual población

La Tabla 5, nos indica que el grado de significancia es 0,000, lo que demuestra que las variables son independientes entre sí (se descartaron dos variables altamente correlacionadas: conductividad y cloruros). Está condición es importante para obtener la Función Discriminante.

Tabla 6. Coeficientes de la Función Discriminante

	Function
	1
Temperatura	.009
pH	-.996
STD	.000
Aceites	.006
Bario	.063
Cadmio	-.171
Cromo	.483
Mercurio	-1.259
Plomo	.384
(Constant)	3.405

Unstandardized coefficients

La Tabla 6, muestra los coeficientes calculados para la función discriminante. La Función Discriminante (D) será, por lo tanto:

$$D = 3{,}405 + 0{,}009 * Temperatura - 0{,}996 * pH + \ldots\ldots + 0{,}384 * plomo$$

Los resultados de la clasificación se muestra en el Anexo 4. Se observa que solo 3 valores difieren de la agrupación asignada con la predicha, lo cual indica que las variables asignadas clasifican muy bien los tipos de aguas estudiadas.

Ahora hagamos un análisis clasificatorio utilizando las tres variables estudiadas ya anteriormente (en el PCA) y que son las que tienen mayor importancia en la clasificación del tipo de agua (pH, temperatura, STD).

Tabla 7. Resumen del Procedimiento de análisis reduciendo variables

Casos		N	Porcentaje
Validos		296	99.3
Excluidos			
	Fuera de rango	0	0.0
	Al menos una de las variables discriminantes perdida	0	0.0
	Tanto Fuera de rango o Al menos una de las variables discriminantes perdida	2	0.7
	Total	2	0.7
Total		298	100.0

Al reducir a sólo tres variables el análisis se observa que solo dos datos han sido excluidos (Tabla 7), a diferencia del caso anterior (todas las variables manifiestas) que obligaron a eliminar tres.

Tabla 8.Resultado de la prueba[&]

Box's M		16.325
F	Approx.	2.682
	df1	6
	df2	171413.6
	Sig.	.013

[&]Prueba de hipótesis nula de matrices de covarianza de igual población

Al aplicar la prueba de hipótesis nula, la significancia menor a 0,05 indica que las variables son independientes entre si (Tabla 8).

Esto nos permite construir una nueva Función Discriminante, solo con las tres variables tomadas en cuenta, siendo los coeficientes obtenidos los mostrados en la Tabla 9.

Tabla 9. Coeficientes de la nueva Función Discriminante

	Function
	1
Temperatura	.007
pH	-1.134
STD	.000
(Constant)	5.011

Unstandardized coefficients

La Nueva Función Discriminante (D) será, por lo tanto:

$$D = 5{,}011 + 0{,}007 * Temperatura - 1{,}134 * pH + 0{,}000 * STD$$

Al aplicar esta nueva función discriminante a los datos, se obtiene el mismo cuadro que se muestra en el Anexo 4, en forma comparativa la clasificación asignada con la predicha. Solo dos datos fueron excluidos de la función discriminante.

CAPÍTULO IV
CONCLUSIONES Y RECOMENDACIONES

- El estudio de los datos mediante el Análisis de Componentes Principales (PCA) nos permite clasificar los resultados según su procedencia: en aguas de producción y en aguas del río Pariñas. Esta observación nos permite decir que es posible un estudio de clasificación de muestras según sus múltiples datos fisicoquímicos. Este trabajo se presenta como un aporte al estudio de clasificación de datos químicos.
- De los estudios con PCA y Análisis Discriminante se ha podido obtener una clasificación de las aguas de producción y de las aguas del río Pariñas.
- De los histogramas se observa que algunas de las variables exceden los valores límites permisibles, por ejemplo un buen número de datos de aceites y grasas exceden el valor máximo de 50 mg/L.
- Además se mostró que es posible hacer una reducción de variables (reducción de número de análisis) para obtener una misma clasificación. Esto también significa que es posible caracterizar las aguas analizadas por la Empresa UNIPETRO ABC S.A.C. con un menor número de variables, lo que significaría menor costo para la empresa. Sin embargo, es el Ministerio de Energía y Minas el que dispone qué variables son las que deben analizarse.

En base a esta Tesis también podríamos recomendar lo siguiente:

- La Empresa UNIPETRO ABC S.A.C. debe de tomar los previsiones del caso para reducir las concentraciones de aquellos parámetros que están superando los límites máximos permisibles, tales como aquellas relacionadas a la salinidad.
- El Ministerio de Energía y Minas en lugar de considerar entre los parámetros de monitoreo, tres variables correlacionadas entre si (STD, conductividad y cloruros) debería de incluir otras como DQO (porque las aguas de refinerías contienen contaminantes orgánicos no biodegradables), Oxígeno Disuelto (ya que la baja concentración de este elemento puede ser un indicador de que el agua tiene una alta carga orgánica, afectando la flora y fauna acuáticas), fenoles (frecuentemente presentes en altas concentraciones en las aguas residuales de la industria petrolera y que causan manchas en los peces y aumentan los cloruros cuando están en bajas concentraciones), o amoniaco (con frecuencia presente en las aguas residuales de las refinerías y que origina una consumo del oxígeno).
- Es posible aplicar el procedimiento seguido en esta tesis para un estudio de clasificación en los casos de: procedencia de productos, adulteración de los mismos, beneficios de un producto, etc. a partir de sus datos químicos.
- Introducción de la asignatura de Estadística en la currícula de la especialidad de Química.

Referencias Bibliográficas

[1] Ramis Ramos, Guillermo
Quimiometría
Editorial Síntesis, 1ª edición, Madrid-España, 2001

[2] Grupo de Quimiometría y Cualimetría de Tarragona, Universidad Rovira i Virgili, Tarragona , España.
Ouimiometría, Una disciplina útil para el análisis químico (Artículo)
http://www.quimica.urv.es

[3] Rui Sánchez, Itziar, PhD, Universidad Rovira i Virgili (Universidad Pública de Tarragona, España.
Introducción a la Química Analítica Avanzada, Curso 1999-2000
http://www.quimica.urv.es

[4] Salvador Figueras, M.
Introducción al Análisis Multivariante, [Curso en línea - 2000]
http://www.5campus.com/leccion/anamul

[5] Peña, Daniel
Análisis de Datos Multivariantes
Editorial Mc Graw Hill – Interamericana, 1ª. Edición, España-Madrid
ISBN 8448136101

[6] Blanco Romia, Marcelo, Universidad Autónoma de Barcelona
"Desarrollo de nuevas tecnologías analíticas en el control de calidad de la industria farmacéutica"
Tesis Doctoral, 2001

[7] Nieto Barajas, Luis E.
"Diplomado en Estadística Aplicada, Módulo 6, Análisis Multivariante"
http://allman.rhon.itam.mx/~lnieto

[8] Porcel García, Marta, Universidad Autónoma de Barcelona, Dpto de Química
"*Aplicación de Técnicas Quimiométricas para el desarrollo de Nuevos Métodos Cinético-Espectrofotométricos de Análisis"*
Tesis Doctoral, 2001

[9] Wold, Svant
"Chemometric; what do we mean with it, and what do we want from it?
Chemometric and Intelligent Systems, 30(1995), pp 109-115.

[10] Ministerio de Energía y Minas
Protocolo de Monitoreo de Calidad de Agua
http://www.minem.gob.pe/archivos/dgaae/legislacion/guias/protocalidaagua.pdf

[11] Empresa Petrolera UNIPETRO ABC S.A.C.
Programa de Adecuación y Manejo Ambiental Lote IX Tomo 1
Enero 1996

[12] Norma Técnica Ecológica Mexicana NTE-CCA-003/88
"Límites máximos permisibles y el procedimiento para la determinación de contaminantes en las descargas de aguas residuales en cuerpos de agua, provenientes de la industria de refinación de petróleo crudo, sus derivados y petroquímica básica"

[13] Empresa Petrolera UNIPETRO ABC. S.A.C., 1999
Cataño Cauti, Víctor
1000 años de Petróleo en el Perú. Apuntes para la historia.

[14] Ministerio de Energía y Minas
Mapa de lotes con contrato para operaciones petroleras en el noroeste del Perú
http://www.minem.gob.pe/archivos/dgh/mapas/graf/mapa2.pdf

[15] Otto, Mathias
Chemometrics. Statistics and Computer Application in Analytical Chemistry
Editorial Wiley, 1ª edición en inglés, Alemania, 1999.

[16] Larrechi, M.S.
"Exploración de resultados multidimensionales: Análisis de agrupaciones mediante métodos jerárquicos"
Departament de Química Analítica i Química Orgànica. Universitat Rovira i Virgili. PI. Imperial Tarraco, Tarragona, España.

ANEXOS

ANEXO 1
TRATAMIENTO MATEMÁTICO PARA EL ANÁLISIS DE LOS COMPONENTES PRINCIPALES (PCA)

TRATAMIENTO MATEMÁTICO PARA EL ANÁLISIS DE LOS COMPONENTES PRINCIPALES (PCA)

Si determinamos m parámetros (llamados **variables**, en lenguaje quimiométrico) correspondientes a n muestras diferentes (llamados **objetos** en quimiometría), los resultados (**variables manifiestas**), podemos expresarlas como una tabla, tal como la Tabla 1A:

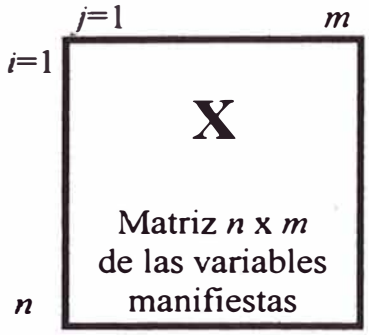
Tabla 1A
Datos objetos-variable genericos

Variable		Nombre 1	Nombre 2, ...	Nombre j, ...	Nombre m
Objetos	No.	1	2, ...	j, ...	m
Nombre 1	1	x_{11}	x_{12}, ...	x_{1j}, ...	x_{1m}
Nombre 2	2	x_{21}	x_{22}, ...	x_{2j}, ...	x_{2m}
.....	...	...	...	...	...
Nombre i	i	x_{i1}	x_{i2}, ...	x_{ij}, ...	x_{im}
.....	...	...	...	...	...
Nombre n	n	x_{n1}	x_{n2}, ...	x_{nj}, ...	x_{nm}

Estos mismos resultado podríamos expresarlo mediante una matriz de datos, **X**,

(A1)
$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}$$

Es decir nuestra matriz dato, **X**, de término general x_{ij}, es una matriz en la que cada fila representa un objeto y cada columna una variable medida:



No olvidemos que un vector *n*-dimensional es una cantidad geométrica que en un sistema usual de coordenadas se representa por una *n*-epla de números reales (una matriz columna de *n* filas). Por ejemplo el vector $\mathbf{x}$ quedaría expresado como:

(A2)
$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}$$

Asimismo, podríamos expresar dicho vector $\mathbf{x}$ por medio de su transpuesta, $\mathbf{x}^T$:

(A3)
$$\mathbf{x}^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_{n-1} & x_n \end{pmatrix}$$

La matriz de datos, $\mathbf{X}$, está formada por los vectores $\mathbf{x}_j$ correspondientes a las variables:

(A4)
$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_j & \cdots & \mathbf{x}_m \end{pmatrix}$$

Los métodos multivariados tienen por primer objetivo primordial permitir al investigador interpretar y visualizar conjuntos grandes de datos tanto en objetos como en variables (es decir le permite usar gráficos en lugar de grandes tablas). El otro gran objetivo primordial de los métodos multivariados es encontrar relaciones entre las variables, entre los objetos y entre ambos. En nuestro caso, el Análisis en Componentes Principales (ACP, o PCA por sus siglas en inglés) cubre aquellos métodos multivariados que se enfocan a estudiar relaciones entre las variables, tratando de disminuir la dimensionalidad de éstas.

PCA encontrará una combinación lineal de las variables independientes que aporten la máxima cantidad de información, que no es otra cosa que la mayor cantidad de variación..

Para lograr esta tarea PCA construye nuevos ejes en los cuales se ubicará el espacio correspondiente a nuestro sistema en estudio. Estos ejes, llamados componentes en lenguaje quimiométrico, serán normalizadas y para trabajar adecuadamente, suele ser necesario escalar la matriz de las variables manifiestas, $\mathbf{X}$, realizando al menos un centrado por columna (restando la media, $\overline{\mathbf{x}_j}$) y con frecuencia también será necesario un autoescalado (dividiendo entre la desviación estándar de cada columna, $\mathbf{s}_j$). Es decir haremos una transformación lineal de nuestra matriz.

Una buena práctica consiste en analizar cada variable por separado, mediante la metodología del Análisis Exploratorio de datos, haciendo el correspondiente análisis unidimensional, calculando los estadísticos que se crean convenientes para un mejor conocimiento de cada variable.

La media para cada variable se obtiene promediando por columnas la matriz de datos:

(A5) $$\bar{\mathbf{x}}_j = \frac{1}{n}\sum_{1}^{n} x_{kj}$$

de donde definimos la matriz fila de medias:

(A6) $$\bar{\mathbf{X}} = \left(\bar{\mathbf{x}}_1 \quad \bar{\mathbf{x}}_2 \quad \ldots \quad \bar{\mathbf{x}}_m\right)$$

Ahora se puede centrar la matriz de datos, restando a cada columna su valor medio:

(A7) $$\mathbf{X}^* = \mathbf{X} - \bar{\mathbf{X}}$$

Para determinar la relación entre los objetos podemos calcular las varianzas y covarianzas de los variables:

(A8) $$s_i^2 = \mathrm{var}(\mathbf{x}_i) = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)^2 = \frac{1}{n}\sum_{k=1}^{n} x_{ki}^2 - \bar{x}_i^2$$

(A9) $$s_{ij} = \mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) = \frac{1}{n}\sum x_{ki}x_{kj} - \bar{x}_i\,\bar{x}_j$$

En la Ec. A9 los primeros sumandos se pueden escribir en forma matricial $\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ mientras que los sustraendos se escriben $\bar{\mathbf{X}}^{\mathrm{T}}\bar{\mathbf{X}}$; por lo tanto la matriz de varianzas - covarianzas es:

(A10) $$\mathbf{V}_{\mathbf{X}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} - \bar{\mathbf{X}}^{\mathrm{T}}\bar{\mathbf{X}}$$

Si empleamos la matriz de datos centrada, entonces:

(A11) $$\mathbf{V}_{\mathbf{X}} = \frac{1}{n}(\mathbf{X}^*)^{\mathrm{T}}\mathbf{X}^*$$

Evidentemente la matriz $\mathbf{V}_{\mathbf{X}}$ es simétrica ya que $\mathbf{V}_{\mathbf{X}}^{\mathrm{T}} = \mathbf{V}_{\mathbf{X}}$:

(A12)
$$\mathbf{V}_X = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{i1} & s_{i2} & \cdots & s_{ij} & \cdots & s_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mj} & \cdots & x_{nm} \end{pmatrix}$$

con $s_{ij} = s_i^2$ y $s_{ij} = s_{ji}$, $i,j = 1, ..., m$

El siguiente paso en PCA es lograr la visualización de los datos. Si solo tuviéramos que graficar los objetos en el espacio formado por 3 variables obtendríamos algo parecido a la figura siguiente:

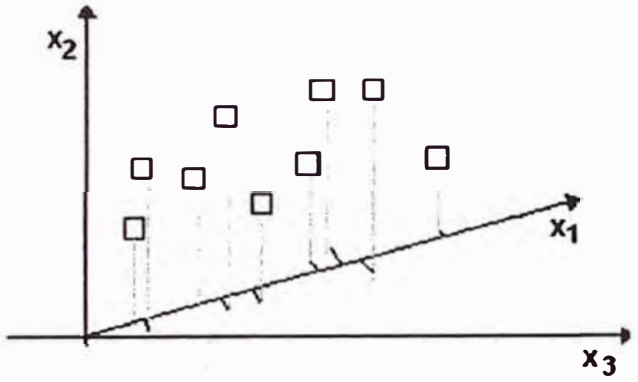


Figura 1A. Representación espacial de objetos en un espacio tridimensional

Sin embargo, en la mayoría de casos de análisis multivariante tendremos una gran cantidad de variables. Tendríamos que graficar en un espacio multidimensional, lo cual no sería posible. PCA buscará entonces reducir el número de dimensiones. Para ello es necesario seleccionar cuidadosamente nuevos planos de observación, mediante la selección de criterios racionales. La expresión algebraica de estos criterios de observación da origen a una serie de poderosas herramientas de análisis de datos, como por ejemplo, la rotación de ejes, seguida del análisis de los resultados obtenidos o "Análisis de los Componentes Principales".

La rotación de rotación de ejes permite encontrar una dirección en el espacio en la cual haya una mejor visión de lo datos. Por ejemplo, la correlación entre los objetos de la Figura 2A, podrían ser mejor representada si rotamos los ejes, como en la Figura 3A:

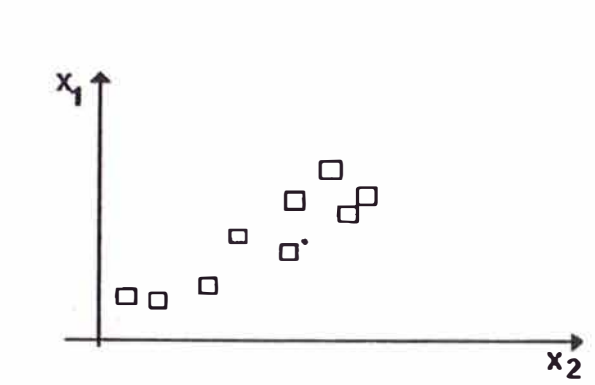


Figura 2A. Objetos en un espacio bidimensional

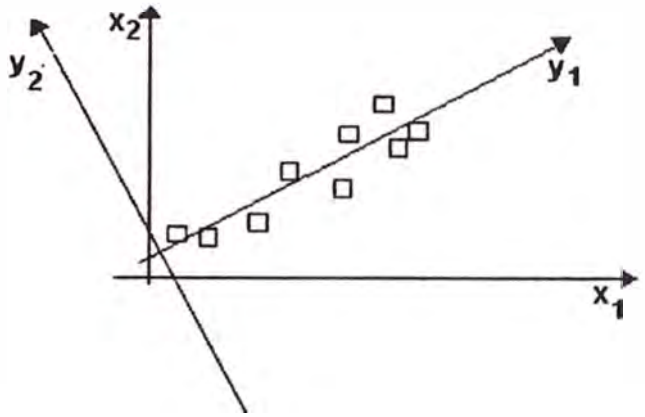


Figura 3A. Los mismos objetos de la Figura 2A, en un espacio bidimensional rotado.

Los nuevos ejes de coordenadas se denominan vectores propios (*eigenvectors*), autovectores, variables latentes, variables oscuras, factores principales o **componentes principales**. Como puede observarse en la Figura 3A, los mejores vectores son aquellos que apuntan en direcciones del espacio caracterizado por distribuciones estructuradas. Un vector describe adecuadamente una tendencia de la nube de datos cuando, en comparación con otras direcciones del espacio, los datos están fuertemente expandidos en la dirección del vector, y en tal caso, el vector representa un buen modelo de la nube multidimensional.

El objetivo de PCA es encontrar estos componentes principales (*Principal Component, PC*). PCA construye los componentes principales asumiendo que variación implica información. Esta variación puede ser clasificada como relevante o irrelevante. En análisis instrumental es menester obtener múltiples datos, muchos de los cuales son irrelevantes o no traen información adicional o traen información redundante sobre el sistema estudiado. PCA tratará entonces de seleccionar un número menor de factores o componentes (generalmente dos o tres) que vendrían a constituir la dimensionalidad intrínseca de los datos.

Los componentes principales gozan de algunas propiedades:

i) Son mutuamente ortogonales, lo que equivale a decir que no están correlacionados entre sí, evitando la redundancia de información.
ii) Los vectores propios pueden ser calculados en orden de varianza decreciente. Así, el primer componente principal trae la mayor información sobre la varianza de los datos, y cada sucesivo componente principal determinado traerá una menor información sobre la varianza.
iii) Aunque se pueden construir tantos componentes principales como variables se hayan determinado, la idea es reducir el número de dimensiones.

Gráficamente se obtendrá la transformación mostrada en la Figura 3A, el que se representan los objetos en la base formada por las variables $\mathbf{x_1}$, $\mathbf{x_2}$, y en la nueva base ortonormal formada por los vectores $\mathbf{y_1}$, $\mathbf{y_2}$, construida por combinación de las variables $\mathbf{x_i}$.

La situación es habitual dentro del análisis multivariante: la de la transformación lineal de variables para simplificar la representación de la nube de puntos. Se trata pues de la transformación de las variables originales:

(A13) $$\mathbf{X} = (\mathbf{x_1} \quad \mathbf{x_2} \quad \ldots \quad \mathbf{x_m})$$

en otras variables:

(A14) $$\mathbf{Y} = (\mathbf{y_1} \quad \mathbf{y_2} \quad \ldots \quad \mathbf{y_m})$$

mediante transformaciones lineales:

(A15) $$\mathbf{y_j} = \varphi_{ij}\,\mathbf{x_1} + \varphi_{2j}\,\mathbf{x_2} + \ldots + \varphi_{mj}\,\mathbf{x_m} \qquad j = 1, 2, \ldots, m$$

O sea que cada objeto , el k-ésimo por ejemplo, se transformaría así:

(A16) $$y_{kj} = \varphi_{1j}\,x_{k1} + \varphi_{2j}\,x_{k2} + \ldots + \varphi_{mj}\,x_{kp}, \qquad j = 1, 2, \ldots, m$$

y en forma matricial se escribiría:

(A17) $$\mathbf{Y} = \mathbf{T}^{\mathbf{T}}\mathbf{X}$$

siendo **X** la matriz de datos originales e **Y** la matriz de datos transformados (la imagen de **X**), y donde las columnas de **T**, la matriz de transformación, son los coeficientes de la transformación. En estas transformaciones las columnas de **T** son las coordenadas de las imágenes **Y** de los vectores de la base.

Fácilmente se comprueba que la matriz fila de medias se transforma de la siguiente manera:

(A18) $$\overline{\mathbf{Y}} = \mathbf{T}^{\mathbf{T}}\overline{\mathbf{X}}$$

Es importante señalar que la reducción de la dimensionalidad no implica que alguna de las variables originales haya sido descartada. Después de optimizar el número de componentes principales, el número de datos no es menor al que había antes de aplicar PCA. Solo se descartarán ciertas combinaciones de las variables originales.

La transformación por componentes principales es una transformación que preserva la varianza. Por lo tanto, las matrices de varianzas-covarianzas $\mathbf{V_X}$ y $\mathbf{V_Y}$ deben ser semejantes. Esto significa que:

(A19) $$\mathbf{V_Y} = \mathbf{T}^{\mathbf{T}}\mathbf{V_X}\mathbf{T}$$

El objetivo del análisis de componentes principales es transformar el espacio de representación P en un nuevo espacio, P', en el que los datos estén **incorrelados** (la matriz de covarianza en ese espacio será *diagonal*). En otras palabras, ***se trata de encontrar un nuevo conjunto de ejes ortogonales en el que la varianza de los datos sea máxima.*** El objetivo final es reducir la dimensionalidad del problema una vez realizada la transformación.

El problema se puede plantear a partir de una ejemplo sencillo. Representamos un conjunto de objetos bidimensionales que presentan cierto grado de correlación (Figura 4A). En esta representación utilizamos una elipse en lugar de la nube de puntos. Si representamos estos objetos en un nuevo espacio generado por las variables $\mathbf{y_1}$ y $\mathbf{y_2}$ (Figura 5A), que se corresponden con los ejes de la elipse y consideramos únicamente $\mathbf{y_1}$, la proyección de los objetos sobre este eje hace que su dispersión sea mayor que sobre cualquier otro eje (y en particular sobre cualquiera de los originales).

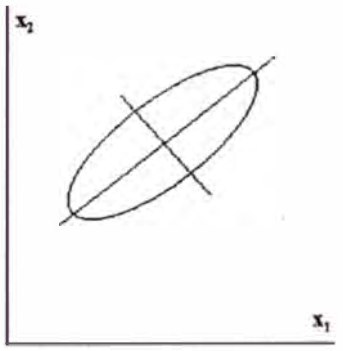


Figura 4A.
$\mathbf{x_1}$ y $\mathbf{x_2}$ están correlados

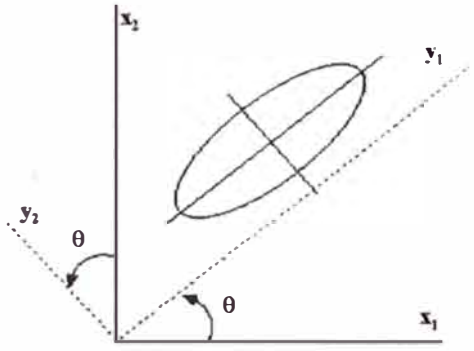


Figura 5A.
$\mathbf{y_1}$ y $\mathbf{y_2}$ están incorrelados

La trasformación consiste, básicamente, en una rotación-traslación de los ejes de P tomando como referencia el centroide de la nube de datos. La consecuencia es

que si los ejes de P' deben ser ortogonales, la distancia Euclídea entre dos puntos se mantiene inalterada con esta transformación. Para que esto sea cierto, la matriz de transformación $\mathbf{T}$ debe ser *ortogonal*, esto es, que $\mathbf{T}^{-1} = \mathbf{T}^{T}$, por lo que

(A20) $$\mathbf{T}^{T}\mathbf{T} = \mathbf{T}\mathbf{T}^{T} = \mathbf{I}$$

En definitiva, buscamos una matriz cuadrada $m \times m$ que sea ortogonal.

Si los datos en P' deben estar incorrelados, la matriz de correlación en P', $\mathbf{V_Y}$, debe ser *diagonal*.

Las condiciones para hallar los componentes principales son por lo tanto:
- Deben ser ortonormales
- Deben expresar la máxima varianza.

El problema que nos ocupa puede formularse como un problema de maximización (varianza en P') con restricciones (ortogonalidad de $\mathbf{T}$). La técnica adecuada es la utilización de los *multiplicadores de Lagrange*, que puede plantearse como sigue.

Si el objetivo es maximizar una función $f(v_1,v_2,...,v_p)$ con la condición $g(v_1,v_2,...,v_p) = 0$ se puede construir una nueva función

(A21) $$F = f(v_1, v_2,..., v_p) - \lambda\, g(v_1, v_2,..., v_p)$$

y maximizar esta función sin restricciones.

En nuestro caso, se trata de maximizar la varianza en P', por lo que $f = \mathbf{V_Y} = \mathbf{T}^{T}\mathbf{V_X}\mathbf{T}$ y la restricción es que $\mathbf{T}^{T}\mathbf{T} = \mathbf{I}$, por lo que $g = \mathbf{T}^{T}\mathbf{T} - \mathbf{I} = \mathbf{0}$. En definitiva, se trata de maximizar

(A22) $$\mathbf{F} = \mathbf{T}^{T}\mathbf{V_X}\mathbf{T} - \lambda(\mathbf{T}^{T}\mathbf{T} - \mathbf{I})$$

y derivando respecto a T,

(A23) $$(\mathbf{V_X} - \lambda\mathbf{I})\mathbf{T} = 0$$

y se tratará de encontrar la solución al sistema de ecuaciones dado por la Ec. A23. En definitiva, $\mathbf{T}$ *debe verificar que* $(\mathbf{V_X} - \lambda\mathbf{I})\mathbf{T} = 0$ *con objeto de que* $\mathbf{T}^{T}\mathbf{V_X}\mathbf{T} = \mathbf{V_Y}$ *sea máxima, sujeta a la restricción de que* $\mathbf{T}^{T}\mathbf{T} = \mathbf{I}$.

Para que la Ec. A23 sea cierta solo pueden ocurrir dos casos:

i. Que $\mathbf{T} = 0$ y en este caso la solución es *trivial* y no interesa.

ii. Que $\mathbf{V_X}$- $\lambda\mathbf{I}$ sea singular (no invertible), esto es, que

(A24) $$|\mathbf{V_X} - \lambda\mathbf{I}| = 0$$

La Ec. A24 es la ***ecuación característica*** de la matriz $\mathbf{V_X}$ y su expresión es una ecuación polinómica de λ. Las soluciones a esta ecuación (los valores de λ) se conocen como los ***autovalores*** de $\mathbf{V_X}$. Cuando se sustituyen en la Ec. A23, se calculan los vectores asociados a cada valor de λ, que se conocen como los ***autovectores*** de $\mathbf{V_X}$.

En decir, cada autovalor (λ_i) es solución a una ecuación del sistema ($\mathbf{V_X} - \lambda\mathbf{I}$) = 0. Así, para cada ecuación, los parámetros de $\mathbf{T}$ asociados a la solución con λ_i es un autovector, φ_i. De esta manera, podemos expresar la matriz de transformación $\mathbf{T}$ como un vector de m vectores columna ($\mathbf{T}$ es un vector de autovectores):

(A25) $$\mathbf{T} = (\varphi_1 \quad \varphi_2 \quad .... \quad \varphi_m)$$

Como $\mathbf{V_X}$ es de orden $m \times m$, tendrá m autovalores asociados, $\lambda_1, \lambda_2, ..., \lambda_m$ y como $\mathbf{V_X}$ es simétrica, todos los autovalores serán reales.

Algoritmo de Cálculo

Después del desarrollo teórico de la sección anterior, en esta sección presentaremos el algoritmo para el cálculo efectivo de la matriz de transformación.

El algoritmo de cálculo de la matriz de transformación $\mathbf{T}$ puede plantearse en 4 pasos:

i. Calcular la matriz de covarianza *global* $\mathbf{V_X}$.
   Para este cálculo se utilizan todos los valores de $\mathbf{X}$. En ningún caso se consideran *prototipos*, ya que no se tiene en cuenta la clase.
ii. Calcular los autovalores de $\mathbf{V_X}$, $\lambda_1, \lambda_2, \ldots, \lambda_m$.
iii. Calcular los autovectores $\varphi_1, \varphi, \ldots, \varphi_m$, asociados a $\lambda_1, \lambda_2, \ldots, \lambda_m$.
iv. Formar la matriz $\mathbf{T} = (\varphi_1 \quad \varphi_2 \quad \ldots \quad \varphi_m)$

Una vez formada la matriz de transformación $\mathbf{T}$ se procede a calcular los nuevos valores, Y, a partir de cada X. Como $\mathbf{Y} = \mathbf{T}^T\mathbf{X}$ y la matriz $\mathbf{T}$ es la matriz formada por los autovectores de $\mathbf{V_X}$, sustituyendo obtenemos:

(A26)
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1m} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{m1} & \varphi_{m2} & \cdots & \varphi_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

o lo que es lo mismo,

(A27)
$$\begin{aligned} y_1 &= \varphi_{11}x_1 + \varphi_{12}x_2 + \ldots + \varphi_{1m}x_m \\ y_2 &= \varphi_{21}x_1 + \varphi_{22}x_2 + \ldots + \varphi_{2m}x_m \\ &\vdots \\ y_m &= \varphi_{m1}x_1 + \varphi_{m2}x_2 + \ldots + \varphi_{mm}x_m \end{aligned}$$

donde

(A28)
$$\varphi_1 = \begin{pmatrix} \varphi_{11} \\ \varphi_{12} \\ \vdots \\ \varphi_{1m} \end{pmatrix}, \quad \varphi_2 = \begin{pmatrix} \varphi_{21} \\ \varphi_{22} \\ \vdots \\ \varphi_{2m} \end{pmatrix}, \quad \cdots, \quad \varphi_m = \begin{pmatrix} \varphi_{m1} \\ \varphi_{m2} \\ \vdots \\ \varphi_{mm} \end{pmatrix}$$

Ejemplo

Como ilustración, mostraremos cómo se aplica la transformación de componentes principales a un conjunto de datos que presenta cierta correlación. En la Figura 6A mostramos los 6 objetos sobre los que se va a efectuar la transformación. Como se observa, las variables x_1 y x_2 presentan una correlación positiva.

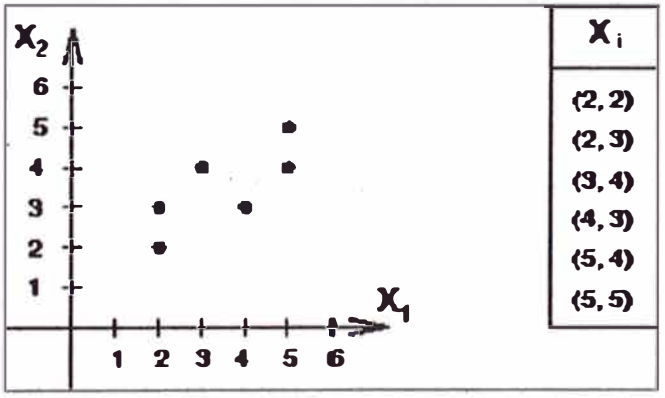


Figura 6A.- Objetos en el espacio original P

Nuestra matriz de datos será:

$$\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 3 \\ 3 & 4 \\ 4 & 3 \\ 5 & 4 \\ 5 & 5 \end{pmatrix}$$

1. Cálculo de $\mathbf{V_X}$

El vector medio $\overline{\mathbf{X}}$ y la matriz de covarianza $\mathbf{V_X}$ se calculan a partir de los vectores columna $\mathbf{x_1}$ y $\mathbf{x_2}$, obteniendo:

$$\overline{\mathbf{X}} = (3{,}50 \quad 3{,}50) \qquad \mathbf{V_X} = \begin{pmatrix} 1{,}9 & 1{,}1 \\ 1{,}1 & 1{,}1 \end{pmatrix}$$

El cálculo de la covarianza se realiza mediante la ecuación:

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Como esta covarianza tiene el inconveniente que depende de la escala en la que se expresan las variables, prefiere usarse el coeficiente de correlación de Pearson (r), que se calcula como:

$$r = \frac{\mathrm{cov}(x, y)}{s_x \, s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right) - \left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

2. Cálculo de los autovalores de $\mathbf{V_X}$

Como $m = 2$ habrán dos autovalores asociados a $\mathbf{V_X}$: λ_1 y λ_2. Serán las soluciones a la ecuación $|\mathbf{V_X} - \lambda \mathbf{I}| = 0$. En particular,

$$\left| \begin{pmatrix} 1{,}9 & 1{,}1 \\ 1{,}1 & 1{,}1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\begin{vmatrix} 1{,}9-\lambda & 1{,}1 \\ 1{,}1 & 1{,}1-\lambda \end{vmatrix} = 0$$

o lo que es igual,

$$\lambda^2 - 3\,\lambda + 0{,}88 = 0$$

y las soluciones son: $\lambda_1 = 2{,}67$ y $\lambda_2 = 0{,}33$

3. Cálculo de los autovectores φ_1 y φ_2 asociados a λ_1 y λ_2

El autovector φ_1, correspondiente a $\lambda_1 = 2{,}67$ se calcula como sigue. Considerando la ecuación $(\mathbf{V_X} - \lambda\mathbf{I})\mathbf{T} = 0$, el autovector φ_1 es la solución a $(\mathbf{V_X} - \lambda_1\mathbf{I})\,\varphi_1 = \mathbf{0}$. Esto es,

$$\left(\begin{pmatrix} 1{,}9 & 1{,}1 \\ 1{,}1 & 1{,}1 \end{pmatrix} - 2{,}67\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)\begin{pmatrix} \varphi_{11} \\ \varphi_{12} \end{pmatrix} = 0$$

$$\begin{pmatrix} 1{,}9-2{,}67 & 1{,}1 \\ 1{,}1 & 1{,}1-2{,}67 \end{pmatrix}\begin{pmatrix} \varphi_{11} \\ \varphi_{12} \end{pmatrix} = 0$$

$$\begin{pmatrix} -0{,}77 & 1{,}1 \\ 1{,}1 & -1{,}57 \end{pmatrix}\begin{pmatrix} \varphi_{11} \\ \varphi_{12} \end{pmatrix} = 0$$

o lo que es igual,

$$-0{,}77\,\varphi_{11} + 1{,}10\,\varphi_{12} = 0$$
$$1{,}10\,\varphi_{11} - 1{,}57\,\varphi_{12} = 0$$

Este sistema de ecuaciones tiene una solución no trivial porque el determinante es cero. Tomando cualquiera de ellas se deduce que

$$\varphi_{11} = 1{,}43\,\varphi_{12}$$

Como la matriz **T** debe ser ortogonal ($\mathbf{T^T} = \mathbf{T^{-1}}$) se requiere que los autovectores estén normalizados, esto es,

$$\varphi_{11}^2 + \varphi_{12}^2 = 1$$

por lo que el sistema de ecuaciones a resolver es:

$$\varphi_{11} = 1{,}43\ \varphi_{12}$$
$$\varphi_{11}^2 + \varphi_{12}^2 = 1$$

Por un lado se deduce que $\phi_{12}^2 = 1 - \varphi_{11}^2$. Por otro lado se deduce que $\varphi_{11}^2 = 1{,}43^2\ \phi_{12}^2$. Así,

$$\varphi_{11}^2 = 1{,}43^2\ (1 - \varphi_{11}^2) = 2{,}05 - 2{,}05\ \varphi_{11}^2$$

Reorganizando términos,

$$3{,}05\ \varphi_{11}^2 = 2{,}05 \Leftrightarrow \varphi_{11} = \sqrt{\frac{2{,}05}{3{,}05}} = 0{,}82$$

y como tenemos que $\varphi_{11} = 1{,}43\ \varphi_{12}$, entonces,

$$\varphi_{12} = \frac{\varphi_{11}}{1{,}43} = \frac{0{,}82}{1{,}43} = 0{,}57$$

El resultado es que el autovector asociado a $\lambda_1 = 2{,}67$ es

$$\varphi_1 = \begin{pmatrix} 0{,}82 \\ 0{,}57 \end{pmatrix}$$

El autovector φ_2, correspondiente a $\lambda_2 = 0{,}33$ se calcula de manera similar. El resultado es:

$$\varphi_2 = \begin{pmatrix} -0{,}57 \\ 0{,}82 \end{pmatrix}$$

Los autovectores que se acaban de calcular están normalizados. Esto implica que son de longitud 1. Efectivamente, para ambos autovectores:

$$\varphi_{11}^2 + \varphi_{12}^2 = 0{,}82^2 + 0{,}57^2 = 0{,}67 + 0{,}33 = 1$$
$$\varphi_{21}^2 + \varphi_{22}^2 = (-0{,}57)^2 + 0{,}82^2 = 0{,}33 + 0{,}67 = 1$$

Las componentes de un autovector indican la dirección de los nuevos ejes respecto al sistema de coordenadas original. La interpretación geométrica del nuevo sistema de coordenadas ($\mathbf{y_1}$, $\mathbf{y_2}$) respecto al original ($\mathbf{x_1}$, $\mathbf{x_2}$) en base a los autovectores φ_1 y φ_2 se detalla en la figura siguiente.

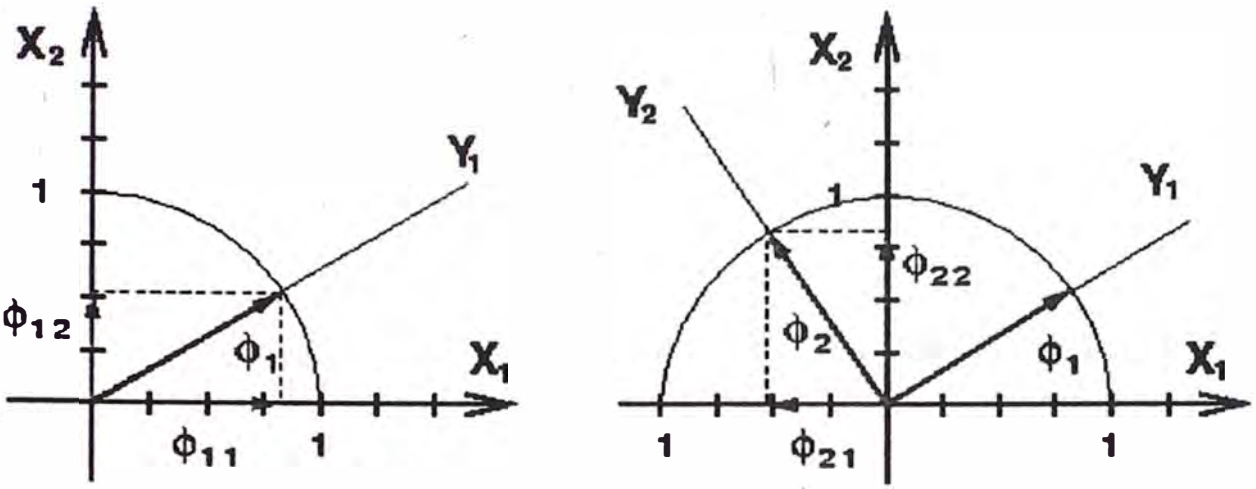


Figura 7A. Los autovectores determinan el nuevo sistema de coordenadas

4. Formar la matriz de transformación T

La matriz de transformación es una matriz cuadrada 2x2 cuyas columnas son los autovectores φ_1 y φ_2:

$$\mathbf{T} = (\varphi_1 \quad \varphi_2) = \begin{pmatrix} 0{,}82 & -0{,}57 \\ 0{,}57 & 0{,}82 \end{pmatrix}$$

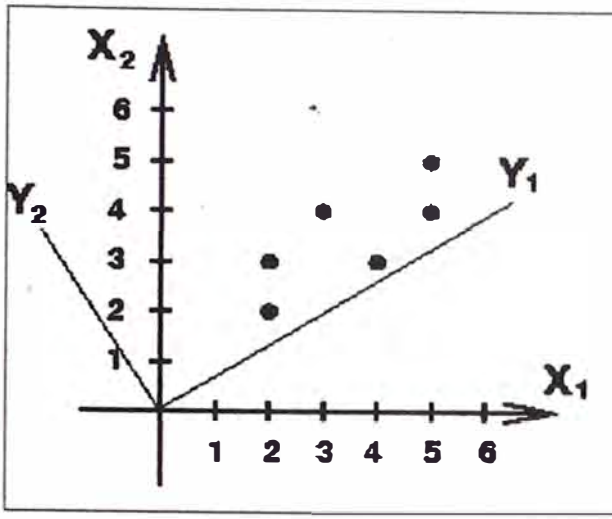
Finalmente, se procede a la transformación de coordenadas para expresar los objetos en las coordenadas del nuevo espacio. La transformación viene dada por la ecuación

$$\mathbf{Y} = \mathbf{T}^{\mathrm{T}}\mathbf{X}$$

o sea,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 0{,}82 & 0{,}57 \\ -0{,}57 & 0{,}82 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Si aplicamos esta transformación a los objetos, el resultado se muestra en la figura siguiente:



X_i	Y_i
(2, 2)	(2,78, 0,50)
(2, 3)	(3,35, 1,32)
(3, 4)	(4,74, 1,57)
(4, 3)	(4,99, 0,18)
(5, 4)	(6,38, 0,43)
(5, 5)	(6,95, 1,25)

Figura 8A. 6 objetos en dos sistemas de coordenadas

Conclusiones del Anexo 1

Como $\mathbf{V_X}$ es simétrica, todos sus autovalores serán reales. Por otra parte, dado que $\mathbf{V_X}$ es definida positiva, sus autovalores están ordenados:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_2$$

La matriz de covarianza

$$\mathbf{V_Y} = \mathbf{T}^{\mathbf{T}} \mathbf{V_X}\, \mathbf{T}$$

será una matriz *diagonal* formada por los *autovalores* de $\mathbf{V_X}$:

$$V_Y = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix}$$

y los valores de la diagonal (*autovalores de* $\mathbf{V_X}$) son las varianzas de los objetos en las respectivas coordenadas transformadas.

La matriz que contiene los coeficientes de la transformación, **T**, es la matriz de *autovectores de* $\mathbf{V_X}$, asumiendo que **T** es *ortogonal*.

Cada autovalor λ_i tiene asociado un autovector φ_i y cada autovector define la dirección de un eje en el espacio transformado, P'. Dado que los autovalores están ordenados (por el valor de varianza en cada eje de P'), y que cada autovalor tiene

asociado un autovector, podemos establecer un **orden** entre las variables transformadas de forma que:

$\mathbf{y}_1$: Primer eje en P' (*primera componente principal*).
La dirección de la máxima varianza de los patrones en P está determinada por este eje.

$\mathbf{y}_2$: Segundo eje en P' (*segunda componente principal*).
La dirección de la máxima varianza en P entre todos los ejes ortogonales a $\mathbf{y}_1$ está determinada por este eje.

Con estas consideraciones, el T***eorema Fundamental del Análisis de Componentes Principales*** se enuncia como sigue:

> *Dado un conjunto de variables* $\mathbf{x}_i$ *(* i = 1, 2,..., m*) con matriz de covarianza* $\mathbf{V_X}$*, no singular, siempre se puede derivar a partir de ellos un conjunto de variables incorreladas* $\mathbf{y}_i$ *(* i = 1, 2,..., m*) mediante un conjunto de transformaciones lineales que corresponden a una rotación rígida cuya matriz de transformación* $\mathbf{T}$ *está formada, por columnas, por los* m *autovectores de* $\mathbf{V_X}$*.*
> *La matriz de covarianza del nuevo conjunto de variables,* $\mathbf{V_Y}$*, es diagonal, y contiene los autovalores de* $\mathbf{V_X}$*.*

La transformación de componentes principales definida por $\mathbf{Y} = \mathbf{T}^T\mathbf{X}$ con la restricción de diagonalidad $\mathbf{V_Y}=E\{(\mathbf{Y}-\mathbf{Y})(\mathbf{Y}-\mathbf{Y})^T\} = \mathbf{T}^T\mathbf{V_X}\mathbf{T}$ se conoce también como ***transformación de Karhunen-Loéwe*** o de ***Hotelling***.

A modo de resumen, y con una interpretación geométrica, los autovalores λ_i representan la varianza de los objetos en el espacio transformado y están relacionados con el *rango* de los objetos en cada uno de los ejes de este espacio mientras que los autovectores φ_i son vectores ortogonales que determinan la *dirección* de estos ejes.

La transformación por componentes principales es una transformación que preserva la varianza.

Si se define la *varianza total* de un conjunto de datos multidimensionales como la suma de las varianzas asociadas a cada atributo, como las varianzas individuales se encuentran en la diagonal de la matriz de covarianza, el cálculo de la varianza global se reduce al cálculo de la *traza* de la matriz de covarianza. Resulta evidente que si $\mathbf{V_Y}$ es la matriz que contiene en su diagonal los autovalores $\lambda_1, \lambda_2, \ldots, \lambda_m$ de $\mathbf{V_X}$, entonces,

$$tr(\mathbf{V}_Y) = \sum_{i=1}^{m} \lambda_i$$

Además, si la transformación de componentes principales preserva la varianza global, entonces,

$$tr(\mathbf{V}_X) = tr(\mathbf{V}_Y) = \sum_{i=1}^{m} \lambda_i$$

Es importante saber que $\mathbf{V}_Y$ contiene los llamados autovalores, varianzas explicadas o "pesos" de cada componente. El término varianza explicada se origina cuando los datos han sido ajustados a un modelo (como el de los componentes principales).

Cuando un vector modela la nube de puntos, la varianza total de los datos, s_T^2 (la suma de cuadrados de las distancias de los puntos a su centroide), queda dividida en dos, la "*explicada*" por el vector y la "*residual*" (aquella que el vector no explica):

$$s_T^2 = s_{\exp}^2 + s_{res}^2$$

Si el vector es un componente principal, su varianza explicada, $s_{\exp}^2$, se denomina "autovalor". En la Figura 9A, la varianza explicada es la varianza del vector, o varianza de las proyecciones de los puntos sobre el vector, que se calcula como la suma de los cuadrados de las distancias de los puntos al centroide en la dirección del vector, dividido por *n*. La varianza residual es la suma de cuadrados de distancias de los puntos al modelo (líneas perpendiculares al vector), dividida también por n.

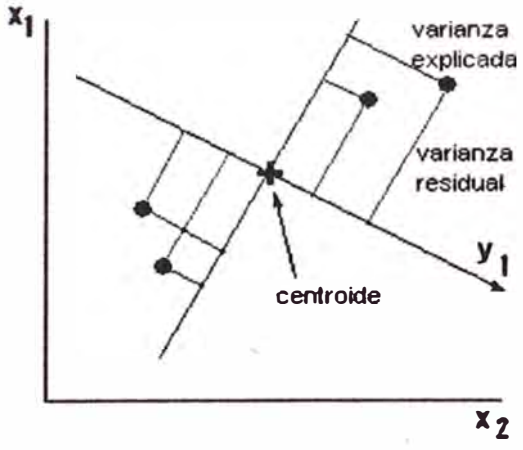


Figura 9A. Significados de la varianza explicada y residual respecto a un vector cualquiera y_1, que pasa por el centroide de cuatro puntos

Para nuestro caso tendremos:

	φ_1	φ_2
Valor propio	2,670	0,330
% varianza	89,016	10,984
% acumulado	89,016	100,000

Finalmente la contribución de cada componente principal será la mostrada en la Figura 10A, es decir el primer componente principal es el más importante,

Figura 10A. Aporte de cada componente principal

ANEXO 2
UBICACIÓN DE INSTALACIONES Y PUNTOS DE MONITOREO AMBIENTAL DE LA EMPRESA UNIPETRO ABC S.A.C.

E 460041.8600

E 463280.8600

N 9500581.8600

N 9495254.000

ALGARROBA

CUESTA

BATANES

LEONES

Borde de Tablazo

BATERIA 401

BAT. 175

MC 4

MC 1

LEYENDA

- CARRETERAS - ACCESOS
- PUNTOS DE MONITOREO AMBIENTAL
- BATERIAS
- MANIFOLD DE CAMPO
- POZO PRODUCTOR
- POZO ABANDONADO PERMANENTEMENTE (APA)
- POZO ABANDONADO TEMPORALMENTE (ATA)
- POZO ABANDONADO DPA

0 100 200 300 400 500 600 700 800 900 1000 m

EMPRESA PETROLERA

UNIPETRO ABC S.A.C.

LOTE IX

LEVANTAMIENTO TOPOGRAFICO UBICACION DE INSTALACIONES Y PUNTOS DE MONITOREO AMBIENTAL

Lev. Topog.: C. Aburto S.	Revisado: J. Peralta	Escala: 1: 10 000
Actualizado: Junio 2005	Archivo:	Figura:

N 9495453.000

E 460048.000

E 463265.000

ANEXO 3
PARÁMETROS FÍSICO-QUÍMICOS ANALIZADOS DE LAS AGUAS DE PRODUCCIÓN DE UNIPETRO ABC S.A.C.

Tabla de Datos de la Empresa UNIPETRO ABC S.A.C.

Lugar	Lugar	Fecha	Temperatura	pH	Cloruros	STD	Aceites y Grasas	Conductividad	Bario	Cadmio	Cromo	Mercurio	Plomo
Poza Api 175	A	Jul-98	68	6.75	*	38954	13.00	61.80	0.002	0.160000	0.120	0.002000	1.060000
	A	Ago-98	72	6.80	25560	37835	16.00	62.80	0.029	0.030000	0.003	0.002000	0.060000
	A	Set-98	70	5.60	975	33620	19.00	51.40	6.850	0.120000	0.040	0.002000	1.090000
	A	Oct-98	69	6.02	22141	38450	6.00	57.70	1.868	0.002400	0.112	0.000200	0.004200
	A	Nov-98	74	6.13	21501	36810	11.00	48.30	0.363	0.001200	0.036	0.000030	0.002000
	A	Mar-99	75	6.87	14226	24830	12.00	31.20	1.712	0.000100	0.024	0.000050	0.003000
	A	Jun-99	75	6.67	16089	29120	17.00	38.60	1.732	0.000150	0.060	0.000040	0.006000
	A	Set-99	68	6.28	16205	28480	3.00	45.10	2.876	0.000030	0.257	0.000010	0.006000
	A	Dic-99	77	6.57	17128	28920	28.00	48.40	1.140	0.000270	2.412	0.000020	0.009000
	A	Abr-00	79	6.62	12496	21210	12.45	35.50	2.003	0.000040	3.226	0.000020	0.003000
	A	Jul-00	75	7.04	22316	39151	19.21	55.00	1.538	0.000020	1.885	0.000180	0.001000
	A	Oct-00	67	6.09	22569	26860	4.00	38.20	1.100	0.000020	1.140	0.000050	0.009000
	A	Dic-00	77	5.79	20329	36055	9.45	29.20	1.557	0.000520	1.259	0.000150	0.001000
	A	Abr-01	81	5.62	14675	27735	38.67	34.40	1.207	0.001570	0.918	0.000110	0.000600
	A	Jul-01	77	6.08	20812	36705	13.00	49.30	2.710	0.000100	1.636	0.000040	0.002000
	A	Oct-01	64	6.22	17274	29430	1.00	41.20	1.798	0.000120	1.336	0.000100	0.002000
	A	Dic-01	79	6.24	10110	21375	16.72	28.70	0.354	0.000740	1.466	0.000020	0.001000
	A	Mar-02	79	6.92	15566	16720	5.00	23.90	0.219	0.000550	0.745	0.000120	0.002000
	A	Jun-02	78	6.10	16725	29680	7.00	39.10	0.195	0.000710	0.712	0.000080	0.005000
	A	Set-02	66	6.42	18914	30610	27.33	41.40	0.390	0.000820	1.320	0.000020	0.001500
	A	Dic-02	73	8.70	17217	28870	90.33	38.60	0.470	0.000730	1.210	0.000020	0.001200
	A	Abr-03	75	5.89	18403	33450	7.40	38.50	0.590	0.000650	0.940	0.000030	0.000900
	A	Jun-03	84	6.78	13606	23256	34.50	30.60	0.620	0.000710	0.970	0.000030	0.001100
	A	Jul-03	68	6.39	10418	13635	11.33	26.90	0.590	0.000620	0.840	0.000010	0.001200
	A	Ago-03	63	6.40	14419	26105	3.17	32.40	0.330	0.000790	1.260	0.000020	0.001700
	A	Set-03	65	5.04	16187	26165	1.17	35.50	0.300	0.000680	1.030	0.000020	0.001400
	A	Oct-03	66	8.51	13782	24530	52.20	31.80	0.873	0.000280	0.460	0.000250	0.001000
	A	Nov-03	70	6.08	16978	29550	14.63	36.50	1.185	0.000100	0.908	0.000080	0.008000
	A	Dic-03	73	5.95	15657	28586	22.84	38.00	0.932	0.000800	1.142	0.000160	0.009000
	A	Ene-04	79	6.63	16354	25035	4.64	35.00	1.303	0.000700	0.646	0.000180	0.008000
	A	Feb-04	79	6.67	12740	23815	13.14	29.40	0.641	0.000300	1.248	0.000140	0.016000
	A	Mar-04	79	6.94	15655	30585	5.00	32.50	1.458	0.000500	1.183	0.000330	0.060000
	A	Mar-04	80	6.39	14078	26160	8.18	30.40	1.520	0.000600	1.594	0.000080	0.042000
	A	Abr-04	76	5.58	22255	40305	3.62	45.90	1.766	0.000600	1.658	0.000070	0.009000
	A	May-04	68	8.04	12496	27425	2.25	41.60	0.930	0.000700	1.134	0.000020	0.008000
	A	Jun-04	68	6.44	19292	37920	3.64	42.00	1.813	0.000800	2.055	0.000020	0.009000
	A	Ago-04	67	5.82	21904	37510	12.50	55.00	2.051	0.000600	2.409	0.000300	0.008000
	A	Sep-04	69	5.96	23108	39785	5.33	55.70	1.487	0.000600	1.792	0.000180	0.009000
	A	Nov-04	71	6.64	20583	36995	13.23	50.70	1.034	0.000500	1.248	0.000050	0.007000
	A	Nov-04	71	6.22	19372	35115	7.90	47.40	1.054	0.000500	1.295	0.000030	0.010000
	A	Dic-04	77	6.78	18618	32740	5.88	47.40	1.165	0.000600	1.339	0.000030	0.010000
	A	Feb-05	91	7.23	10152	24690	9.00	38.60	0.829	0.000600	0.947	0.000060	0.007000
	A	Mar-05	52	6.82	15188	33780	9.83	46.80	0.921	0.000700	1.060	0.000050	0.007000
Agua MC-2	B	Jul-98	68	6.79	*	17218	6.00	27.90	0.002	0.100000	0.130	0.002000	0.800000
	B	Ago-98	72	6.92	11005	19235	9.00	32.20	5.250	0.002000	0.002	0.002000	0.010000
	B	Set-98	70	5.95	518	17970	13.00	28.60	1.070	0.020000	0.020	0.002000	0.800000
	B	Oct-98	69	5.86	11071	19010	12.00	31.40	0.870	0.001300	0.058	0.000300	0.002000
	B	Nov-98	74	6.05	11218	18830	13.00	26.80	0.479	0.000400	0.038	0.000030	0.003000
	B	Mar-99	78	6.30	10621	18410	21.00	28.00	0.753	0.000300	0.019	0.000050	0.001000
	B	Jun-99	75	6.27	12052	17250	22.00	23.80	0.708	0.000020	0.020	0.000020	0.011000
	B	Set-99	68	5.80	10887	19250	21.00	29.20	1.644	0.000030	0.280	0.000010	0.004000
	B	Dic-99	77	5.37	10804	18580	26.00	29.80	1.264	0.000020	2.504	0.000010	0.004000
	B	Abr-00	79	6.39	12320	18250	4.59	29.60	1.003	0.000040	3.753	0.000030	0.003000
	B	Jul-00	75	6.92	19405	34651	27.27	53.10	0.862	0.000020	1.728	0.000060	0.001000
	B	Oct-00	67	6.95	11285	17960	26.00	25.70	0.469	0.000010	1.050	0.000060	0.010000
	B	Dic-00	77	5.16	12659	22555	12.99	27.20	0.659	0.000150	1.511	0.000030	0.001000
	B	Abr-01	81	8.34	10236	19570	10.50	22.90	0.611	0.001720	1.150	0.000100	0.000500
	B	Jul-01	84	6.54	10469	20575	44.00	28.70	1.060	0.000150	1.740	0.000030	0.001000
	B	Oct-01	64	6.09	12454	18880	7.00	25.30	0.784	0.000100	1.511	0.000090	0.001000
	B	Dic-01	82	6.05	9309	21360	46.67	27.30	0.664	0.000070	1.754	0.000010	0.002000
	B	Mar-02	79	5.40	10334	20010	6.30	25.30	0.719	0.000300	1.322	0.000060	0.005000
	B	Jun-02	78	8.60	10621	19972	14.70	27.20	0.643	0.000400	0.522	0.000030	0.006000
	B	Set-02	66	6.24	12109	22405	38.33	28.30	0.685	0.000080	1.540	0.000010	0.002700
	B	Dic-02	72	4.66	33069	24750	71.67	28.50	0.720	0.000090	1.260	0.000020	0.001900

	B	Abr-03	77	5.48	10553	25575	11.20	24.90	0.650	0.000070	1.320	0.000020	0.001500
	B	Jun-03	81	7.01	12393	17940	35.17	23.90	0.720	0.000060	1.290	0.000020	0.001700
	B	Jul-03	68	5.80	10274	26970	9.67	22.00	0.650	0.000030	0.940	0.000020	0.002600
	B	Ago-03	63	6.18	11329	27745	1.11	25.50	0.662	0.000082	1.460	0.000010	0.002500
	B	Sep-03	65	5.95	10202	18380	1.50	23.20	0.580	0.000070	1.300	0.000020	0.002000
	B	Oct-03	66	6.18	10551	19033	12.20	24.40	0.578	0.000760	0.502	0.000150	0.003100
	B	Nov-03	70	5.55	10370	20425	7.78	23.80	0.638	0.000200	1.268	0.000060	0.005000
	B	Dic-03	73	5.71	10501	20525	20.22	24.50	0.542	0.000300	1.308	0.000050	0.010000
	B	Ene-04	79	5.73	10833	17365	7.14	24.10	0.985	0.000500	0.914	0.000050	0.009000
	B	Feb-04	79	6.29	10301	19945	19.10	24.30	0.998	0.000500	1.047	0.000120	0.008000
	B	Mar-04	79	6.22	23079	37595	5.90	41.40	1.793	0.000800	1.322	0.000300	0.086000
	B	Mar-04	80	6.12	10730	21520	34.40	24.10	1.559	0.000900	2.020	0.000100	0.038000
	B	Abr-04	76	5.90	10943	22675	3.49	23.80	1.101	0.000800	1.915	0.000060	0.011000
	B	May-04	68	5.72	10460	24705	2.62	24.40	0.472	0.000600	1.010	0.000010	0.009000
	B	Jun-04	68	5.98	10161	22465	3.06	25.20	1.016	0.000700	2.801	0.000010	0.010000
	B	Ago-04	67	5.62	10686	19075	12.44	27.60	1.006	0.000600	2.831	0.000050	0.009000
	B	Sep-04	69	6.26	10677	19175	4.65	28.00	0.601	0.000700	2.031	0.000160	0.008000
	B	Nov-04	71	6.31	10840	18985	12.59	28.10	0.458	0.000400	1.488	0.000030	0.007000
	B	Nov-04	71	5.95	10827	21490	7.92	26.90	0.499	0.000400	1.438	0.000020	0.009000
	B	Dic-04	77	5.91	10793	20815	14.78	25.20	0.556	0.000500	1.738	0.000030	0.008000
	B	Feb-05	91	7.05	10299	23605	7.00	30.70	0.466	0.000700	1.396	0.000070	0.010000
	B	Mar-05	55	6.94	9585	26020	10.67	30.20	0.535	0.000600	1.392	0.000060	0.014000
Paniñas entrada	C	Jul-98	68	6.82	*	1410	1.00	4.60	0.003	0.020000	0.020	0.002000	0.110000
	C	Ago-98	72	7.00	781	3950	1.00	5.00	0.060	0.020000	0.002	0.002000	0.040000
	C	Set-98	69	7.63	61	3250	2.00	4.80	0.190	0.020000	0.020	0.002000	0.050000
	C	Oct-98	69	8.28	765	2040	1.00	3.30	0.005	0.000300	0.005	0.000100	0.002000
	C	Nov-98	74	8.19	997	2420	1.00	3.50	0.007	0.001100	0.008	0.000100	0.002000
	C	Mar-99	78	7.38	565	2658	1.00	2.40	0.017	0.000100	0.003	0.000000	0.000040
	C	Jun-99	88	8.48	2114	2180	2.00	3.10	1.492	0.000100	0.032	0.000100	0.009000
	C	Set-99	68	7.24	877	2770	3.00	3.90	0.180	0.000030	0.054	0.000010	0.006000
	C	Dic-99	77	7.20	680	25560	37.84	16.00	0.063	29.470000	0.030	0.003000	0.002000
	C	Abr-00	79	7.00	560	975	33.62	19.00	0.051	6.850000	0.120	0.040000	0.002000
	C	Jul-00	74	6.90	602	22141	38.45	6.00	0.058	1.868000	0.002	0.112000	0.000200
	C	Oct-00	67	7.40	613	21501	36.61	11.00	0.048	0.363000	0.001	0.036000	0.000030
	C	Dic-00	77	7.50	687	14226	24.83	12.00	0.031	1.712000	0.000	0.024000	0.000050
	C	Abr-01	81	7.50	687	16069	29.12	17.00	0.039	1.732000	0.000	0.060000	0.000040
	C	Jul-01	86	6.80	628	16205	28.48	3.00	0.045	2.876000	0.000	0.257000	0.000010
	C	Oct-01	84	7.70	657	17128	28.92	28.00	0.048	1.140000	0.000	2.412000	0.000020
	C	Dic-01	86	7.90	662	12496	21.21	12.45	0.036	2.003000	0.000	3.226000	0.000020
	C	Mar-02	79	7.50	704	22316	39.15	19.21	0.055	1.538000	0.000	1.685000	0.000180
	C	Jun-02	79	6.70	609	22563	26.86	4.00	0.038	1.100000	0.000	1.140000	0.000050
	C	Set-02	66	7.70	579	20329	36.06	9.45	0.029	1.557000	0.001	1.259000	0.000150
	C	Dic-02	72	8.10	562	14675	27.74	38.67	0.034	1.207000	0.002	0.918000	0.000110
	C	Abr-03	78	7.70	608	20812	36.71	13.00	0.049	2.710000	0.000	1.636000	0.000040
	C	Jun-03	86	6.40	622	17274	29.43	1.00	0.041	1.796000	0.000	1.336000	0.000100
	C	Jul-03	68	7.90	624	10110	21.38	16.72	0.029	0.354000	0.001	1.466000	0.000020
	C	Ago-03	63	8.33	755	2545	2.33	2.83	0.015	0.001600	0.030	0.000020	0.006000
	C	Sep-03	65	7.94	680	2115	0.00	2.60	0.018	0.001300	0.040	0.000020	0.007000
	C	Oct-03	66	8.03	948	2650	0.48	3.20	0.007	0.000560	0.028	0.007000	0.000500
	C	Nov-03	70	7.79	764	2575	5.26	2.40	0.043	0.000200	0.168	0.000710	0.002000
	C	Dic-03	73	7.54	778	2545	8.87	2.90	0.026	0.000700	0.170	0.000040	0.010000
	C	Ene-04	79	7.40	800	2200	4.09	2.90	0.096	0.000400	0.031	0.003510	0.009000
	C	Feb-04	79	8.12	790	3280	6.26	2.70	0.065	0.000700	0.177	0.002840	0.003000
	C	Mar-04	79	8.17	765	2555	0.83	2.90	0.045	0.000300	0.182	0.001200	0.008000
	C	Mar-04	80	8.03	764	2230	3.18	2.80	0.053	0.000500	0.217	0.001700	0.010000
	C	Abr-04	76	8.02	775	2560	2.27	2.80	0.057	0.000600	0.233	0.001400	0.009000
	C	May-04	68	7.91	771	2635	0.33	2.90	0.038	0.000500	0.193	0.000780	0.008000
	C	Jun-04	68	7.96	734	2210	1.74	3.15	0.058	0.000500	0.323	0.001360	0.004000
	C	Ago-04	67	8.26	781	2165	0.18	2.73	0.077	0.000200	0.297	0.000880	0.005000
	C	Sep-04	69	8.42	766	2485	0.18	2.96	0.027	0.000400	0.218	0.001730	0.007000
	C	Nov-04	71	8.04	762	2185	1.52	2.70	0.028	0.000300	0.170	0.000640	0.006000
	C	Nov-04	71	8.28	761	2115	0.42	3.14	0.021	0.000300	0.151	0.000230	0.007000
	C	Dic-04	77	8.10	910	2570	1.89	3.87	0.029	0.000600	0.205	0.000070	0.010000
	C	Feb-05	94	8.75	780	2090	2.00	3.20	0.018	0.000600	0.139	0.000130	0.009000
	C	Mar-05	98	8.54	735	2455	3.33	3.43	0.028	0.000700	0.154	0.000100	0.007000
Batería 175	D	Jul-98	68	6.71	*	38912	18.00	61.40	0.140	0.180000	0.200	0.002000	1.080000

	D	Ago-98	72	6.00	19880	30960	20.00	52.80	17.120	0.200000	0.002	0.002000	0.080000
	D	Set-98	69	6.62	1058	36050	53.00	57.30	8.640	0.180000	0.040	0.002000	1.140000
	D	Oct-98	69	6.20	22141	36920	14.00	58.50	2.689	0.003200	0.077	0.001600	0.001700
	D	Nov-98	74	8.66	21750	36960	17.00	48.90	0.255	0.001400	0.073	0.000100	0.008000
	D	Mar-99	78	6.41	23013	40602	18.00	44.00	7.239	0.000060	0.080	0.000060	0.004000
	D	Jun-99	75	6.13	17126	25540	21.00	35.50	1.804	0.000050	0.092	0.000030	0.008000
	D	Set-99	68	6.09	17031	29640	24.00	46.70	1.492	0.000030	0.254	0.000010	0.002000
	D	Dic-99	77	6.77	17560	29740	25.00	49.50	3.384	0.000120	2.254	0.000030	0.008000
	D	Abr-00	79	6.76	20881	33890	22.59	53.80	1.957	0.000040	3.166	0.000010	0.033000
	D	Jul-00	74	7.12	20212	38870	47.80	54.60	2.104	0.000020	1.692	0.000060	0.001000
	D	Oct-00	66	6.94	26331	27035	2.00	38.80	1.112	0.000040	1.156	0.000080	0.007000
	D	Dic-00	77	5.15	30038	49750	3.97	41.60	1.924	0.000320	1.328	0.000150	0.001000
	D	Abr-01	81	5.74	15687	28540	11.67	36.50	1.146	0.001750	0.995	0.000080	0.021000
	D	Jul-01	79	8.47	21952	41425	2.48	58.50	3.330	0.000150	1.708	0.000030	0.001000
	D	Oct-01	64	6.01	17359	30075	1.00	42.60	1.812	0.000150	1.382	0.000050	0.001000
	D	Dic-01	76	6.02	17481	33880	16.72	47.60	1.434	0.000910	1.446	0.000020	0.001000
	D	Mar-02	79	6.92	17252	28715	4.50	43.10	1.036	0.000170	1.014	0.000120	0.004000
	D	Jun-02	76	6.50	20059	35337	6.10	49.10	1.012	0.000140	0.904	0.000030	0.005000
	D	Set-02	66	6.38	17613	27010	38.33	39.10	1.165	0.001900	1.030	0.000050	0.016000
	D	Dic-02	72	7.37	17928	28875	13.33	38.30	1.050	0.001500	0.980	0.000040	0.018000
	D	Abr-03	78	4.77	19207	36015	22.83	40.20	1.000	0.001100	0.860	0.000030	0.016000
	D	Jun-03	84	6.79	13471	25630	28.33	31.10	1.050	0.000900	0.880	0.000020	0.015000
	D	Jul-03	68	6.25	20584	36995	8.50	38.90	0.920	0.000600	0.620	0.000010	0.011000
	D	Ago-03	63	6.39	14350	26465	2.50	32.80	1.090	0.002000	1.100	0.000050	0.018000
	D	Sep-03	65	4.71	17411	32615	2.17	40.90	1.010	0.001800	0.990	0.000030	0.020000
	D	Oct-03	66	6.51	13782	24630	52.20	31.80	0.873	0.000280	0.460	0.000250	0.001000
	D	Nov-03	70	6.08	16978	29550	14.63	36.50	1.185	0.000100	0.908	0.000080	0.008000
	D	Dic-03	73	5.95	15657	28586	22.84	36.00	0.932	0.000800	1.142	0.000180	0.009000
	D	Ene-04	79	6.63	16354	25035	4.64	35.00	1.303	0.000700	0.646	0.000180	0.008000
	D	Feb-04	79	6.67	12740	23815	13.14	29.40	0.641	0.000300	1.248	0.000140	0.016000
	D	Mar-04	79	6.94	15655	30585	5.00	32.50	1.458	0.000500	1.183	0.000330	0.060000
	D	Mar-04	80	6.39	14078	26160	8.18	30.40	1.520	0.000600	1.594	0.000080	0.042000
	D	Abr-04	76	5.58	22255	40305	3.62	45.90	1.766	0.000600	1.658	0.000070	0.009000
	D	May-04	68	6.04	12498	27425	2.25	41.60	0.930	0.000700	1.134	0.000020	0.008000
	D	Jun-04	68	6.44	19292	37920	3.64	42.00	1.813	0.000800	2.055	0.000020	0.009000
	D	Ago-04	87	5.82	21904	37510	12.50	55.00	2.051	0.000600	2.409	0.000300	0.008000
	D	Sep-04	69	5.96	23108	39785	5.33	55.70	1.487	0.000600	1.792	0.000180	0.009000
	D	Nov-04	71	6.64	20583	36995	13.23	50.70	1.034	0.000500	1.248	0.000050	0.007000
	D	Nov-04	71	6.22	19372	35115	7.90	47.40	1.054	0.000500	1.295	0.000030	0.010000
	D	Dic-04	77	6.78	18618	32740	5.88	47.40	1.165	0.000600	1.339	0.000030	0.010000
	D	Feb-05	89	7.36	16632	29550	10.00	59.50	0.974	0.000700	1.116	0.000050	0.009000
	D	Mar-05	52	7.04	18579	39845	11.17	55.00	1.193	0.000600	1.166	0.000060	0.007000
Poza Api 401	E	Jul-98	68	6.74	*	19314	17.00	35.40	0.020	0.120000	0.120	0.002000	0.460000
	E	Ago-98	72	6.80	17040	24780	16.00	42.00	30.470	0.010000	0.002	0.002000	0.010000
	E	Set-98	70	6.85	701	24730	31.00	39.70	10.910	0.190000	0.030	0.002000	1.040000
	E	Oct-98	69	6.48	14550	25520	6.00	36.90	3.732	0.001800	0.025	0.000100	0.007000
	E	Nov-98	74	7.17	14957	26310	9.00	35.60	0.313	0.001000	0.026	0.000070	0.005000
	E	Mar-99	78	8.30	13389	24768	16.00	29.00	4.497	0.000100	0.047	0.000020	0.007000
	E	Jun-99	75	6.79	14589	22890	16.00	29.70	2.460	0.000030	0.070	0.000030	0.010000
	E	Set-99	68	6.88	11947	22030	3.00	33.80	1.668	0.000030	0.187	0.000010	0.005000
	E	Dic-99	77	6.77	13847	23470	33.00	39.00	4.440	0.000300	1.788	0.000030	0.005000
	E	Abr-00	79	6.95	13573	21960	17.04	36.10	0.564	0.000040	2.024	0.000010	0.005000
	E	Jul-00	73	6.75	21340	38800	24.45	54.60	2.788	0.000020	1.284	0.000030	0.001000
	E	Oct-00	66	6.87	13616	25645	9.00	34.20	1.992	0.000010	0.956	0.000060	0.027000
	E	Dic-00	77	8.87	10299	22990	72.88	28.00	0.234	0.000400	0.960	0.000050	0.001000
	E	Abr-01	81	6.43	9740	18060	44.40	23.50	1.179	0.001620	0.317	0.000050	0.000170
	E	Jul-01	81	6.50	14184	27945	71.00	35.80	1.975	0.000460	1.510	0.000050	0.008000
	E	Oct-01	64	6.45	13861	24515	32.14	33.30	1.864	0.000150	1.140	0.000050	0.002000
	E	Dic-01	84	6.77	11121	18515	29.61	24.70	0.201	0.001100	0.580	0.000020	0.005000
	E	Mar-02	79	7.18	9945	19265	5.40	26.20	1.984	0.000180	0.860	0.000030	0.007000
	E	Jun-02	78	6.70	14352	26135	20.20	33.10	1.840	0.000200	0.780	0.000020	0.005000
	E	Set-02	66	7.39	13410	23690	79.00	31.50	1.900	0.000170	1.120	0.000040	0.003000
	E	Dic-02	73	6.60	15537	26540	342.00	31.70	1.450	0.000120	1.050	0.000030	0.005000
	E	Abr-03	78	6.84	16498	31830	77.00	35.30	1.760	0.000190	1.110	0.000030	0.007000
	E	Jun-03	84	7.42	13516	21095	69.33	29.80	1.660	0.000140	0.970	0.000020	0.009000
	E	Jul-03	68	6.73	2433	5590	25.17	6.10	1.520	0.000100	0.860	0.000020	0.006000

	E	Ago-03	63	6.39	12290	27525	1.83	27.30	1.600	0.000150	1.050	0.000040	0.007300
	E	Sep-03	65	5.94	9250	18390	1.33	23.00	1.500	0.000130	0.950	0.000030	0.002000
	E	Oct-03	66	6.50	15263	24306	10.80	31.60	1.220	0.001150	0.274	0.000050	0.002100
	E	Nov-03	70	5.62	16894	28315	6.09	31.90	1.302	0.000400	0.937	0.000130	0.003000
	E	Dic-03	73	5.62	10425	19750	15.40	24.70	0.547	0.000300	1.314	0.000080	0.007000
	E	Ene-04	79	6.52	14877	23780	31.77	32.90	1.683	0.000300	0.948	0.000040	0.009000
	E	Feb-04	79	6.96	6126	12450	5.82	15.00	0.981	0.000800	0.911	0.000110	0.008000
	E	Mar-04	79	5.84	10423	21775	5.41	23.70	0.795	0.000500	1.499	0.000250	0.087000
	E	Mar-04	80	8.37	13999	25020	7.09	30.20	1.583	0.000700	1.632	0.000080	0.015000
	E	Abr-04	76	7.28	14904	28260	2.37	31.80	2.317	0.000700	1.217	0.000030	0.009000
	E	May-04	68	8.30	12391	23665	4.45	29.20	0.832	0.000800	1.202	0.000020	0.009000
	E	Jun-04	68	7.14	8847	17565	9.70	21.60	0.330	0.000700	1.178	0.000010	0.012000
	E	Ago-04	66	5.68	10650	19015	13.37	28.00	0.926	0.000600	2.692	0.000020	0.010000
	E	Sep-04	69	6.64	14339	25465	21.20	31.60	1.715	0.000600	1.201	0.000070	0.009000
	E	Nov-04	71	6.64	12709	21900	4.35	30.40	0.563	0.000700	1.349	0.000020	0.008000
	E	Nov-04	71	5.98	19553	35845	12.22	48.20	0.995	0.000700	1.212	0.000030	0.008000
	E	Dic-04	77	6.65	14220	25590	10.35	37.20	1.551	0.000600	1.097	0.000020	0.011000
	E	Feb-05	90	7.32	14565	27150	6.00	42.00	1.924	0.000600	0.841	0.000050	0.008000
	E	Mar-05	49.5	6.46	11207	26940	29.17	35.20	0.552	0.000500	1.058	0.000050	0.009000
Pañas salida	F	Jul-98	68	6.78	*	1033	2.00	3.00	0.004	0.020000	0.060	0.002000	0.100000
	F	Ago-98	72	6.82	958	2350	2.00	4.70	0.027	0.010000	0.002	0.002000	0.030000
	F	Set-98	69	7.93	61	2390	2.00	3.40	0.040	0.040000	0.002	0.002000	0.270000
	F	Oct-98	69	8.14	803	2280	1.00	3.60	0.005	0.000300	0.007	0.000200	0.005400
	F	Nov-98	74	8.17	935	2490	1.00	3.40	0.005	0.000100	0.005	0.000050	0.002000
	F	Mar-99	78	7.47	586	2048	1.00	2.40	0.036	0.000100	0.004	0.000030	0.009000
	F	Jun-99	88	8.55	1902	2130	2.00	3.10	0.028	0.000020	0.003	0.000040	0.012000
	F	Set-99	68	8.21	852	2640	2.00	4.20	0.180	0.000030	0.054	0.000020	0.006000
	F	Dic-99	77	7.69	1033	2680	2.00	4.10	0.091	0.000020	0.440	0.000020	0.020000
	F	Abr-00	79	8.09	1044	2540	3.64	4.90	0.007	0.000040	0.612	0.000100	0.005000
	F	Jul-00	73	7.17	482	909	8.07	2.20	0.003	0.000200	0.328	0.000070	0.002000
	F	Oct-00	66	7.30	1165	3240	0.40	4.20	0.049	0.000010	0.204	0.000070	0.012000
	F	Dic-00	77	7.77	1180	5370	1.18	15.40	0.019	0.000150	0.332	0.000200	0.003000
	F	Abr-01	81	7.39	366	920	3.50	1.00	0.022	0.001660	0.036	0.000350	0.009000
	F	Jul-01	84	7.46	844	2630	3.00	4.10	0.017	0.000100	0.212	0.000080	0.002000
	F	Oct-01	64	7.27	725	2255	1.00	2.40	0.017	0.000120	0.247	0.000030	0.001000
	F	Dic-01	79	7.65	843	2825	4.18	3.33	0.026	0.000270	0.246	0.000100	0.001000
	F	Mar-02	79	7.26	311	1335	5.50	1.67	0.018	0.000410	0.109	0.000400	0.009000
	F	Jun-02	77	7.00	868	2690	3.50	3.90	0.021	0.000350	0.214	0.000040	0.007000
	F	Set-02	68	8.51	731	1690	4.17	2.70	0.027	0.001300	0.040	0.000060	0.008000
	F	Dic-02	72	7.80	806	2260	5.67	2.90	0.020	0.003000	0.027	0.000020	0.009000
	F	Abr-03	79	7.91	1020	2915	6.60	3.40	0.015	0.004000	0.039	0.000020	0.012000
	F	Jun-03	84	8.22	971	2335	5.00	3.30	0.020	0.003000	0.046	0.000020	0.011000
	F	Jul-03	69	7.03	945	3290	8.67	2.90	0.018	0.002100	0.033	0.000010	0.009000
	F	Ago-03	63	8.33	755	2545	2.33	2.83	0.015	0.001600	0.030	0.000020	0.006000
	F	Sep-03	65	7.94	680	2115	0.00	2.60	0.018	0.001300	0.040	0.000020	0.007000
	F	Oct-03	66	8.03	948	2650	0.48	3.20	0.007	0.000560	0.028	0.007000	0.000500
	F	Nov-03	70	7.79	764	2575	5.26	2.40	0.043	0.000200	0.168	0.000710	0.002000
	F	Dic-03	73	7.54	778	2545	8.87	2.90	0.026	0.000700	0.170	0.000040	0.010000
	F	Ene-04	79	7.40	800	2200	4.09	2.90	0.098	0.000400	0.031	0.003510	0.009000
	F	Feb-04	78	8.12	790	3280	6.26	2.70	0.085	0.000700	0.177	0.002840	0.003000
	F	Mar-04	79	8.17	785	2555	0.83	2.90	0.045	0.000300	0.182	0.001200	0.008000
	F	Mar-04	80	8.26	948	2650	2.08	3.40	0.054	0.000500	0.263	0.001500	0.011000
	F	Abr-04	76	8.25	954	3025	1.76	3.30	0.058	0.000600	0.253	0.001100	0.010000
	F	May-04	68	7.46	958	3125	0.22	3.40	0.043	0.000600	0.201	0.000970	0.007000
	F	Jun-04	68	8.25	923	2670	1.20	3.46	0.081	0.000500	0.368	0.000910	0.005000
	F	Ago-04	66	8.34	923	2535	0.71	3.52	0.050	0.000300	0.308	0.000730	0.005000
	F	Sep-04	69	8.46	992	2790	0.60	3.80	0.029	0.000400	0.245	0.001270	0.006000
	F	Nov-04	71	8.09	1024	755	0.98	3.80	0.038	0.000400	0.192	0.000290	0.006000
	F	Nov-04	71	8.07	992	2785	0.79	4.07	0.025	0.000400	0.181	0.000220	0.007000
	F	Dic-04	77	7.87	726	2130	1.05	3.23	0.025	0.000600	0.175	0.000100	0.010000
	F	Feb-05	93	8.33	1045	2950	2.00	3.70	0.026	0.000600	0.192	0.000080	0.010000
	F	Mar-05	51	8.24	941	2975	4.33	5.01	0.025	0.000600	0.130	0.000090	0.006000

ANEXO 4
APLICACIÓN DEL SOFTWARE SPSS PARA EL ANÁLISIS CLASIFICATORIO

Casewise Statistics											
	Case Number	Actual Group	Highest Group					Second Highest Group			Discriminant Scores
			Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
				p	df						
Original	1	1	1	.411	1	1.000	.677	2	.000	22.192	1.952
	2	1	1	.500	1	1.000	.455	2	.000	20.817	1.804
	3	1	1	.115	1	1.000	2.478	2	.000	29.836	2.704
	4	1	1	.109	1	1.000	2.572	2	.000	30.159	2.733
	5	1	1	.188	1	1.000	1.736	2	.000	27.099	2.447
	6	1	1	.445	1	.990	.584	2	.010	9.759	.365
	7	1	1	.934	1	.999	.007	2	.001	14.484	1.047
	8	1	1	.807	1	1.000	.060	2	.000	17.078	1.374
	9	1	1	.981	1	1.000	.001	2	.000	15.299	1.153
	10	1	1	.402	1	.987	.702	2	.013	9.305	.292
	11	1	1	.574	1	1.000	.316	2	.000	19.806	1.692
	12	1	1	.779	1	1.000	.079	2	.000	17.383	1.411
	13	1	1	.096	1	1.000	2.771	2	.000	30.831	2.794
	14	1	1	.316	1	1.000	1.003	2	.000	23.910	2.131
	15	1	1	.160	1	1.000	1.973	2	.000	28.014	2.534
	16	1	1	.699	1	1.000	.149	2	.000	18.269	1.516
	17	1	1	.697	1	.998	.152	2	.002	12.241	.740
	18	1	1	.098	1	.754	2.741	2	.246	4.984	-.526
	19	1	1	.520	1	1.000	.414	2	.000	20.536	1.773
	20	1	1	.765	1	1.000	.090	2	.000	17.536	1.429

	21	1	1	.876	1	.999	.024	2	.001	13.925	.973
	22	1	1	.207	1	1.000	1.591	2	.000	26.516	2.391
	23	1	1	.443	1	.990	.588	2	.010	9.742	.363
	24	1	1	.146	1	.870	2.117	2	.130	5.919	-.325
	25	1	1	.859	1	.999	.031	2	.001	13.769	.952
	26	1	1	.166	1	1.000	1.914	2	.000	27.791	2.513
	27	1	1	.653	1	.997	.202	2	.003	11.828	.681
	28	1	1	.552	1	1.000	.354	2	.000	20.097	1.724
	29	1	1	.507	1	1.000	.440	2	.000	20.719	1.793
	30	1	1	.655	1	.997	.200	2	.003	11.843	.683
	31	1	1	.537	1	.994	.382	2	.006	10.693	.512
	32	1	1	.837	1	.999	.042	2	.001	13.558	.924
	33	1	1	.964	1	.999	.002	2	.001	14.769	1.085
	34	1	1	.019	1	1.000	5.507	2	.000	38.872	3.476
	35	1	1	.688	1	1.000	.161	2	.000	18.398	1.531
	36	1	1	.289	1	1.000	1.126	2	.000	24.495	2.191
	37	1	1	.087	1	1.000	2.937	2	.000	31.380	2.843
	38	1	1	.070	1	1.000	3.288	2	.000	32.506	2.943
	39	1	1	.447	1	1.000	.577	2	.000	21.604	1.890
	40	1	1	.300	1	1.000	1.075	2	.000	24.253	2.166
	41	1	1	.852	1	1.000	.035	2	.000	16.607	1.317
	42	1	1	.281	1	.967	1.164	2	.033	7.892	.051
	43	1	1	.931	1	1.000	.007	2	.000	15.794	1.216
	44	1	1	.126	1	.834	2.337	2	.166	5.567	-.399
	45	1	1	.151	1	.879	2.059	2	.121	6.019	-.305

	46	1	1	.629	1	.997	.233	2	.003	11.596	.647
	47	1	1	.782	1	.998	.077	2	.002	13.038	.852
	48	1	1	.633	1	.997	.228	2	.003	11.629	.652
	49	1	1	.436	1	.989	.607	2	.011	9.668	.351
	50	1	1	.374	1	.984	.789	2	.016	9.000	.241
	51	1	1	.849	1	.999	.036	2	.001	13.672	.939
	52	1	1	.774	1	1.000	.082	2	.000	17.430	1.416
	53	1	1	.373	1	.984	.794	2	.016	8.983	.239
	54	1	1	.825	1	1.000	.049	2	.000	16.884	1.351
	55	1	1	.101	1	.767	2.683	2	.233	5.063	-.508
	56	1	1	.344	1	1.000	.896	2	.000	23.374	2.076
	57	1	1	.496	1	.993	.463	2	.007	10.287	.449
	58	1	1	.435	1	.989	.609	2	.011	9.657	.349
	59	1	1	.558	1	.995	.343	2	.005	10.906	.544
	60	1	1	.877	1	.999	.024	2	.001	13.934	.974
	61	1	1	.677	1	1.000	.174	2	.000	18.531	1.546
	62	1	1	.340	1	.979	.909	2	.021	8.614	.176
	63	1	1	.714	1	.998	.135	2	.002	12.399	.763
	64	1	1	.087	1	1.000	2.932	2	.000	31.365	2.842
	65	1	1	.366	1	1.000	.817	2	.000	22.966	2.034
	66	1	1	.107	1	.784	2.599	2	.216	5.179	-.483
	67	1	1	.530	1	1.000	.395	2	.000	20.399	1.758
	68	1	1	.806	1	1.000	.060	2	.000	17.090	1.376
	69	1	1	.636	1	.997	.225	2	.003	11.656	.656
	70	1	1	.511	1	.993	.433	2	.007	10.434	.472

	71	1	1	.820	1	1.000	.052	2	.000	16.939	1.357
	72	1	1	.937	1	1.000	.006	2	.000	15.740	1.209
	73	1	1	.810	1	.999	.058	2	.001	13.308	.890
	74	1	1	.550	1	.995	.357	2	.005	10.826	.532
	75	1	1	.176	1	1.000	1.833	2	.000	27.479	2.484
	76	1	1	.817	1	.999	.053	2	.001	13.373	.898
	77	1	1	.910	1	1.000	.013	2	.000	16.014	1.243
	78	1	1	.635	1	1.000	.226	2	.000	19.040	1.605
	79	1	1	.955	1	.999	.003	2	.001	14.678	1.073
	80	1	1	.988	1	.999	.000	2	.001	14.999	1.114
	81	1	1	.476	1	.992	.509	2	.008	10.080	.416
	82	1	1	.437	1	.989	.603	2	.011	9.682	.353
	83	1	1	.918	1	.999	.011	2	.001	14.329	1.027
	84	1	1	.927	1	.999	.008	2	.001	14.413	1.038
	85	1	1	.322	1	.976	.980	2	.024	8.400	.140
	86	1	1	.394	1	.986	.727	2	.014	9.213	.277
	87	2	2	.517	1	.994	.421	1	.006	10.495	-2.110
	88	2	2	.459	1	.991	.549	1	.009	9.904	-2.017
	89	2	2	.946	1	1.000	.005	1	.000	15.649	-2.826
	90	2	2	.351	1	1.000	.870	1	.000	23.243	-3.691
	91	2	2	.449	1	1.000	.573	1	.000	21.575	-3.515
	92	2	2	.831	1	.999	.046	1	.001	13.500	-2.545
	93	2	2	.310	1	1.000	1.032	1	.000	24.051	-3.775
	94	2	2	.751	1	.998	.100	1	.002	12.755	-2.442
	95	2	1(**)	.295	1	.970	1.097	2	.030	8.070	.082

	96	2	2	.636	1	.997	.224	1	.003	11.664	-2.286
	97	2	1(**)	.276	1	.965	1.189	2	.035	7.828	.039
	98	2	1(**)	.076	1	.661	3.142	2	.339	4.476	-.643
	99	2	2	.194	1	.925	1.686	1	.075	6.707	-1.460
	100	2	2	.128	1	.838	2.313	1	.162	5.604	-1.238
	101	2	1(**)	.127	1	.836	2.327	2	.164	5.583	-.396
	102	2	2	.197	1	.927	1.668	1	.073	6.743	-1.467
	103	2	2	.470	1	.991	.522	1	.009	10.023	-2.036
	104	2	1(**)	.086	1	.705	2.956	2	.295	4.703	-.590
	105	2	1(**)	.433	1	.989	.615	2	.011	9.634	.345
	106	2	2	.100	1	.761	2.709	1	.239	5.028	-1.113
	107	2	2	.527	1	.994	.399	1	.006	10.603	-2.127
	108	2	2	.076	1	.657	3.158	1	.343	4.457	-.981
	109	2	1(**)	.338	1	.979	.919	2	.021	8.583	.171
	110	2	2	.728	1	.998	.121	1	.002	12.538	-2.411
	111	2	2	.329	1	1.000	.954	1	.000	23.667	-3.735
	112	2	2	.571	1	1.000	.320	1	.000	19.839	-3.324
	113	2	2	.543	1	1.000	.370	1	.000	20.220	-3.367
	114	2	2	.754	1	1.000	.099	1	.000	17.657	-3.072
	115	2	2	.989	1	1.000	.000	1	.000	15.222	-2.772
	116	2	2	.884	1	.999	.021	1	.001	14.003	-2.612
	117	2	2	.581	1	1.000	.305	1	.000	19.717	-3.311
	118	2	2	.493	1	1.000	.470	1	.000	20.921	-3.444
	119	2	2	.579	1	1.000	.308	1	.000	19.740	-3.313
	120	2	2	.592	1	1.000	.287	1	.000	19.569	-3.294

	121	2	2	.645	1	1.000	.212	1	.000	18.911	-3.219
	122	2	2	.574	1	1.000	.316	1	.000	19.806	-3.321
	123	2	2	.361	1	1.000	.835	1	.000	23.060	-3.672
	124	2	2	.296	1	1.000	1.091	1	.000	24.330	-3.803
	125	2	2	.527	1	1.000	.399	1	.000	20.431	-3.390
	126	2	2	.362	1	1.000	.831	1	.000	23.036	-3.670
	127	2	2	.534	1	1.000	.387	1	.000	20.340	-3.380
	128	2	2	.197	1	1.000	1.667	1	.000	26.824	-4.050
	129	2	2	.196	1	1.000	1.670	1	.000	26.835	-4.051
	130	1	1	.388	1	1.000	.746	2	.000	22.579	1.993
	131	1	1	.394	1	1.000	.726	2	.000	22.469	1.982
	132	1	1	.504	1	1.000	.447	2	.000	20.766	1.798
	133	1	1	.216	1	1.000	1.531	2	.000	26.269	2.367
	134	1	1	.451	1	1.000	.568	2	.000	21.548	1.884
	135	1	1	.147	1	1.000	2.105	2	.000	28.504	2.580
	136	1	1	.881	1	1.000	.022	2	.000	16.306	1.280
	137	1	1	.560	1	1.000	.340	2	.000	19.990	1.713
	138	1	1	.907	1	.999	.014	2	.001	14.226	1.013
	139	1	1	.727	1	1.000	.122	2	.000	17.951	1.478
	140	1	1	.664	1	1.000	.189	2	.000	18.689	1.565
	141	1	1	.502	1	.993	.450	2	.007	10.352	.459
	142	1	1	.000	1	1.000	14.768	2	.000	59.768	4.973
	143	1	1	.342	1	1.000	.904	2	.000	23.417	2.081
	144	1	1	.140	1	1.000	2.181	2	.000	28.783	2.606
	145	1	1	.489	1	1.000	.480	2	.000	20.983	1.822

	171	1	1	.469	1	.991	.524	2	.009	10.011	.406
	172	1	1	.631	1	1.000	.230	2	.000	19.081	1.610
	173	1	1	.211	1	.937	1.561	2	.063	6.962	-.120
	174	1	1	.477	1	.992	.505	2	.008	10.097	.419
	175	1	1	.432	1	.989	.618	2	.011	9.623	.344
	176	1	1	.772	1	.998	.084	2	.002	12.950	.840
	177	1	1	.340	1	.979	.910	2	.021	8.608	.176
	178	1	1	.917	1	.999	.011	2	.001	14.317	1.025
	179	1	1	.379	1	.984	.773	2	.016	9.053	.250
	180	1	1	.263	1	.961	1.255	2	.039	7.662	.010
	181	1	1	.434	1	.989	.611	2	.011	9.650	.348
	182	1	1	.258	1	.959	1.282	2	.041	7.595	-.003
	183	1	1	.401	1	1.000	.706	2	.000	22.358	1.970
	184	1	1	.460	1	.991	.546	2	.009	9.918	.391
	185	1	1	.344	1	.980	.895	2	.020	8.657	.184
	186	1	1	.346	1	.980	.889	2	.020	8.674	.187
	187	1	1	.979	1	1.000	.001	2	.000	15.323	1.156
	188	1	1	.692	1	.998	.157	2	.002	12.194	.734
	189	1	1	.208	1	.935	1.587	2	.065	6.908	-.130
	190	1	1	.093	1	.736	2.823	2	.264	4.874	-.551
	191	1	1	.680	1	.997	.170	2	.003	12.078	.717
	192	1	1	.125	1	.831	2.354	2	.169	5.540	-.405
	193	1	1	.772	1	.998	.084	2	.002	12.944	.839
	194	1	1	.974	1	1.000	.001	2	.000	15.373	1.162
	195	1	1	.085	1	.704	2.964	2	.296	4.694	-.592

	196	1	2(**)	.232	1	.949	1.426	1	.051	7.258	-1.564
	197	1	1	.988	1	.999	.000	2	.001	14.997	1.114
	198	1	1	.644	1	.997	.213	2	.003	11.741	.668
	199	1	1	.642	1	.997	.216	2	.003	11.719	.665
	200	1	1	.324	1	1.000	.971	2	.000	23.750	2.115
	201	1	1	.920	1	1.000	.010	2	.000	15.908	1.230
	202	1	1	.650	1	.997	.206	2	.003	11.794	.676
	203	1	2(**)	.083	1	.692	3.012	1	.308	4.633	-1.023
	204	1	1	.739	1	1.000	.111	2	.000	17.819	1.463
	205	1	1	.886	1	.999	.021	2	.001	14.023	.986
	206	1	1	.391	1	.986	.737	2	.014	9.179	.271
	207	1	1	.770	1	.998	.086	2	.002	12.929	.837
	208	1	1	.059	1	.551	3.579	2	.449	3.985	-.762
	209	1	1	.926	1	.999	.009	2	.001	14.403	1.037
	210	1	1	.634	1	.997	.227	2	.003	11.637	.653
	211	1	1	.400	1	.986	.708	2	.014	9.282	.288
	212	1	1	.166	1	1.000	1.921	2	.000	27.818	2.516
	213	1	1	.672	1	.997	.180	2	.003	12.000	.706
	214	1	1	.354	1	.981	.859	2	.019	8.770	.203
	215	1	1	.804	1	.999	.062	2	.001	13.248	.881
	216	2	2	.513	1	.993	.428	1	.007	10.460	-2.105
	217	2	2	.438	1	.989	.601	1	.011	9.689	-1.983
	218	2	2	.618	1	1.000	.249	1	.000	19.247	-3.258
	219	2	2	.454	1	1.000	.561	1	.000	21.501	-3.507
	220	2	2	.467	1	1.000	.528	1	.000	21.296	-3.485

	221	2	2	.962	1	.999	.002	1	.001	14.753	-2.711
	222	2	2	.271	1	1.000	1.212	1	.000	24.889	-3.859
	223	2	2	.426	1	1.000	.635	1	.000	21.948	-3.555
	224	2	2	.887	1	1.000	.020	1	.000	16.242	-2.900
	225	2	2	.551	1	1.000	.356	1	.000	20.114	-3.355
	226	2	2	.816	1	.999	.054	1	.001	13.363	-2.526
	227	2	2	.776	1	.998	.081	1	.002	12.982	-2.473
	228	2	2	.958	1	.999	.003	1	.001	14.710	-2.706
	229	2	2	.969	1	.999	.001	1	.001	14.819	-2.720
	230	2	2	.872	1	.999	.026	1	.001	13.892	-2.598
	231	2	2	.841	1	.999	.040	1	.001	13.595	-2.557
	232	2	2	.946	1	1.000	.005	1	.000	15.649	-2.826
	233	2	2	.830	1	.999	.046	1	.001	13.492	-2.543
	234	2	2	.521	1	.994	.411	1	.006	10.542	-2.117
	235	2	2	.211	1	1.000	1.562	1	.000	26.398	-4.008
	236	2	2	.730	1	1.000	.119	1	.000	17.920	-3.104
	237	2	2	.724	1	1.000	.125	1	.000	17.986	-3.111
	238	2	2	.465	1	1.000	.534	1	.000	21.333	-3.489
	239	2	2	.537	1	.994	.380	1	.006	10.702	-2.142
	240	2	2	.329	1	1.000	.954	1	.000	23.667	-3.735
	241	2	2	.571	1	1.000	.320	1	.000	19.839	-3.324
	242	2	2	.543	1	1.000	.370	1	.000	20.220	-3.367
	243	2	2	.754	1	1.000	.099	1	.000	17.657	-3.072
	244	2	2	.989	1	1.000	.000	1	.000	15.222	-2.772
	245	2	2	.884	1	.999	.021	1	.001	14.003	-2.612

	246	2	2	.581	1	1.000	.305	1	.000	19.717	-3.311
	247	2	2	.493	1	1.000	.470	1	.000	20.921	-3.444
	248	2	2	.441	1	1.000	.594	1	.000	21.708	-3.529
	249	2	2	.455	1	1.000	.558	1	.000	21.484	-3.505
	250	2	2	.919	1	.999	.010	1	.001	14.338	-2.657
	251	2	2	.400	1	1.000	.709	1	.000	22.377	-3.601
	252	2	2	.333	1	1.000	.939	1	.000	23.590	-3.727
	253	2	2	.290	1	1.000	1.118	1	.000	24.459	-3.816
	254	2	2	.401	1	1.000	.706	1	.000	22.358	-3.599
	255	2	2	.547	1	1.000	.363	1	.000	20.164	-3.361
	256	2	2	.683	1	1.000	.166	1	.000	18.455	-3.166
	257	2	2	.465	1	1.000	.534	1	.000	21.331	-3.489
	258	2	2	.363	1	1.000	.829	1	.000	23.026	-3.669
	259	1	1	.687	1	.998	.162	2	.002	12.146	.727
	260	1	1	.364	1	.983	.824	2	.017	8.883	.222
	261	1	1	.348	1	.980	.880	2	.020	8.703	.192
	262	1	1	.576	1	.995	.312	2	.005	11.086	.571
	263	1	1	.426	1	.989	.635	2	.011	9.556	.333
	264	1	1	.629	1	1.000	.233	2	.000	19.103	1.612
	265	1	1	.574	1	.995	.316	2	.005	11.065	.568
	266	1	1	.784	1	.998	.075	2	.002	13.058	.855
	267	1	1	.870	1	1.000	.027	2	.000	16.416	1.293
	268	1	1	.881	1	.999	.022	2	.001	13.974	.980
	269	1	1	.462	1	1.000	.542	2	.000	21.383	1.866
	270	1	1	.184	1	.916	1.768	2	.084	6.545	-.200

	271	1	1	.820	1	.999	.052	2	.001	13.401	.902
	272	1	1	.967	1	.999	.002	2	.001	14.800	1.089
	273	1	1	.714	1	1.000	.134	2	.000	18.098	1.496
	274	1	1	.661	1	.997	.192	2	.003	11.902	.691
	275	1	1	.515	1	1.000	.424	2	.000	20.607	1.781
	276	1	1	.702	1	.998	.146	2	.002	12.292	.748
	277	1	1	.288	1	.969	1.128	2	.031	7.987	.068
	278	1	1	.700	1	.998	.148	2	.002	12.273	.745
	279	1	1	.336	1	1.000	.927	2	.000	23.533	2.093
	280	1	1	.536	1	.994	.383	2	.006	10.688	.511
	281	1	1	.355	1	.981	.856	2	.019	8.778	.204
	282	1	1	.229	1	1.000	1.449	2	.000	25.926	2.333
	283	1	1	.732	1	.998	.117	2	.002	12.570	.787
	284	1	1	.234	1	.949	1.415	2	.051	7.282	-.060
	285	1	1	.248	1	1.000	1.336	2	.000	25.442	2.286
	286	1	1	.905	1	.999	.014	2	.001	14.207	1.011
	287	1	1	.772	1	1.000	.084	2	.000	17.456	1.420
	288	1	1	.807	1	1.000	.060	2	.000	17.074	1.374
	289	1	1	.853	1	1.000	.034	2	.000	16.592	1.315
	290	1	1	.584	1	1.000	.300	2	.000	19.673	1.677
	291	1	1	.053	1	1.000	3.735	2	.000	33.881	3.062
	292	1	1	.229	1	.947	1.449	2	.053	7.206	-.074
	293	1	1	.205	1	1.000	1.609	2	.000	26.592	2.398
	294	1	1	.623	1	.996	.242	2	.004	11.534	.638
	295	1	1	.100	1	.761	2.708	2	.239	5.029	-.516

	296	1	1	.074	1	.649	3.189	2	.351	4.420	-.656
** Misclassified case											