UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
"Z A R A G O Z A"

COORDENADAS PRINCIPALES, UNA HERRAMIENTA DE ANALISIS ESTADISTICO MULTIVARIADO APLICABLE A ESTUDIOS BIOLOGICOS.

T E S I S

QUE PARA OBTENER EL TITULO DE:

B I O L O G O

P R E S E N T A

GENARO OCHOA DE LA ROSA

U N A M
FES
ZARAGOZA

MEXICO, D. F. NOVIEMBRE 1993

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

RESUMEN

La necesidad de utilizar técnicas adecuadas para realizar análisis estadísticos acordes con los objetivos de trabajo conduce a un incremento en el uso de la estadística multivariada. Dentro de estas técnicas se encuentra coordenadas principales como una herramienta útil para la comparación de entidades.

En este trabajo se hace una revisión del fundamento matemático y de las características del análisis de coordenadas principales, con el propósito de mostrar la técnica de manera simplificada y la forma en que se aplica. Además de sugerir algunos casos en los cuales se puede aplicar y la forma de utilizar el paquete NTSYS-PC para la realización de los cálculos necesarios.

CAPITULO I: INTRODUCCION

Generalmente se considera a la estadística o al análisis estadístico, como un paso de la investigación que permite obtener conclusiones acerca de un determinado caso en estudio llegando al extremo de mencionar que una investigación no es válida si no existe análisis estadístico. Sin embargo, el análisis estadístico debe considerarse una herramienta que puede conducir al diseño de un plan de experimentación, ya que por medio de este es posible identificar las variables de respuesta que intluyen sobre el fenómeno en estudio, además de que permite eliminar redundancia en los datos obtenidos.

En Biología, cuando se investiga un fenómeno generalmente se obtiene una gran cantidad de datos que, además de la problemática inherente a su abundancia, tienen diferencias en cuanto a las unidades en que se expresa cada variable, lo cual no permite un manejo de las diferentes variables a la vez. Por lo general se busca agrupar los datos de acuerdo a las semejanzas que presentan y caracterizan a cada observación; por ejemplo en edafología se pueden determinar las variables pH, color, salinidad, textura, etc., de varios perfiles (entidades). Para lo cual es necesario realizar análisis multivariados que permitan eliminar redundancia en los datos y presenten en forma simplificada su estructura, lo que permite obtener la asociación natural de los datos (Ocegueda, 1991), para tal efecto el análisis de coordenadas principales puede ser útil.

Coordenadas principales es una técnica desarrollada por Gower (1966), que se ha aplicado en varios estudios entre los cuales pueden mencionarse:

Lefkovitch (1976) desarrolla un método de Clasificación Jerárquica para el análisis de coordenadas principales.

Legrende y Chodorosky (1977) dan un ejemplo de la aplicación ecológica de la técnica, a partir de un estudio de similaridad del zooplancton de 20 estanques de las islas del río San Lawrence.

Bates (1978) muestra el uso del análisis de coordenadas principales sobre muestras de vegetación, en el estudio de la influencia de la disponibilidad de metales sobre la vegetación de briofitas de cuatro tipos de rocas de Skye y Rhum.

Adam (1978) realiza un estudio utilizando análisis de cluster y coordenadas principales para determinar las variaciones geográficas en la vegetación de las marismas Británicas.

Huntley and Birks (1978) utilizan la técnica para el estudio sobre el pasado y presente de la vegetación de la reserva natural de Morrone Birkwoods de Escocia.

Williamson (1978), apartir de datos de presencia ausencia, realiza una comparación entre el análisis de componentes principales y la técnica de coordenadas principales.

Gauch (1981) presenta una investigación realizada por **Field & Robb (1970)**, en la cual estudian los moluscos y las lapas de veintiun cuadrantes localizados sobre las rocas costeras de Bahía Falsa en Sudáfrica.

Howard y Howard (1988) utilizando el análisis de coordenadas principales abarcan una mezcla de atributos cuantitativos y cualitativos para 140 hondonadas del distrito ingles Lake.

A pesar de que existen varios estudios que demuestran la utilidad de la técnica, ésta no es de uso generalizado en Biología, esto radica en la complejidad matemática que representa para el investigador; por lo cual surge la necesidad de realizar una descripción de los fundamentos matemáticos de la técnica con el objetivo de fomentar la utilización de esta herramienta. Esta descripción corresponde realizarla a gente que tenga conocimiento del comportamiento de los fenómenos biológicos, de tal manera que comprenda el significado biológico de cada uno de los datos implicados en el estudio y que de esta manera obtenga una mejor interpretación de los resultados obtenidos.

Dado su potencial de aplicación en el estudio y descripción de fenómenos biológicos se realizó el siguiente trabajo planteando como objetivo:

"Analizar los fundamentos teóricos y prácticos de la técnica de coordenadas principales, con el propósito de mostrarla como una herramienta de análisis multivariado aplicable a estudios biológicos."

Para lograr lo anterior se presentan siete Capítulos, incluyendo el presente, que muestran diferentes aspectos de la técnica de análisis de coordenadas principales, cuyo contenido está estructurado de la siguiente manera.

El **Capítulo II** presenta a coordenadas principales como una técnica de ordenación dentro del análisis estadístico multivariado, indicando su potencial de aplicación, así como el tipo de datos que maneja.

En el siguiente capítulo se desglosa el fundamento matemático de la técnica y los requerimientos necesarios para su aplicación, se describe cada uno de los pasos que comprende la técnica y se ejemplifica las operaciones que se realizan utilizando una matriz de datos, también se muestra el gráfico final y su interpretación.

El comportamiento de la técnica al utilizar diferentes medidas de distancia, se muestra en el **Capítulo IV** y auxiliándose con la técnica de cluster para ayudar a interpretar y visualizar mejor los resultados.

El **Capítulo V** comprende la aplicación de coordenadas principales a un caso de estudio biológico, concerniente a "La Taxonomía, Distribución y Descripción de los Suelos del Parque Nacional "El Chico", Hidalgo."

Con el propósito de facilitar los cálculos necesarios para la utilización de la técnica en el **Capítulo VI** se desglosan las instrucciones de uso del paquete NTSYS-PC.

En el **Capítulo VII** se dan las conclusiones del trabajo, mostrando las ventajas de su aplicación, así como algunas de las situaciones en las cuales es recomendable su aplicación.

CAPITULO II: COORDENADAS PRINCIPALES, UNA TECNICA DE ORDENACION EN EL ANALISIS ESTADISTICO MULTIVARIADO.

En consideración a que los datos provenientes de estudios biológicos raramente se ajustan a modelos estadísticos preestablecidos (métodos univariados), en los últimos años se ha tenido un enorme incremento en el desarrollo y aplicación de métodos multivariados de distribución libre en biología.

Dos razones pueden explicar el desarrollo de los métodos multivariados. La primera es indudablemente el incremento en la disponibilidad de recursos computacionales, lo cual hace posible el manejo de grandes matrices de datos típicas de los estudios biológicos. La segunda, quizá menos apreciable, es el cambio en el énfasis de los trabajos teóricos sobre análisis multivariado, contribuyendo contínuamente al desarrollo de modelos estadísticos formales y teorías de distribución de asociaciones, al lado de técnicas descriptivas para la exploración de patrones en el conjunto de datos con el fin de propiciar exposiciones y resúmenes concisos.

Seguramente la importancia de estos métodos consiste en que permite dilucidar las complejas interacciones observadas en las comunidades estudiadas. Para estos datos es de interes identificar similitudes en la respuesta de las entidades, a través de las diferentes variables, pero los procedimientos estándar de estimación estadística y pruebas de hipótesis son inapropiadas, ya que solo permiten el manejo de una o dos variables a la vez. Al realizar un análisis de datos provenientes de una comunidad biológica se tiene que hacer por partes (relacionando una o dos variables al mismo tiempo) y obtener resultados parciales, que después se intentan conjuntar en una conclusión general; en contraste, el análisis multivariado realiza un estudio global de todas las variables, para obtener conclusiones generales que abarca la información proveniente de todas las variables a la vez.

El análisis estadístico multivariado considera un conjunto de i-entidades sobre las cuales se observan o miden j-variables. Las variables pueden ser continuas o discontinuas. También pueden encontrarse mezclas de tipos de variables lo cual resulta en una gran complejidad. Debido a esto, se requiere un estudio que cumpla diferentes propósitos, entre los cuales se pueden mencionar como más importantes los siguientes:

a) Simplificación estructural. El objeto es "ver el bosque a través de los árboles". La explicación del fenómeno bajo estudio con el menor número de transformaciones posibles. Es la vía de representación del complejo bajo estudio.

b) Clasificación. La cuestión es obtener grupos de entidades que se forman de acuerdo a su semejanza.

c) Agrupación de variables. Mientras la clasificación es concerniente con la agrupación de entidades, se puede estar interesado en colocar las variables en grupos conocidos.

d) Análisis de interdependencia. Se busca examinar la interdependencia de variables, las posibles variaciones de independencia a colinelidad.

e) Análisis de dependencia. En el análisis de dependencia una o más variables se simplifican para determinar su dependencia con otras, como en un análisis de regresión.

f) Construcción y pruebas de hipótesis.

Formalmente se puede definir al análisis multivariado como la rama de la estadística a la cual le conciernen las relaciones entre un conjunto de variables dependientes y las entidades que ellas soportan (Kendall, 1982).

El análisis multivariado tiene tres papeles básicos en el estudio de entidades biológicas (Gauch, 1982):

1.- Ayudar a describir la estructura de los datos.

2.- Auxiliar en la elaboración de objetivos, resumiendo los datos, lo cual facilita la comprensión y proporciona un medio para una comunicación efectiva de los resultados.

3.- Contribuir a la generación de hipótesis.

En el análisis multivariado existen varios grupos de técnicas entre las cuales pueden mencionarse: clasificación, análisis de gradientes, ordenación, etc.

Ordenación es el término usado para referirse al conjunto de técnicas multivariadas que intentan, primeramente, representar las relaciones entre muestras y especies tan fielmente como sea posible en el menor espacio dimensional.

El término ordenación deriva de cada uno de los intentos para organizar un grupo de objetos, por ejemplo en el tiempo o a lo largo de un gradiente ambiental. Hoy en día el término es de uso más general, y se refiere al arreglo en un número de dimensiones, preferiblemente pocas, que se aproxime a un patrón de respuesta del conjunto de entidades.

El objetivo usual de la ordenación es auxiliar la generación de hipótesis acerca de las relaciones entre la composición de entidades en un sitio y el subyacente factor ambiental. Además de permitir descubrir la estructura de los datos, facilitando la interpretación del caso de estudio.

Con métodos directos de ordenación el experimentador puede especificar los factores ambientales de interés y tener un conocimiento independiente sobre el valor de cada variable para cada entidad (para una ordenación de entidades) o la respuesta de las entidades para cada variable (para una ordenación de variables).

Entre los métodos de ordenación se encuentran: el análisis de componentes principales, análisis de correspondencias, análisis de discriminantes y coordenadas principales. La utilización de alguno de ellos depende de los objetivos de trabajo planteados en el estudio.

Coordenadas principales es una técnica de análisis estadístico multivariado la cual maneja entidades a ser representadas en un gráfico de dimensionalidad reducida, preservando tanto como sea, posible la distancia de las relaciones entre ellas.

Cuando se realiza el análisis de un conjunto de datos se dispone de una gran variedad de técnicas para realizarlo, dichas técnicas se agrupan en dos tipos: técnicas tipo **R** y técnicas tipo **Q**, la diferencia fundamental entre ellas estriba en que las **Q** operan sobre una matriz de tamaño **i x i** con elementos que miden la asociación entre individuos. Las técnicas **R** trabajan con una matriz **j x j** que define las relaciones existentes entre variables (Gower, 1966).

El método de coordenadas principales, descrito por Gower (1966), es una generalización del análisis de componentes principales. Que concibe una matriz de asociación **Q** positiva semidefinida, correspondiente a las distancias métricas a ser factorizadas. En otras palabras, una matriz que describe las relaciones entre las entidades en un espacio euclidiano. Se entendiende como matriz de asociación **Q**, a una matriz de tamaño **n x n**, en la cual se realiza la correlación entre entidades tomando valores no negativos (valores positivos incluyendo al cero).

Esta técnica también recibe otros nombres como: escalamiento clásico y *step-across* (Kempton y Digby, 1987).

Coordenadas principales es una técnica de gran utilidad en casos de estudio biológicos, ya que frecuentemente se presenta el registro de datos cuantitativos, cualitativos o mezclas de ambos tipos de datos. Para el caso del análisis de variables cuantitativas el mejor método es sin duda el análisis de componentes principales, pero cuando se presentan variables de tipo cualitativa o mezclas de variables cuantitativas y cualitativas el análisis de coordenadas principales es el método más apropiado, siempre y cuando se elija la medida de distancia adecuada a la estructura de los datos, lo que le proporciona una gran ventaja sobre el análisis de componentes principales.

Para decidir que medida de distancia es la adecuada para los datos que se tiene, deben tomarse en cuenta varios aspectos:

1.- Conocer las propiedades del conjunto de datos y los efectos que tengan sobre cada medida.

2.- La escala de medición de cada una de las variables.

Una vez conocidos estos aspectos se puede seleccionar una medida de distancia adecuada al conjunto de datos.

La elección de una medida de asociación adecuada para el análisis de coordenadas principales, permite trabajar con otras medidas de distancia además de la euclidiana. lo cual le proporciona a la técnica gran flexibilidad de manejo y por lo tanto potencial de uso; debido a que en un gran número de casos los estudios biológicos registran mezclas de tipos de variables.

El producto final del análisis de coordenadas principales es un gráfico que representa en la menor dimensionalidad posible la estructura de los datos, es decir las relaciones existentes entre las diferentes entidades. Esto facilita la interpretación de los resultados, auxiliando en la elaboración de hipótesis de trabajo específicas hacia aquellos aspectos que tengan un mayor efecto sobre el objeto de estudio, o permite la obtención de conclusiones con un enfoque que globaliza la interacción de todas las variables al mismo tiempo.

De lo anterior se puede decir que, coordenadas principales es una técnica de ordenación que dentro del análisis estadístico multivariado presenta un gran potencial de uso, ya que las características de las variables generalmente son diferentes, por lo cual es necesario utilizar una metodología que permita correlacionar diferentes tipos de datos, esta técnica es el análisis de coordenadas principales.

CAPITULO III: FUNDAMENTO MATEMATICO DE LA TECNICA

Coordenadas principales es una técnica que se basa en una serie de principios matemáticos que permiten cumplir los requerimientos para su aplicación, una de sus principales ventajas es que permite manejar diferentes tipos de variables, que de acuerdo con sus características, pueden dividirse en: Binarias, Cualitativas y Cuantitativas. En la realización de un estudio biológico, uno de los primeros factores que ocurren al analizar datos es el tipo de variables que se presenta, ya que estos influye en la selección de la herramienta estadística a utilizar.

Para empezar la información obtenida se registra en una matriz de datos multivariados **X**, donde cada $x_{i,j}$ representa el valor de la j-ésima variable para la i-ésima entidad. Lo que genera una matriz de datos multivariados.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,j} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,j} \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,j} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{i,1} & x_{i,2} & x_{i,3} & \cdots & x_{i,j} \end{bmatrix}$$

Esta matriz es el punto de partida, no solamente para el análisis de coordenadas principales, sino para varios métodos estadísticos multivariados, como: Análisis de Discriminantes, Correspondencias y Componentes Principales, entre otros.

La matriz inicial se transforma a una matriz de distancias métricas, que tiene por objetivo cuantificar diferencias entre entidades; si el tipo de datos no permite la utilización de una medida de distancia la técnica presenta la flexibilidad de utilizar índices de similaridad o inclusive el coeficiente de Gower que permite mezclar variables cuantitativas y cualitativas.

En el caso de datos cualitativos se puede utilizar un índice de similaridad y calcular su complemento por la fórmula:

$$d_{i,j} = 1 - S_{i,j}$$

donde $S_{i,j}$ es la similaridad entre la i-ésima y j-ésima entidad, de esta forma se obtienen valores de distancia.

Para empezar a trabajar la técnica se debe fijar una medida de distancia δ, de manera que las relaciones de distancia entre las entidades se preserve tanto como sea posible, con dicha medida se construye una matriz simétrica de distancias cuadradas **D**.

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} \cdots\cdots d_{1j} \\ d_{21} & d_{22} \cdots\cdots d_{2j} \\ \vdots & \vdots \\ d_{i1} & d_{i2} \cdots\cdots d_{ij} \end{bmatrix}$$

En esta matriz los elementos $d_{i,j}$ representan la distancia cuadrada entre la i-ésima y j-ésima entidad, cumpliendo con las siguientes condiciones:

$d^2_{i,j} = d^2_{j,i}$ Condición que implica la simetría de la matriz de distancias cuadradas.

$d^2_{i,i} = 0$ Para toda i = j, indica que la distancia entre la misma entidad es cero.

La primera restricción a cumplir es que las relaciones de distancia, entre las entidades se preserven tanto como sea posible, por lo cual se establece la siguiente ecuación:

$$d^2(i,j) = \sum_{k=1}^{1} (c_{k,i} - c_{k,j})^2$$

donde **k** denota el **k**-ésimo renglón de la matriz de coordenadas principales **C**. En la cual $d_{i,j}$ representa la distancia en términos de los datos originales y las $c_{k,i}$, $c_{k,j}$ representan las coordenadas de las entidades transformadas, que se conocen como coordenadas principales.

De tal manera que la distancia cuadrada entre la **i**-ésima y **j**-ésima entidad está dada por:

$$d^2(i,j) = \sum_{k=1}^{1} (c^2_{k,i} + c^2_{k,j} - 2c_{k,i}c_{k,j})$$

Si se considera la formación de una matriz, **A** de asociación; simétrica de tamaño **i x j** que cumpla con la condición de ser igual a la premultiplicación de la matriz de coordenadas **C** por su transpuesta **C'**, se tiene:

$$A=C'C= \begin{bmatrix} c_{11} & c_{21} & \cdots & c_{k1} \\ c_{12} & c_{22} & \cdots & c_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ c_{11} & c_{21} & \cdots & c_{k1} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{11} \\ c_{21} & c_{22} & \cdots & c_{21} \\ \cdots & \cdots & \cdots & \cdots \\ c_{k1} & c_{k2} & \cdots & c_{k1} \end{bmatrix}$$

Al realizar las operaciones:

$$a_{11} = c_{11}c_{11} + c_{21}c_{21} + \cdots + c_{k1}c_{k1}$$

$$a_{11} = c^2_{11} + c^2_{21} + \cdots + c^2_{k1}$$

La cual se puede expresar de la siguiente manera, ya que el indice que se modifica es el **k**:

$$a_{11} = \sum_{k=i}^{i} c_{k1}^2$$

Para a_{12}

$$a_{12} = c_{11} c_{12} + c_{21} c_{22} + \ldots + c_{k1} c_{k2}$$

De forma análoga:

$$a_{12} = \sum_{k=1}^{i} c_{k1} c_{k2}$$

$$a_{11} = c_{11} c_{11} + c_{21} c_{21} + \ldots + c_{k1} c_{k1}$$

$$a_{11} = \sum_{k=1}^{i} c_{k1} c_{k1}$$

$$a_{21} = c_{12} c_{11} + c_{22} c_{21} + \ldots + c_{k2} c_{k1}$$

$$a_{21} = \sum_{k=1}^{i} c_{k2} c_{k1}$$

$$a_{22} = c_{12} c_{12} + c_{22} c_{22} + \ldots + c_{k2} c_{k2}$$

$$a_{22} = c_{12}^2 + c_{22}^2 + \ldots + c_{k2}^2$$

$$a_{22} = \sum_{k=1}^{i} c_{k2}^2$$

$$a_{21} = c_{12} c_{11} + c_{22} c_{21} + \ldots + c_{k2} c_{k1}$$

$$a_{21} = \sum_{k=1}^{i} c_{k2} c_{k1}$$

$$a_{k1} = c_{11} c_{11} + c_{21} c_{21} + \ldots + c_{k1} c_{k1}$$

$$a_{k1} = \sum_{k=1}^{1} c_{k1} c_{k1}$$

$$a_{k2} = c_{11} c_{12} + c_{21} c_{22} + \ldots + c_{k1} c_{k2}$$

$$a_{k2} = \sum_{k=1}^{1} c_{k1} c_{k2}$$

$$a_{k1} = c_{11} c_{11} + c_{21} c_{21} + \ldots + c_{k1} c_{k1}$$

$$a_{k1} = c_{11}^2 + c_{21}^2 + \ldots + c_{k1}^2$$

$$a_{k1} = \sum_{k=1}^{1} c_{k1}^2$$

Obteniendo la matriz:

$$A = \begin{bmatrix} \Sigma c_{k1}^2 & \Sigma c_{k1} c_{k2} & \ldots & \Sigma c_{k1} c_{kn} \\ \Sigma c_{k2} c_{k1} & \Sigma c_{k2}^2 & \ldots & \Sigma c_{k2} c_{kn} \\ \ldots & \ldots & \ldots & \ldots \\ \Sigma c_{kn} c_{k1} & \Sigma c_{kn} c_{k2} & \ldots & \Sigma c_{kn}^2 \end{bmatrix}$$

Donde se tiene que:

$$a_{i,j} = \sum_{k=1}^{1} c_{ki} c_{kj} \quad ; \quad a_{i,i} = \sum_{k=1}^{1} c_{k,i}^2 \quad ; \quad a_{j,j} = \sum_{k=1}^{1} c_{k,j}^2$$

De esta manera, una fórmula alternativa para $d^2_{i,j}$ es:

$$d^2_{i,j} = a_{i,i} + a_{j,j} - 2a_{i,j} \quad \ldots\ldots\ldots(1)$$

Se puede encontrar $a_{i,j}$ como una función de $d^2_{i,j}$, a partir de la ecuación (1):

$$a_{i,j} = -\frac{1}{2}\left[d^2_{i,j} + a_{i,i} + a_{j,j} \right] \ldots\ldots\ldots(2)$$

Ahora hay que fijar el orígen de las coordenadas, con respecto a las hileras, en la media igual a cero, estableciendo la siguiente igualdad $\sum_i a_{i,j} = 0$, y como A es simétrica, esta condición también se cumple para las columnas de manera que $\sum_j a_{i,j} = 0$. Y la ecuación (1) se puede expresar de la siguiente forma:

$$\sum_i d^2_{i,j} = \sum_i a_{i,i} + \sum_i a_{j,j} - 2 \sum_i a_{i,j}$$

Además se debe cumplir con las condiciones de independencia, para lo cual son necesarias las siguientes operaciones

$$\sum_i a_{i,i} = x \qquad \sum_j a_{j,j} = na_{j,j} \qquad \sum_i a_{i,j} = 0$$

$$\sum_i d^2_{i,j} = x + na_{j,j} \qquad \text{donde} \qquad a_{j,j} = \frac{1}{n}\left(\sum_i d^2_{i,j} - x\right)$$

de igual forma:

$$\sum_j d^2_{i,j} = x + na_{i,i} \qquad \text{donde} \qquad a_{i,i} = \frac{1}{n}\left(\sum_j d^2_{i,j} - x\right)$$

se observa que $\sum_i a_{i,i} = \sum_i a_{j,j}$, están denotados por x.

Entonces, la ecuación (2) se expresa como:

$$a_{i,j} = \frac{1}{2}\left(-d^2_{i,j} + \frac{1}{n}\left(\sum_i d^2_{i,j} - x\right) + \frac{1}{n}\left(\sum_j d^2_{i,j} - x\right)\right)$$

$$= -\frac{1}{2}d^2_{i,j} + \frac{1}{2n}\sum_i d^2_{i,j} + \frac{1}{2n}\sum_j d^2_{i,j} - \frac{x}{n} \qquad \ldots\ldots(3)$$

como se mostró anteriormente:

$$a_{i,i} = \sum_k c^2_{k,i}$$

por lo tanto:

$$\sum_i a_{i,i} = \sum_i \sum_k c^2_{k,i}$$

Esta última suma es igual a $(\frac{1}{n})$ veces la suma de las distancias cuadradas entre todos los pares de puntos. Que es:

$$\sum_{i}\sum_{k} c^2_{k,l} = \frac{1}{n}\sum_{i<j} d^2_{i,j} = \frac{1}{n}\sum_{i}\sum_{j} d^2_{i,j}$$

$\sum_{i<j}$ especifica que cada par de puntos se debe considerar solamente una vez. La forma $\sum_{i}\sum_{j}$ especifica que $d^2_{i,j}$ y $d^2_{j,i}$ es la misma. Por la cual la fórmula (3) ahora queda como:

$$a_{i,j} = -\frac{1}{2} d^2_{i,j} + \frac{1}{2n}\sum_{i=1}^{n} d^2_{i,j} + \frac{1}{2n}\sum_{j=1}^{n} d^2_{i,j} - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{i,j}$$

Donde:

$\sum_{j=1}^{n} d^2_{i,j}$ es la suma de los elementos en el **i**-ésimo renglón de **D**.

$\sum_{i=1}^{n} d^2_{i,j}$ es la suma de los elementos en la **j**-ésima columna de **D**.

$\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{i,j}$ es la suma del total de elementos de **D**.

A partir de **D** puede encontrarse una matriz de asociación **A** de tamaño **i x j**, simétrica, que muestra las relaciones entre las entidades estudiadas y que permite cumplir con las condiciones y restricciones pre-establecidas.

$$A = \begin{bmatrix} a_{11} & a_{12} \cdots\cdots a_{1j} \\ a_{21} & a_{22} \cdots\cdots a_{2j} \\ \vdots & \vdots \\ a_{i1} & a_{i2} \cdots\cdots a_{ij} \end{bmatrix}$$

Como **A** es una matriz cuadrada simétrica se puede encontrar una matriz ortogonal **U** tal que $A=U\Lambda U'$, donde Λ es la matriz diagonal con elementos λ, que son las raíces latentes o eigenvalores de **A**.

Para encontrar eigenvalores (λ) y eigenvectores (**u**), de una matriz **A**, se plantea la ecuación:

$$\lambda IU - AU = 0$$

o

$$(\lambda I - A)U = 0$$

donde **I** garantiza la conformabilidad de la operación.

Esta ecuación tiene una solución no trivial, si y sólo si el determinante de la matriz ($\lambda I - A$) es cero.

De tal manera que se puede construir la ecuación característica de la matriz **A**, y calcular sus eigenvalores, llamados también valores característicos, valores propios o raíces latentes:

$$\det(\lambda I - A) = 0$$

En donde se encontrará un polinomio con tantas raíces latentes (λ) como entidades se estén comparando, y al menos una de esas raíces es cero.

Las λ's representan la variación explicada de cada uno de los ejes en el nuevo sistema de coordenadas, de tal manera que la suma de los elementos de la diagonal, la traza de la matriz Λ, representa el total de variación y si se desea conocer el porcentaje de varianza que explican cada uno de los ejes, sólo se necesita el siguiente cálculo:

$$\%\ \text{de varianza del eje}_i = (\lambda_i\ /\ \text{traza de}\ \Lambda\ )\ \text{x}\ 100$$

Por lo tanto se tiene que:

$$A = (U\Lambda^{1/2})\ (\Lambda^{1/2}U')$$

$$A = C'\,C$$

$$C = \Lambda^{1/2}\,U'$$

Donde C es la matriz con elementos $c_{i,j}$ que representan las coordenadas principales en un nuevo sistema de ejes.

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & \cdots & c_{1,j} \\ c_{2,1} & c_{2,2} & c_{2,3} & \cdots & c_{2,j} \\ c_{3,1} & c_{3,2} & c_{3,3} & \cdots & c_{3,j} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ c_{i,1} & c_{i,2} & c_{i,3} & \cdots & c_{i,j} \end{bmatrix}$$

De esta manera se cumple la siguiente condicion:

$$A = C'C$$

Lo cual quiere decir que la matriz de coordenadas principales multiplicada por su transpuesta, debe ser igual a la matriz de asociación que se obtiene a partir de las distancias métricas de los datos originales.

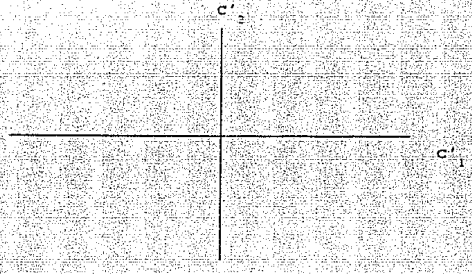
Además, se cumple que la distancia original debe preservarse tanto como sea posible en el nuevo espacio de coordenadas $c_{i,j}$, de tal manera que se conserva la estructura de los datos. Esto se refleja en la siguiente fórmula:

$$d^2_{i,j} = \sum_k (c_{i,1} - c_{k,j})^2$$

donde k denota el k-ésimo renglón de la matriz C.

De la matriz C resultante se eligen dos vectores de tal manera que su combinación de porcentaje de varianza explicada sea la mayor, y estos dos vectores comprenden las nuevas coordenadas en un espacio bidimensional.

Las nuevas coordenadas se representan finalmente en un gráfico de dos dimensiones.



donde c'_1 y c'_2 representan los vectores de coordenadas principales 1 y 2. En el cual se procede a interpretar por agrupamiento o como un paso intermedio para la realización de alguna técnica de agrupamiento, como cluster, a partir de las relaciones de distancia de las nuevas coordenadas generadas.

Cuando existen eigenvalores negativos implica que las coordenadas no se encuentran en los números reales y por lo tanto se tiene que construir en un espacio con dimensiones no reales que corresponde al campo de los números complejos.

Las operaciones necesarias para la aplicación de la técnica son:

1.- Registrar la información colectada en una matriz X de datos multivariados, de manera que las columnas representen las variables y las hileras a las entidades.

2.- A partir de la matriz **X** calcular una matriz **D**, utilizando cualquiera de las medidas de distancia métricas o indices de similaridad, de acuerdo con el tipo de datos.

Cuando se trate de la aplicación de índices de similaridad se debe calcular la matriz complemento con la siguiente fórmula:

$$d_{i,j} = 1 - s_{i,j}$$

3.- Encontrar los elementos de la matriz simétrica de asociación **A**, con la ecuación:

$$a_{i,j} = -\frac{1}{2} d^{-}_{i,j} + \frac{1}{2n} \sum_{i=1}^{n} d^{2}_{i,j} - \frac{1}{2n} \sum_{j=1}^{n} d^{2}_{i,j} - \frac{1}{2n^{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} d^{2}_{i,j}$$

$\sum_{j=1}^{n} d^{2}_{i,j}$ es la suma de los elementos en el **i**-ésimo renglón de **D**.

$\sum_{i=1}^{n} d^{2}_{i,j}$ es la suma de los elementos en la **j**-ésima columna de **D**.

$\sum_{i=1}^{n} \sum_{j=1}^{n} d^{2}_{i,j}$ es la suma del total de elementos de **D**.

4.- Realizar un eigenanálisis de **A** para la obtención de los eigenvalores (λ) y los eigenvectores (**u**), con los cuales se generan las matrices Λ y **U**.

5.- Obtener la matriz de coordenadas principales **C** a partir de la siguiente fórmula:

$$C = \Lambda^{1/2} U'$$

6.- Hacer la grafica de los dos primeros vectores de **c'** de tal manera que cada uno representa la coordenada principal de una entidad en ese eje.

7.- Interpretar el fenómeno en base a agrupamiento de los puntos cercanos, o continuar con alguna técnica de clasificación.

Con el propósito de mostrar la aplicación de la técnica de análisis de coordenadas principales, utilizando una matriz de datos, se realizan los cálculos correspondientes.

En este caso se tiene una matriz **X** de tamaño **4 x 3**.

$$\mathbf{X} = \begin{bmatrix} 5.00 & 2.00 & 1.00 \\ 10.00 & 14.00 & 2.00 \\ 6.00 & 3.00 & 10.00 \\ 7.00 & 12.00 & 4.00 \end{bmatrix}$$

Matriz 3.1: Datos originales

Despues se procede a calcular una matriz **D**, correspondiente a la relaciones de distancia entre las entidades de la matriz **X**, para la construcción de dicha matriz se utiliza la medida de distancia de Manhattan, que presenta la siguiente expresión matemática:

$$d_{i,j} = \sum_{k=1}^{p} | X_{i,k} - X_{j,k} |$$

Donde $d_{i,j}$ se expresa como el cálculo de la sumatoria de las diferencias absolutas de las entidades **i** y **j** para las **p** variables.

Calculando la distancia entre la entidad **1** y **1** se tiene:

$$d_{1,1} = \sum | X_{1,k} - X_{1,k} |$$
$$d_{1,2} = | X_{1,1} - X_{1,1} | + | X_{1,2} - X_{1,2} | + | X_{1,3} - X_{1,3} |$$
$$d_{1,2} = | 5.00 - 5.00 | + | 2.00 - 2.00 | + | 1.00 - 1.00 |$$
$$d_{1,1} = 0 + 0 + 0$$
$$d_{1,1} = 0$$

Obtención de las demas distancias:

$$d_{1,2} = \sum | X_{1,k} - X_{2,k} |$$
$$d_{1,2} = | X_{1,1} - X_{2,1} | + | X_{1,2} - X_{2,2} | + | X_{1,3} - X_{2,3} |$$
$$d_{1,2} = | 5.00 - 10.00 | + | 2.00 - 14.00 | + | 1.00 - 2.00 |$$
$$d_{1,3} = 5 + 12 + 1$$
$$d_{1,2} = 18$$

$$d_{1,3} = \sum | X_{1,k} - X_{3,k} |$$
$$d_{1,3} = | X_{1,1} - X_{3,1} | + | X_{1,2} - X_{3,2} | + | X_{1,3} - X_{3,3} |$$
$$d_{1,3} = | 5.00 - 6.00 | + | 2.00 - 1.00 | + | 1.00 - 10.00 |$$
$$d_{1,3} = 1 + 1 + 9$$
$$d_{1,3} = 11$$

$$d_{1,4} = \sum | X_{1,k} - X_{4,k} |$$
$$d_{1,4} = | X_{1,1} - X_{4,1} | + | X_{1,2} - X_{4,2} | + | X_{1,3} - X_{4,3} |$$
$$d_{1,4} = | 5.00 - 7.00 | + | 2.00 - 12.00 | + | 1.00 - 4.00 |$$
$$d_{1,4} = 2 + 10 + 3$$
$$d_{1,4} = 15$$

$$d_{2,2} = \sum | x_{2,k} - x_{2,k} |$$
$$d_{2,2} = | x_{2,1} - x_{2,1} | + | x_{2,2} - x_{2,2} | + | x_{2,3} - x_{2,3} |$$
$$d_{2,2} = | 10.00 - 10.00 | + | 14.00 - 14.00 | + | 2.00 - 2.00 |$$
$$d_{2,2} = 0 + 0 + 0$$
$$d_{2,2} = 0$$

$$d_{2,3} = \sum | x_{2,k} - x_{3,k} |$$
$$d_{2,3} = | x_{2,1} - x_{3,1} | + | x_{2,2} - x_{3,2} | + | x_{2,3} - x_{3,3} |$$
$$d_{2,3} = | 10.00 - 6.00 | + | 14.00 - 3.00 | + | 2.00 - 10.00 |$$
$$d_{2,3} = 4 + 11 + 8$$
$$d_{2,3} = 23$$

$$d_{2,4} = \sum | x_{2,k} - x_{4,k} |$$
$$d_{2,4} = | x_{2,1} - x_{4,1} | + | x_{2,2} - x_{4,2} | + | x_{2,3} - x_{4,3} |$$
$$d_{2,4} = | 10.00 - 7.00 | + | 14.00 - 12.00 | + | 2.00 - 4.00 |$$
$$d_{2,4} = 3 + 2 + 2$$
$$d_{2,4} = 7$$

$$d_{3,3} = \sum | x_{3,k} - x_{3,k} |$$
$$d_{3,3} = | x_{3,1} - x_{3,1} | + | x_{3,2} - x_{3,2} | + | x_{3,3} - x_{3,3} |$$
$$d_{3,3} = | 6.00 - 6.00 | + | 3.00 - 3.00 | + | 10.00 - 10.00 |$$
$$d_{3,3} = 0 + 0 + 0$$
$$d_{3,3} = 0$$

$$d_{3,4} = \sum | x_{3,k} - x_{4,k} |$$
$$d_{3,4} = | x_{3,1} - x_{4,1} | + | x_{3,2} - x_{4,2} | + | x_{3,3} - x_{4,3} |$$
$$d_{3,4} = | 6.00 - 7.00 | + | 3.00 - 12.00 | + | 10.00 - 4.00 |$$
$$d_{3,4} = 1 + 9 + 6$$
$$d_{3,4} = 16$$

$$d_{4,4} = \sum | x_{4,k} - x_{4,k} |$$

$$d_{4,4} = | x_{4,1} - x_{4,1} | + | x_{4,2} - x_{4,2} | + | x_{4,3} - x_{4,3} |$$

$$d_{4,4} = | 7.00 - 7.00 | + | 12.00 - 12.00 | + | 4.00 - 4.00 |$$

$$d_{4,4} = 0 + 0 + 0$$

$$d_{4,4} = 0$$

Como $d_{1,4} = d_{4,1}$; se construye la matriz:

$$D = \begin{bmatrix} 0 & 18 & 11 & 15 \\ 18 & 0 & 23 & 7 \\ 11 & 23 & 0 & 16 \\ 15 & 7 & 16 & 0 \end{bmatrix}$$

Matriz 3.2: Matriz de distancias

Con la cual se calcula la matriz de distancias cuadradas:

$$D^2 = \begin{bmatrix} 0 & 324 & 121 & 225 \\ 324 & 0 & 529 & 49 \\ 121 & 529 & 0 & 256 \\ 225 & 49 & 256 & 0 \end{bmatrix}$$

Matriz 3.3: Matriz de distancias al cuadrado

A partir de la matriz D^2 se encuentran los elementos de la matriz de asociación **A**, con la ecuación:

$$a_{i,j} = -\frac{1}{2} d^2_{i,j} + \frac{1}{2n} \sum_{i=1}^{n} d^2_{i,j} + \frac{1}{2n} \sum_{j=1}^{n} d^2_{i,j} - \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d^2_{i,j}$$

Mostrando el cálculo de cada uno de los elementos de **A**:

$$a_{1,1} = -\frac{1}{2} d^2_{1,1} + \frac{1}{2(4)} \sum_{i=1}^{4} d^2_{i,1} + \frac{1}{2(4)} \sum_{j=1}^{4} d^2_{1,j} - \frac{1}{2(4)^2} \sum_{i=1}^{4} \sum_{j=1}^{4} d^2_{i,j}$$

$$a_{1,1} = \frac{1}{2}(0) + \frac{1}{8}(0+324+121+225) + \frac{1}{8}(0+324+121+225) - \frac{1}{32}(0+324+121+225) + (0+324+121+225))$$

$$a_{1,1} = 0 + 83.75 + 83.75 - 94$$

$$a_{1,1} = 73.75$$

$$a_{2,1} = -162 + 112.75 + 83.75 - 94$$

$$a_{2,1} = -59.5$$

$$a_{3,1} = -60.5 + 113.25 + 83.75 - 94$$

$$a_{3,1} = -42.5$$

$$a_{4,1} = -112.5 + 66.25 + 83.75 - 94$$

$$a_{4,1} = -56.5$$

$$a_{1,2} = -162 + 83.75 + 112.75 - 94$$

$$a_{1,2} = -59.5$$

$$a_{2,2} = -0 + 112.75 + 112.75 - 94$$

$$a_{2,2} = 131.5$$

$$a_{3,2} = -264.5 + 113.25 + 112.75 - 94$$

$$a_{3,2} = -132.5$$

$$a_{4,2} = -24.5 + 66.25 + 112.75 - 94$$

$$a_{4,2} = 60.5$$

$$a_{1,3} = -60.5 + 83.75 + 113.25 - 94$$

$a_{1,1} = 12.5$

$a_{2,1} = -264.5 + 112.75 + 113.25 - 94$

$a_{2,1} = -132.5$

$a_{3,1} = -0 + 113.25 + 113.25 - 94$

$a_{3,1} = 132.5$

$a_{4,1} = -120 + 66.25 + 113.5 - 94$

$a_{4,1} = -42.5$

$a_{1,4} = -112.5 + 83.75 + 66.25 - 94$

$a_{1,4} = -56.5$

$a_{2,4} = -24.5 + 112.75 + 66.25 - 94$

$a_{2,4} = 60.5$

$a_{3,4} = -120 + 113.25 + 66.25 - 94$

$a_{3,4} = -42.5$

$a_{4,4} = 0 + 66.25 + 66.25 - 94$

$a_{4,4} = 38.5$

Con estos valores se construye la matriz:

$$\mathbf{A} = \begin{bmatrix} 73.5 & -59.5 & 42.5 & -56.5 \\ -59.5 & 131.5 & -132.5 & 60.5 \\ 42.5 & -132.5 & 132.5 & -42.5 \\ -56.5 & 60.5 & -42.5 & 38.5 \end{bmatrix}$$

Matriz 3.4: Matriz de asociación

Posteriormente con el auxilio del paquete estadístico NTSYS-PC versión 1.60 se procede a calcular los eigenvalores y eigenvectores que permiten la obtención de las coordenadas en el nuevo sistema de ejes.

$$\Lambda = \begin{bmatrix} 316.45119 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 68.12552 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 8.57676 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 \end{bmatrix}$$

Matriz 3.5: Eigenvalores

Aquí se cumple la condición de que al menos una de las raíces latentes (λ) es cero, de lo cual se desprende la reducción en al menos una dimensión.

En esta matriz se calcula la variación explicada de cada uno de los ejes en el nuevo sistema de coordenadas, el cálculo se realiza con la siguiente ecuación:

$$\text{\% de varianza del eje } i = \left[\frac{\lambda_i}{\text{traza de } \Lambda} \right] X\ 100$$

Calculando la varianza para cada uno de los ejes:

$$\text{\% de varianza del eje } 1 = \frac{316.45119}{393.15341} X\ 100 = 80.49\ \%$$

$$\text{\% de varianza del eje } 2 = \frac{68.12552}{393.15341} X\ 100 = 17.32\ \%$$

$$\text{\% de varianza del eje } 3 = \frac{8.57676}{393.15341} X\ 100 = 2.19\ \%$$

Como se puede observar los dos primeros ejes explican el 97.81 % de la información original.

$$U = \begin{bmatrix} 0.33625 & 0.75033 & 0.27192 & -0.20410 \\ -0.84693 & 0.15743 & -0.55380 & -0.60424 \\ 0.41368 & -0.46418 & -0.39740 & 0.76923 \\ -0.30300 & -0.44358 & 0.67929 & 0.03910 \end{bmatrix}$$

Matriz 3.6: Eigenvectores

A partir de los eigenvalores y eigenvectores se obtiene la matriz de coordenadas principales **C**, con la fórmula:

$$C = \Lambda^{1/2} U'$$

Obteniendo la matriz:

$$C = \begin{bmatrix} 5.98165 & 6.19306 & 0.79635 \\ -11.50834 & 1.29941 & -1.62187 \\ 10.91678 & -3.43126 & -1.16383 \\ -5.39009 & -3.66122 & 1.98936 \end{bmatrix}$$

Matriz 3.7: Matriz **C**

Donde las nuevas coordenadas se encuentran en los vectores **c'**, de tal manera que en un gráfico de dos dimensiones el eje 1 está representado por el vector c'_1 y el eje dos por el vector c'_2, de la siguiente manera:

$$\begin{bmatrix} c'_1 \\ c'_2 \end{bmatrix} = \begin{bmatrix} (& 5.98165 & -11.50834 & 10.9167 & -5.39009 &) \\ (& 6.19306 & 1.29941 & -3.4126 & -3.66122 &) \end{bmatrix}$$

Matriz 3.8: Vectores de coordenadas principales c'_1 y c'_2

Las cuales generan el siguiente gráfico:

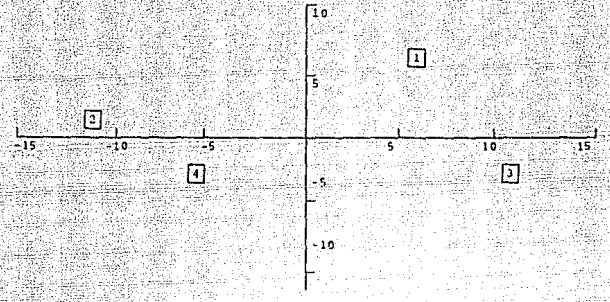


Gráfico 3.1. Coordenadas principales c'_1 y c'_2 para las cuatro entidades

Al realizar la agrupación entre las entidades más cercanas se obtiene que las entidades **2** y **4** son las que guardan mayor similaridad, posteriormente las entidades **1** y33, despues las **4** y **1**, **3** y **4**, **1** y **2** y que las más disimilares son las entidades **2** y **3**.

Al comparar estas observaciones, con los datos de distancia calculados al principio, se comprueba que las relaciones de distancia se mantienen, ya que en los valores numericos se detecta una situación similar.

Las relaciones de distancia se conservan entre la matriz **D** y la matriz **C** con la diferencia que a partir de la matriz **C** se puede construir un gráfico en el cual estas relaciones se observan.

Si en el gráfico no es posible detectar las relaciones de distancia entre las entidades se puede continuar con alguna técnica de clasificación, como cluster.

CLUSTER PARA LAS CUATRO ENTIDADES

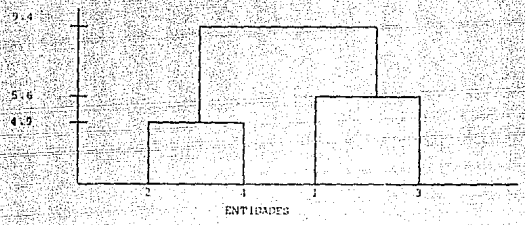


Gráfico 3.2 : Cluster para las cuatro entidades

En el cual se corrobora que las entidades **2** y **4** forman el primer cluster a una distancia de **4.9** unidades, posteriormente a **5.6** se forma otro cluster entre las entidades **1** y **3**, observándose dos cluster a esta distancia y finalmente todas las entidades forman un solo cluster a una distancia de **9.34**.

CAPITULO IV: ANALISIS DE COORDENADAS PRINCIPALES CON DIFERENTES MEDIDAS DE DISTANCIA.

Al analizar un estudio biológico se debe seleccionar una medida de distancia que permita una visualización adecuada de la estructura de los datos y por lo tanto una mejor interpretación de las relaciones existentes entre las entidades sujetas a estudio.

Rivera (1992) realiza una revisión de las diferentes medidas de distancia utilizadas en biología, considerando que la mayoría de las medidas de distancia se derivan básicamente de la distancia Euclidiana y de la distancia de Manhattan, ya que conceptualmente presentan una forma matemática semejante, midiendo la hipotenusa, o los catetos de un triángulo.

Las medidas que propone son:

Medida de distancia euclidiana.
Métrica de Manhattan.
Coeficiente de correlación producto momento.

La selección de cualesquiera de las medidas de distancia corresponde al investigador y depende de las propiedades del conjunto de datos y los efectos que tengan sobre cada medida. Esta selección se realiza en función de aquella que represente de la mejor manera la asociación natural de los datos.

Ya que el análisis de coordenadas principales se basa en las relaciones existentes entre las entidades sujetas a estudio y que la medida de distancia que utiliza el análisis es la de Manhattan, se analiza el comportamiento de la técnica al utilizar diferentes medidas de distancia como punto de partida.

Se trabaja con una matriz de datos de 10 x 6, con los cuales se realiza el análisis de coordenadas principales con el propósito de detectar la posibilidad de utilizar cualesquiera de las medidas de distancia .

$$X = \begin{bmatrix} 15 & 25 & 23 & 32 & 50 & 19 \\ 2 & 12 & 0 & 9 & 34 & 27 \\ 4 & 16 & 10 & 28 & 45 & 16 \\ 3 & 23 & 30 & 0 & 60 & 22 \\ 19 & 3 & 22 & 5 & 56 & 45 \\ 23 & 17 & 47 & 29 & 12 & 32 \\ 21 & 19 & 20 & 45 & 29 & 36 \\ 15 & 0 & 0 & 0 & 0 & 0 \\ 12 & 56 & 45 & 15 & 2 & 36 \\ 18 & 29 & 54 & 0 & 10 & 0 \end{bmatrix}$$

Matriz 4.1: Datos originales

MEDIDA DE DISTANCIA EUCLIDIANA

La primera medida de distancia a trabajar es la euclidiana, con la cual después de aplicarla a la matriz X se obtiene:

$$D = \begin{bmatrix} 0.00 & & & & & & & & & \\ 41.42 & 0.00 & & & & & & & & \\ 20.52 & 26.89 & 0.00 & & & & & & & \\ 36.47 & 42.47 & 38.66 & 0.00 & & & & & & \\ 44.07 & 40.96 & 38.66 & 35.91 & 0.00 & & & & & \\ 48.23 & 59.86 & 55.47 & 63.01 & 59.31 & 0.00 & & & & \\ 31.30 & 47.03 & 36.64 & 60.18 & 51.71 & 36.03 & 0.00 & & & \\ 70.99 & 47.73 & 59.51 & 75.21 & 75.46 & 67.61 & 70.42 & 0.00 & & \\ 65.85 & 72.12 & 72.85 & 71.97 & 80.52 & 44.25 & 60.86 & 81.82 & 0.00 & \\ 71.07 & 81.89 & 79.41 & 74.97 & 81.89 & 51.83 & 75.28 & 70.33 & 60.75 & 0.00 \end{bmatrix}$$

Matriz 4.2: Matriz de distancia tipo Manhattan

A partir de esta matriz D se realiza el cálculo de asociación para obtener una matriz A:

A=

703									
-15	980								
548	535	814							
380	281	352	1146						
176	446	188	844	1591					
-266	-757	-587	-747	-418	1088				
379	-98	253	-599	-23	413	1035			
-803	714	0	-771	-687	-377	-597	2728		
-493	-788	-923	-573	-1123	888	-11	-660	2645	
-609	-1299	-1181	-553	-993	764	-751	455	1041	3128

Matriz 4.3: Asociación entre las entidades

Al realizar el eigen análisis se obtiene la matriz Λ y la matriz U con los siguientes valores:

6699.73									
0.00	1980.79								
0.00	0.00	2489.13							
0.00	0.00	0.00	1561.23						
0.00	0.00	0.00	0.00	1147.79					
0.00	0.00	0.00	0.00	0.00	223.62				
0.00	0.00	0.00	0.00	0.00	0.00	0.00			
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Matriz 4.4: Eigenvalores

Con estos valores se procede a calcular el porciento de variación explicada.

% variación eje 1 = 41.60 %
% variación eje 2 = 24.72 %
% variación eje 3 = 15.45 %
% variación eje 4 = 9.69 %
% variación eje 5 = 7.12 %
% variación eje 6 = 1.38 %
% variación eje 7 = 0.00 %
% variación eje 8 = 0.00 %
% variación eje 9 = 0.00 %
% variación eje 10 = 0.00 %

U=

0.17	-0.21	0.02	-0.19	0.48	-0.14	-0.04	-0.02	-0.01	-0.02
0.29	0.23	0.12	0.30	-0.13	-0.20	-0.08	-0.13	-0.14	-0.13
0.29	0.03	0.15	-0.03	0.42	0.18	-0.13	-0.14	-0.15	-0.15
0.25	-0.15	-0.52	0.30	0.16	0.36	0.50	0.52	0.52	0.52
0.33	-0.14	-0.32	-0.14	-0.05	-0.25	0.31	0.32	0.32	0.32
-0.24	-0.18	0.20	-0.27	-0.21	0.71	-0.23	-0.20	-0.19	-0.20
0.05	-0.23	0.49	-0.34	-0.03	-0.31	-0.51	-0.49	-0.48	-0.49
-0.06	0.81	0.15	0.04	-0.06	0.10	-0.05	-0.16	-0.18	-0.16
-0.47	-0.30	0.20	0.68	-0.05	-0.17	-0.26	-0.20	-0.20	-0.20
-0.57	0.14	-0.50	-0.33	0.17	-0.28	0.49	0.49	0.50	0.49

Matriz 4.5: Eigenvectores

A partir de las matrices 4.4 y 4.5 se encuentran las nuevas coordenadas que se registran en la matriz **C**:

$$\mathbf{C} = \begin{bmatrix}
14.1 & -13.6 & 1.1 & -7.5 & 16.1 & -2.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
23.6 & 14.7 & 6.1 & 11.7 & -4.5 & -2.9 & 0.0 & 0.0 & 0.0 & 0.0 \\
23.5 & 1.6 & 7.3 & -1.4 & 14.1 & 2.6 & 0.0 & 0.0 & 0.0 & 0.0 \\
20.5 & -9.6 & -26.0 & 11.7 & 5.5 & 5.4 & 0.0 & 0.0 & 0.0 & 0.0 \\
26.9 & -9.1 & -15.8 & -5.7 & -22.1 & -3.7 & 0.0 & 0.0 & 0.0 & 0.0 \\
-22.9 & -11.3 & 10.0 & -10.8 & -10.4 & 10.6 & 0.0 & 0.0 & 0.0 & 0.0 \\
4.5 & -14.3 & 24.6 & -13.6 & -0.9 & -4.6 & 0.0 & 0.0 & 0.0 & 0.0 \\
-4.8 & 51.4 & 7.6 & 1.8 & -1.9 & 1.5 & 0.0 & 0.0 & 0.0 & 0.0 \\
-38.3 & -18.8 & 9.9 & 24.8 & -1.6 & -2.5 & 0.0 & 0.0 & 0.0 & 0.0 \\
-47.0 & 8.9 & -24.7 & -13.1 & 5.8 & -4.2 & 0.0 & 0.0 & 0.0 & 0.0
\end{bmatrix}$$

Matriz 4.6: Matriz **C**

Para obtener las nuevas cordenadas se obtiene la traspuesta de la matriz **C**, de manera que cada vector corresponde a las coordenadas principales en el nuevo sistema de ejes.

$$\begin{bmatrix} c'_1 \\ c'_2 \\ c'_3 \\ c'_4 \\ c'_5 \\ c'_6 \\ c'_7 \\ c'_8 \\ c'_9 \\ c'_{10} \end{bmatrix} = \begin{bmatrix}
14.1 & 23.6 & 23.5 & 20.5 & 26.9 & -22.9 & 4.5 & -4.8 & -38.3 & -47.0 \\
-13.6 & 14.7 & 1.6 & -9.6 & -9.1 & -11.3 & -14.3 & 51.4 & -18.8 & 8.9 \\
1.1 & 6.1 & 7.3 & -26.0 & -15.8 & 10.0 & 24.6 & 7.6 & 9.9 & -24.7 \\
-7.5 & 11.9 & -1.4 & 11.7 & -5.7 & -10.8 & -13.6 & 1.8 & 24.8 & -13.1 \\
16.1 & -4.5 & 14.1 & 5.5 & -22.1 & -10.4 & -0.9 & -1.9 & -1.6 & 5.8 \\
-2.0 & -2.9 & 2.6 & 5.4 & -3.7 & 10.6 & -4.6 & 1.5 & -2.5 & -4.3 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0
\end{bmatrix}$$

Matriz 4.7 : Coordenadas principales

A partir de esta matriz se obtiene el gráfico, seleccionando la combinación de vectores que represente el mayor porcentaje de variación explicada, que en este caso corresponde a los vectores c'_1 y c'_2, que abarcan el 66.38 % de variación.

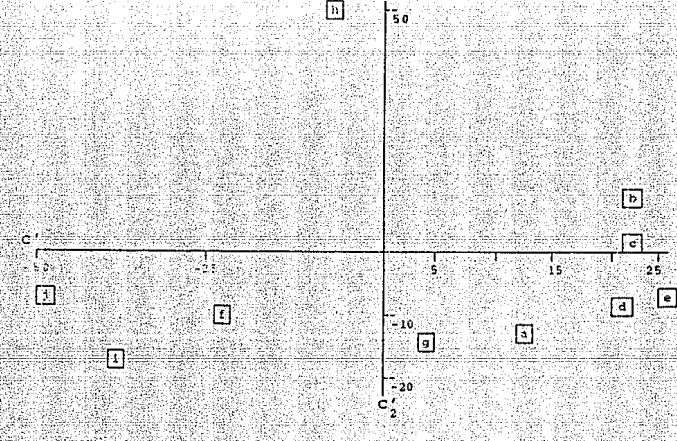


Gráfico 4.1: Coordenadas principales c'_1 y c'_2 utilizando la medida de distancia euclidiana

Con el auxilio de un cluster se pueden identificar grupos de entidades cercanas:

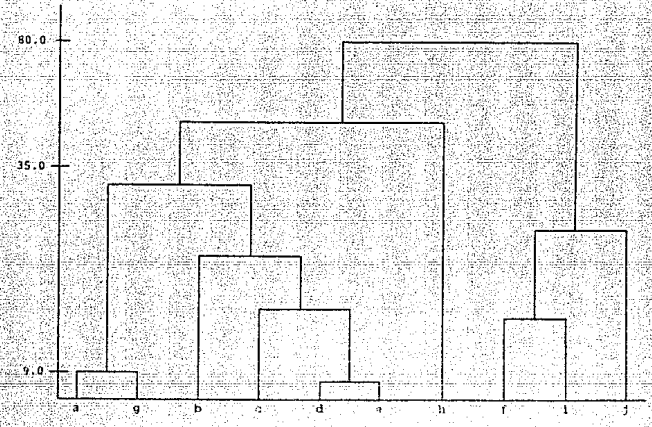


Gráfico 4.2: Cluster para las diez entidades utilizando la medida de distancia euclidiana

A una distancia de **35.00** unidades se observa la formación de 3 cluster el primero constituido por las entidades **ag** y **bcde**, el segundo por **fi** y **j** y el tercero por la entidad **h**.

MEDIDA DE DISTANCIA TIPO MANHATTAN

La siguiente medida a utilizar, es la tipo manhattan, con la cual se obtiene la matriz:

$$D = \begin{bmatrix}
0.00 & & & & & & & & & \\
16.00 & 0.00 & & & & & & & & \\
7.50 & 9.50 & 0.00 & & & & & & & \\
11.00 & 13.66 & 12.83 & 0.00 & & & & & & \\
14.33 & 15.33 & 17.16 & 12.66 & 0.00 & & & & & \\
15.66 & 20.00 & 17.83 & 21.66 & 20.00 & 0.00 & & & & \\
11.00 & 16.00 & 13.83 & 26.33 & 16.33 & 11.33 & 0.00 & & & \\
24.83 & 15.83 & 21.00 & 24.50 & 22.50 & 24.16 & 25.83 & 0.00 & & \\
23.00 & 24.33 & 26.50 & 24.00 & 29.00 & 13.33 & 21.33 & 26.16 & 0.00 & \\
26.50 & 24.50 & 30.00 & 24.50 & 30.50 & 17.83 & 28.50 & 21.00 & 21.83 & 0.00
\end{bmatrix}$$

Matriz 4.8: Distancias tipo manhattan

Partiendo de esta matriz se calcula la matriz de asociación correspondiente:

$$A = \begin{bmatrix}
94 & & & & & & & & & \\
-19 & 122 & & & & & & & & \\
81 & -28 & 134 & & & & & & & \\
55 & 37 & 49 & 139 & & & & & & \\
33 & 32 & -4 & 77 & 177 & & & & & \\
-17 & -80 & -38 & -107 & -66 & 115 & & & & \\
56 & 2 & 30 & -68 & 30 & 63 & 139 & & & \\
-117 & 80 & -14 & -86 & -20 & -90 & -120 & 286 & & \\
-69 & -86 & -140 & -70 & -101 & 116 & -19 & -50 & 296 & \\
-96 & -166 & -180 & -24 & -169 & 105 & -129 & 130 & 116 & 414
\end{bmatrix}$$

Matriz 4.9: Matriz de asociación

Con el eigenanálisis se obtienen la matriz Λ y la matriz U:

$$\Lambda = \begin{bmatrix}
658.03 & & & & & & & & & \\
0.00 & 508.99 & & & & & & & & \\
0.00 & 0.00 & 276.46 & & & & & & & \\
0.00 & 0.00 & 0.00 & 197.43 & & & & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 145.44 & & & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.76 & & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 35.03 & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 25.42 & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 10.51 & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.0
\end{bmatrix}$$

Matriz 4.10: Eigenvalores

Calculando la variación explicada para cada eje:

% variación eje 1 = 43.00 %
% variación eje 2 = 23.88 %
% variación eje 3 = 12.86 %
% variación eje 4 = 9.19 %
% variación eje 5 = 6.76 %
% variación eje 6 = 4.45 %
% variación eje 7 = 1.63 %
% variación eje 8 = 1.18 %
% variación eje 9 = 0.50 %
% variación eje 10 = 0.00 %

$$U = \begin{bmatrix}
0.21 & -0.19 & -0.25 & 0.26 & -0.03 & 0.03 & 0.58 & 0.54 & 0.22 & -0.38 \\
0.25 & 0.25 & 0.28 & -0.14 & -0.31 & -0.26 & -0.51 & 0.47 & 0.16 & -0.54 \\
0.32 & 0.05 & 0.06 & 0.38 & -0.36 & -0.31 & 0.15 & -0.45 & -0.41 & 0.20 \\
0.18 & 0.08 & -0.63 & -0.22 & -0.25 & 0.48 & -0.25 & -0.23 & 0.09 & -0.09 \\
0.29 & 0.04 & -0.13 & -0.41 & 0.66 & -0.28 & 0.06 & -0.01 & -0.27 & 0.22 \\
-0.22 & -0.35 & 0.16 & 0.23 & -0.13 & 0.40 & -0.28 & 0.26 & -0.56 & -0.05 \\
0.16 & -0.36 & 0.31 & 0.21 & 0.29 & 0.09 & -0.18 & -0.37 & 0.58 & 0.56 \\
-0.16 & 0.68 & 0.36 & 0.01 & 0.11 & 0.39 & 0.31 & -0.34 & 0.04 & 0.24 \\
-0.35 & -0.35 & 0.21 & -0.52 & -0.36 & 0.10 & 0.28 & -0.11 & 0.02 & -0.28 \\
-0.64 & 0.17 & -0.38 & 0.28 & 0.14 & -0.11 & -0.17 & -0.02 & 0.14 & 0.10
\end{bmatrix}$$

Matriz 4.11: Eigenvectores

Que permiten calcular la matriz C.

$$C = \begin{bmatrix}
6.3 & -4.4 & -4.2 & 3.7 & -0.6 & 0.3 & 3.4 & 2.7 & 0.7 & 0.0 \\
7.3 & 5.7 & 4.7 & -1.9 & -3.7 & -2.6 & -3.0 & 2.4 & 0.5 & 0.0 \\
9.4 & 1.1 & 1.1 & 5.4 & -4.3 & -3.3 & 0.9 & -2.3 & -1.3 & 0.0 \\
5.3 & 1.8 & -10.4 & -3.1 & -1.1 & 4.7 & -1.5 & -1.2 & 0.3 & 0.0 \\
8.7 & 0.9 & -2.3 & -8.3 & 7.9 & -2.7 & 0.4 & -0.1 & -0.4 & 0.0 \\
-5.5 & -7.9 & 2.8 & 3.3 & 1.6 & 3.9 & -1.7 & 1.3 & -1.8 & 0.0 \\
4.7 & -8.3 & 5.1 & 3.0 & 3.6 & 0.9 & -1.1 & -1.9 & 1.9 & 0.0 \\
-4.9 & 15.4 & 5.9 & 0.3 & 1.4 & 3.9 & 1.9 & -0.4 & 0.1 & 0.0 \\
-11.4 & -8.2 & 3.5 & -8.2 & -4.4 & -1.5 & 1.7 & -0.6 & 0.1 & 0.0 \\
-18.9 & 3.9 & -6.3 & 3.9 & 1.7 & -4.0 & -0.9 & -0.1 & 0.5 & 0.0
\end{bmatrix}$$

Matriz 4.12: Matriz C

Al obtener la transpuesta de C se obtiene C' que corresponde a la matriz de coordenadas principales.

$$\begin{bmatrix} c'_1 \\ c'_2 \\ c'_3 \\ c'_4 \\ c'_5 \\ c'_6 \\ c'_7 \\ c'_8 \\ c'_9 \\ c'_{10} \end{bmatrix} = \begin{bmatrix} 6.3 & 7.1 & 9.4 & 5.3 & 8.7 & -6.5 & 4.7 & -4.9 & -11.4 & -18.9 \\ -4.4 & 5.8 & 1.1 & 1.8 & 0.9 & -7.9 & -8.3 & 15.4 & -8.2 & 3.9 \\ -4.2 & 1.7 & 1.2 & -10.1 & -2.1 & 2.8 & 5.1 & 5.9 & 3.5 & -6.3 \\ 3.7 & -1.9 & 5.4 & -3.1 & -6.3 & 3.3 & 1.0 & 0.2 & -8.2 & 3.9 \\ -0.6 & -3.7 & -4.3 & -3.1 & -7.9 & 1.6 & 3.6 & 1.4 & -4.4 & 1.7 \\ 0.3 & -2.6 & -3.3 & 4.7 & -2.7 & 3.9 & 0.9 & 3.0 & -1.0 & -4.0 \\ 3.4 & -3.0 & 0.9 & -1.5 & 0.4 & -1.7 & -1.1 & 1.9 & 1.7 & -0.9 \\ 2.7 & 2.4 & -2.3 & -1.2 & -0.1 & 1.3 & -1.9 & -0.4 & -0.6 & -0.1 \\ 0.7 & 0.5 & -1.3 & 0.3 & -0.9 & -1.8 & 1.9 & 0.1 & 0.1 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

Matriz 4.13 : Coordenadas principales

Utilizando los dos primeros vectores de coordenadas para formar el gráfico, se representa el **66.88 %** de variación:

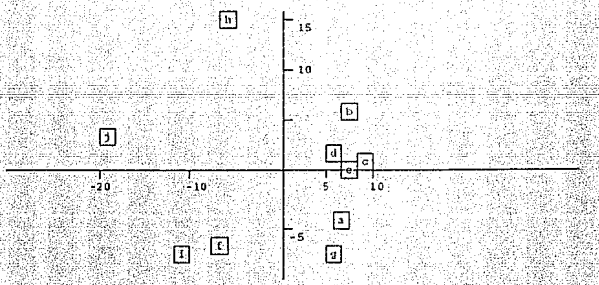


Gráfico 4.3: Coordenadas principales c'_1 y c'_2 utilizando la medida de distancia manhattan

Con el correspondiente cluster:

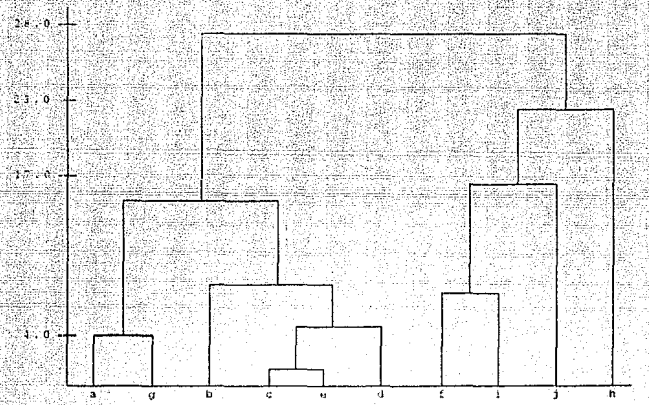


Gráfico 4.4: Cluster para las coordenadas principales utilizando la medida de distancia manhattan

En el cual se observa que a una distancia de **17.00** se forman tres cluster, uno con las entidades **ag** y **bced**, otro con **fi** y **j** y un tercero por la entidad **h**.

Si se compara el gráfico 4.2 con el 4.3 se observa que representan un agrupamiento de las entidades de manera similar, ya que se pueden formar los mismos grupos de entidades. La unica variación es en cuanto a los valores de distancia en los cuales se encuentran expresados cada uno de los gráficos.

COEFICIENTE DE CORRELACION PRODUCTO MOMENTO

Calculando ahora el coeficiente de correlación producto momento, se obtiene la siguiente matriz de similaridad:

$$S = \begin{bmatrix}
1.00 & & & & & & & & & \\
0.62 & 1.00 & & & & & & & & \\
0.97 & 0.74 & 1.00 & & & & & & & \\
0.71 & 0.66 & 0.61 & 1.00 & & & & & & \\
0.43 & 0.76 & 0.48 & 0.73 & 1.00 & & & & & \\
-0.50 & -0.54 & -0.52 & -0.25 & -0.17 & 1.00 & & & & \\
0.27 & 0.35 & 0.46 & -0.24 & 0.10 & 0.03 & 1.00 & & & \\
-0.48 & -0.43 & -0.53 & -0.45 & -0.14 & -0.14 & -0.34 & 1.00 & & \\
-0.50 & -0.31 & -0.52 & -0.10 & -0.42 & 0.30 & -0.41 & -0.36 & 1.00 & \\
-0.45 & -0.73 & -0.66 & -0.04 & -0.30 & 0.33 & -0.85 & 0.50 & 0.30 & 1.00
\end{bmatrix}$$

Matriz 4.14: Similaridad entre las entidades

Al calcular la matriz complemento se tiene:

$$D = \begin{bmatrix}
0.00 & & & & & & & & & \\
0.18 & 0.00 & & & & & & & & \\
0.03 & 0.26 & 0.00 & & & & & & & \\
0.29 & 0.34 & 0.39 & 0.00 & & & & & & \\
0.57 & 0.23 & 0.53 & 0.27 & 0.00 & & & & & \\
0.51 & 0.47 & 0.48 & 0.72 & 0.84 & 0.00 & & & & \\
0.73 & 0.65 & 0.54 & 0.76 & 0.90 & 0.90 & 0.00 & & & \\
0.52 & 0.58 & 0.48 & 0.55 & 0.87 & 0.86 & 0.66 & 0.00 & & \\
0.51 & 0.69 & 0.48 & 0.90 & 0.48 & 0.60 & 0.39 & 0.64 & 0.00 & \\
0.56 & 0.28 & 0.35 & 0.92 & 0.70 & 0.67 & 0.15 & 0.50 & 0.71 & 0.00
\end{bmatrix}$$

Matriz 4.15: Distancias para el coeficiente de correlación

Generando la matriz de asociación correspondiente:

$$A = \begin{bmatrix}
0.05 & & & & & & & & & \\
-0.02 & 0.02 & & & & & & & & \\
0.04 & -0.02 & 0.00 & & & & & & & \\
0.09 & 0.06 & 0.03 & 0.22 & & & & & & \\
-0.01 & 0.11 & -0.01 & 0.01 & 0.24 & & & & & \\
0.13 & 0.04 & 0.02 & 0.02 & -0.08 & 0.28 & & & & \\
-0.10 & -0.06 & -0.01 & -0.04 & -0.15 & -0.21 & 0.25 & & & \\
0.00 & -0.04 & -0.01 & 0.06 & -0.14 & -0.12 & 0.01 & 0.21 & & \\
-0.04 & -0.13 & -0.02 & -0.29 & 0.09 & -0.05 & 0.14 & 0.00 & 0.18 & \\
-0.05 & 0.04 & 0.00 & -0.21 & -0.04 & -0.01 & 0.19 & 0.05 & -0.08 & 0.14
\end{bmatrix}$$

Matriz 4.16: Asociación entre las entidades

Al realizar el eigenanálisis se obtienen las matrices:

$$\Lambda = \begin{bmatrix}
0.70 & & & & & & & & & \\
0.00 & 0.49 & & & & & & & & \\
0.00 & 0.00 & 0.38 & & & & & & & \\
0.00 & 0.00 & 0.00 & 0.31 & & & & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.30 & & & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.13 & & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.04 & & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00
\end{bmatrix}$$

Matriz 4.17: Eigenvalores

Valores con los cuales al calcular el porciento de variación explicada para cada eje se obtiene:

% variación eje 1 = 27.8 %
% variación eje 2 = 19.5 %
% variación eje 3 = 15.2 %
% variación eje 4 = 12.4 %
% variación eje 5 = 11.9 %
% variación eje 6 = 6.0 %
% variación eje 7 = 5.2 %
% variación eje 8 = 1.5 %
% variación eje 9 = 0.5 %
% variación eje 10 = 0.0 %

$$U = \begin{bmatrix} 0.21 & -0.08 & -0.06 & 0.19 & 0.12 & -0.05 & -0.15 & 0.67 & 0.55 & 0.32 \\ 0.18 & 0.00 & 0.08 & -0.47 & -0.03 & 0.03 & -0.71 & -0.32 & 0.14 & 0.31 \\ 0.05 & -0.05 & -0.08 & 0.80 & 0.04 & 0.10 & -0.21 & 0.43 & -0.80 & 0.32 \\ 0.40 & -0.61 & 0.14 & 0.13 & -0.40 & 0.36 & 0.30 & -0.18 & 0.00 & 0.33 \\ 0.23 & 0.40 & 0.63 & -0.17 & 0.28 & -0.19 & 0.35 & 0.00 & -0.05 & 0.28 \\ 0.32 & 0.32 & -0.68 & 0.03 & 0.26 & 0.10 & 0.23 & -0.27 & 0.00 & 0.34 \\ -0.60 & -0.11 & 0.11 & 0.05 & 0.42 & 0.54 & 0.02 & -0.09 & 0.11 & 0.29 \\ -0.17 & -0.47 & -0.06 & 0.20 & 0.22 & -0.69 & 0.01 & -0.25 & -0.06 & 0.32 \\ -0.22 & 0.46 & 0.09 & 0.57 & -0.46 & -0.08 & -0.19 & -0.15 & 0.01 & 0.29 \\ -0.24 & 0.04 & 0.20 & -0.56 & -0.46 & -0.14 & 0.22 & 0.19 & 0.09 & 0.31 \end{bmatrix}$$

Matriz 4.18: Eigenvectores

Calculando la matriz C se obtiene lo siguiente:

$$C =$$

0.17	-0.00	0.04	0.10	-0.07	-0.02	0.00	0.14	0.07	0.00
0.15	0.00	0.00	0.00	-0.02	0.01	-0.26	-0.07	0.01	0.00
0.04	-0.03	-0.04	0.09	0.01	0.03	-0.07	0.09	-0.10	0.00
0.33	-0.36	0.11	0.07	-0.22	0.14	0.11	-0.04	0.00	0.00
0.19	0.26	0.39	-0.09	0.15	-0.07	0.13	0.00	0.00	0.00
0.26	0.22	0.42	0.02	0.14	0.04	0.09	-0.06	0.00	0.00
-0.56	-0.07	0.07	0.12	0.23	0.21	0.01	-0.02	0.01	0.00
-0.14	-0.33	-0.03	0.11	0.12	-0.27	0.09	-0.05	0.00	0.00
-0.19	0.32	0.05	0.11	-0.25	-0.02	-0.07	-0.03	0.00	0.00
0.32	0.03	-0.13	0.00	-0.25	-0.05	0.12	0.04	0.01	0.00

Matriz 4.19: Matriz C

Apartir de la cual se obtiene la matriz de coordenadas principales C':

$$C' =$$

0.17	0.15	0.04	0.33	0.19	0.26	-0.56	-0.14	-0.19	-0.32
-0.00	0.00	-0.03	-0.36	0.26	0.22	-0.07	-0.33	0.32	0.03
0.04	0.00	-0.04	0.11	0.39	0.42	0.07	-0.03	0.05	-0.13
0.10	-0.26	0.09	0.07	-0.09	0.02	0.02	0.11	0.11	-0.30
0.00	-0.01	0.02	-0.22	0.15	0.11	0.23	0.12	-0.25	-0.25
-0.07	0.01	0.03	0.14	-0.07	0.04	0.21	-0.27	-0.02	-0.05
-0.05	-0.26	-0.07	0.11	0.13	0.09	0.01	0.00	-0.07	0.13
0.14	-0.07	0.03	-0.04	0.00	-0.06	-0.02	-0.05	-0.05	0.04
0.07	0.01	-0.10	0.00	0.00	0.00	0.01	0.00	0.00	0.01
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Matriz 4.20: Coordenadas principales

Utilizando los dos primeros vectores resulta la siguiente gráfica, que representa el 47.3 % de variación explicada:

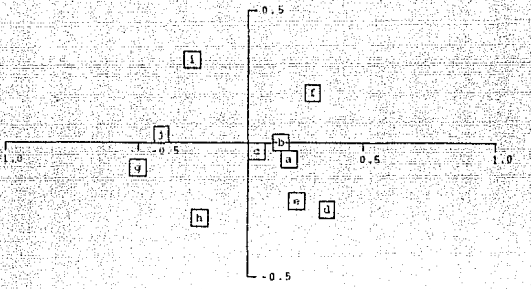


Gráfico 4.5: Coordenadas principales c_1' y c_2' utilizando el coeficiente de correlación producto momento

En el correspondiente cluster se observa la formación de tres grupos a una distancia de **0.24**, en los cuales el primero esta formado por las entidades **bcfj** y **eih**, el segundo por las entidades **dg** y el tercero por la entidad **a**.

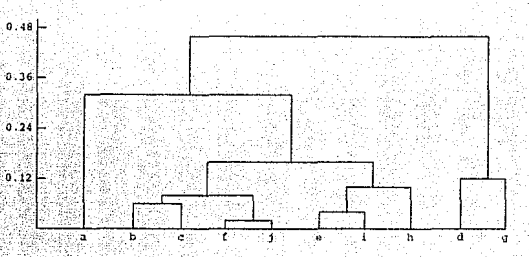


Gráfico 4.6: Cluster para las 10 entidades utilizando el coeficiente de correlación

Si se comparan los resultados obtenidos al aplicar las tres medidas de distancia, se encuentra que los cluster de las coordenadas principales calculadas por las medidas Euclidiana y Manhattan forman el mismo tipo de cluster, en cambio el Coeficiente de Correlación Producto Momento forma una agrupación diferente.

Esto se debe al planteamiento de una relación de distancia entre las entidades que debe preservarse tanto al inicio como en las coordenadas finales. Para cumplir esto, se establece una igualdad entre cualquier medida de distancia (utilizada para obtener la matriz **D** y la medida Euclidiana (Utilizada para obtener las coordenadas principales); cuando la medida utilizada inicialmente es la Manhattan o la Euclidiana las interpretación de los valores es semejante, por ejemplo un valor de cero indica que se trata de la misma entidad, a diferencia de cuando se utiliza el coeficiente de correlación producto momento en el cual los valores se interpretan de manera inversa, es decir un valor de cero indica que se tratan de dos entidades diferentes.

Por otra parte al obtener el complemento de los valores absolutos del coeficiente de correlación se obtienen valores de distancia, pero estos se encuentran en un rango menos amplio y por lo tanto se reduce la visualización de la estructura de los datos.

Otra de las restricciones de la técnica indica que las matrices **D** o **S** pueden tomar valores no-negativos, es decir los reales positivos y cero, situación que el coeficiente de correlación no cumple. Es por estas razones que no se recomienda la utilización de el coeficiente de correlación producto momento dado que se contrapone con diferentes aspectos de la técnica.

CAPITULO V: COORDENADAS PRINCIPALES UNA HERRAMIENTA PARA LA TAXONOMIA, DISTRIBUCION Y DESCRIPCION DE SUELOS.

Los datos que se utilizan en este capítulo derivan del estudio de la "Taxonomia, distribución y descripción de los suelos del Parque Nacional " El Chico ", Hidalgo, realizado por Romeo De Paz Colmenares.

Se analizó un conjunto de datos pertenecientes a seis perfiles de suelo, dichos datos corresponde a los parámetros físico-químicos determinados en el laboratorio. Con los resultados obtenidos se encontró la relación de similaridad entre los perfiles descritos con el propósito de contribuir a la determinación de la distribución de los suelos dentro del parque nacional "El Chico", en el estado de Hidalgo.

Se realizó el análisis de coordenadas principales con el propósito de obtener un gráfico que permitiera la identificación de las entidades (perfiles) similares y obtener el mapa de distribución de suelos.

La conservación de los suelos requiere del uso de cada unidad de terreno conforme a sus potenciales y estado en que se encuentre. Un estudio de suelo ayuda a definir sus características importantes, proporciona mapas de suelos que contienen muchos tipos de información, pero tal vez la de mayor valor es el tipo de suelo, pendiente y grado de erosión que se registra para cada área delineada sobre el mismo mapa.

Estos mapas son la base para desarrollar otros que indiquen la gran variedad de usos, y son necesarios en la programación de un plan de manejo o prácticas de conservación del suelo. Además, proporcionan un inventario y recuento de existencias de los recursos nacionales del suelo.

El parque nacional " El Chico " se encuentra situado en el extremo occidental de la sierra de Pachuca, forma el parteaguas que divide a la cuenca del Pánuco de la cuenca endorréica conocida como Valle de México. Pertenece a los municipios de Mineral del Chico y Mineral del Monte, ubicados en la parte sureste del estado de Hidalgo, a 26 Km de la ciudad de Pachuca, a los 20°13' latitud norte y los 98°44' longitud oeste.

Limita al norte con el poblado El Puente y Carboneras, al sur con los ejidos Estanzuela, El Cerezo y Pueblo Nuevo, y al oeste con terrenos de El Puente y Estanzuela.

La superficie total es de 2739 Ha. la altura máxima en la peña de las Ventanas a 3090 msnm, y la mínima a 2320 msnm, en el poblado Mineral del Chico. (SEDUE, 1988).

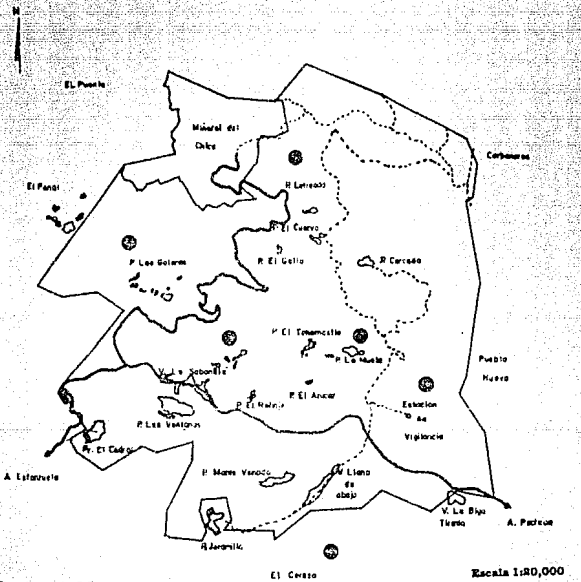
El clima según Köpen modificado por García (1970), es de tipo **C(m)b(i'')g** con las siguientes caracteristicas; templado húmedo con invierno benigno fresco y largo, con dos máximos de lluvias separados por estaciones secas, con régimen de lluvias de verano. La temperatura anual del mes más caliente se presenta antes del solsticio de verano,temperaturas máximas anuales entre 10° a 14°C y las mínimas de -6° a -9°C. La precipitación total varia de 600 a 1500 mm. La vegetación es de Bosque de Oyamel, de Encino, Tlaxcal y Pino.

Al estar situado sobre parte de la sierra de Pachuca, con rocas que afloran de procedencia volcánica, se encuentran suelos derivados de cenizas volcánicas o andosoles.

METODO

Primero se seleccionaron las zonas y rutas de muestreo (mapa 5.1), utilizando como criterios: homogeneidad de los perfiles realizados en anteriores ciclos de muestreo, tipo de vegetación, geología y características fisiográficas empleando mapas cartográficos e información fotográfica aérea.

Mapa 5.1. Ubicación de los Sitios de Muestreo.



Se realizó la descripción de los perfiles en los sitios seleccionados en base al criterio utilizado por Cuanalo (1981), y se tomaron muestras de cada horizonte para determinar los parametros físico-químicos en el laboratorio.

Dentro del análisis físico-químico se determinó textura (por el método de Boyoucos), densidad aparente (método de la probeta) y densidad real (picnómetro), pH potencial y activo (método electrométrico), materia orgánica (Walkey Black), CICT (EDTA), porcentaje de saturación de bases, humedad, bases totales y salinidad Condutividad electrica).

Posteriormente se utilizaron los datos obtenidos del análisis de laboratorio para realizar el análisis de coordenadas principales con el propósito de encontrar las similitudes entre los perfiles descritos y con apoyo de estudios realizados anteriormente poder elaborar un mapa de distribución de los tipos de suelos del parque.

Para el manejo de los datos en la técnica, todos los valores se expresan como variables cuantitativas con el proposito de evitar mezcla de tipo de variables y por lo tanto un mayor número de transformaciones, la textura se expresa como porcentaje de arena, limo y arcilla y no como clase textural.

RESULTADOS

Se ubicaron seis sitios de muestreo con los criterios antes mencionados, los cuales se indican en el mapa 5.1 (muestreo 90-1).

Con los datos obtenidos en campo, la descripción de campo y el análisis de laboratorio (que se muestran en el cuadro 5.1), se integró una descripción para cada perfil realizado, con lo cual se determinó el tipo de suelo al que pertenecían cada uno.

Se obtuvieron tres perfiles característicos de suelos andosólicos Húmicos (perfiles 1, 3 y 5), dos característicos de andosoles Ócricos (perfiles 4 y 2) y solo uno característico de andolosoles Vítricos (perfil 6).

Los resultados de los análisis de laboratorio se muestran en el cuadro 5.1.

Cuadro 5.1: Resultados de los análisis de laboratorio

Perfil	Horizonte	Textura			C.I.C.T.	%
		arcilla	limo	arena	me / 100 g	M.O.
1	A	9.44	33.28	57.28	36.37	21.37
1	B	21.44	17.28	71.28	28.79	10.33
1	C	10.72	24.00	65.28	29.30	7.30
2	A_1	6.72	36.00	57.28	36.06	21.55
2	A_2	10.72	23.28	66.00	28.06	22.41
2	C_1	3.44	26.56	70.28	31.30	10.33
2	C	7.44	31.28	61.68	32.38	10.51
3	A	7.44	27.28	65.28	28.51	20.86
3	C	2.16	20.56	79.38	32.02	9.13
4	A	5.44	22.56	72.00	30.54	19.99
4	AB	4.72	23.00	67.26	25.76	12.23
4	B	5.44	23.28	71.28	25.76	12.23
4	C	1.44	24.56	74.00	24.90	13.61
5	A	17.44	14.56	68.00	38.22	11.02
5	C	10.44	20.56	70.00	34.02	6.09
6	A	15.28	36.72	48.00	29.15	11.89
6	AB	22.00	36.72	41.28	29.79	12.76
6	B	19.28	63.44	17.28	32.23	10.51
6	C	15.28	24.72	60.00	31.15	6.26

Cuadro 5.1: continuación

D.A. g/c.c	D.R. g/c.c	% E.P.	Humedad (% H_2O)			pH		
			P. Tot.		H.H.	s-H_2O 1:1	s-KCl 1:1	s-KCl 1:2
0.59	1.74	66.09	87.61	85.17	7.16	4.9	4.8	4.85
0.79	2.06	61.65	85.96	84.17	6.95	5.0	4.7	4.80
0.93	1.46	37.16	71.23	67.45	4.36	4.5	4.2	4.30
0.77	1.83	56.83	83.43	74.72	6.28	4.4	5.0	5.10
0.90	1.45	51.72	78.09	64.15	6.96	4.3	4.8	4.90
0.98	2.00	55.45	76.50	62.77	8.77	4.2	4.8	4.90
0.94	2.01	53.92	91.46	81.60	9.41	4.3	4.8	4.80
0.68	1.77	61.58	0.0	0.0	0.0	4.9	5.0	5.10
1.05	3.40	57.66	68.24	54.10	11.36	5.7	5.8	5.50
0.95	1.86	48.92	66.57	58.33	8.74	4.7	4.9	5.00
1.95	1.98	46.96	65.07	62.94	12.16	4.3	4.6	4.65
1.01	2.29	55.49	74.14	64.90	10.86	4.3	4.5	4.60
1.09	1.82	50.01	71.72	56.14	8.94	4.3	4.7	4.75
0.61	1.58	61.69	72.20	57.36	7.14	5.0	4.8	4.80
0.80	1.78	55.04	69.43	48.27	7.84	4.4	4.6	4.20
0.94	1.45	49.19	81.62	56.05	4.98	5.2	5.0	5.00
1.03	1.90	47.45	90.09	62.24	7.57	4.8	4.3	4.45
1.05	1.97	46.70	68.67	45.35	3.46	5.0	4.1	4.10
0.95	1.65	42.42	54.97	44.40	3.73	4.3	3.7	4.10

Cuadro 5.1: continuación

Bases totales		V	Cationes extraibles (me/100g)				C.E.
$CO_3^{=}$	HCO_3^{-} 1 X 10^6	$\frac{me}{100\ g}$	Ca^{+2}	Mg^{+2}	K^{+}	Na^{+}	mmhos
0.0	2.85	8.52	0.37	0.86	1.09	0.74	0.39
0.0	2.56	13.18	0.91	2.11	1.30	0.79	0.24
0.0	3.06	23.14	0.78	3.59	1.26	0.93	0.46
0.0	6.13	25.10	1.07	5.78	1.02	0.78	0.93
0.0	8.36	14.18	0.10	1.65	1.20	0.71	0.20
0.0	5.22	9.78	0.19	0.91	1.18	0.72	0.26
0.0	2.99	16.49	0.28	2.32	1.13	0.74	0.42
0.0	0.0	34.75	1.32	5.27	2.53	0.77	0.55
0.0	6.13	15.28	0.86	1.88	1.84	0.79	0.27
0.0	3.06	14.74	0.80	1.16	1.82	0.75	0.34
0.0	6.13	10.95	0.78	0.42	1.46	0.71	0.24
0.0	6.13	8.21	0.24	0.23	1.04	0.75	0.16
0.0	3.06	9.42	0.76	0.41	1.35	0.76	0.26
0.0	3.06	17.29	0.21	1.20	2.55	0.79	0.57
0.0	9.19	15.45	0.68	0.75	1.28	0.76	0.27
0.0	3.06	14.76	0.19	1.25	1.61	0.91	0.34
0.0	6.98	9.31	0.19	0.24	1.56	0.79	0.17
0.0	9.19	10.97	0.24	1.05	1.41	0.83	0.40
0.0	6.13	15.59	0.57	2.00	1.49	0.79	0.28

Los datos de laboratorio se sometieron al análisis de coordenadas principales para corroborar la similutud de los perfiles y obtener la distribución de los tipos de suelo del Parque Nacional "El Chico", la información que proporciona el análisis es la siguiente:

EIGENVALORES						% λ c/eje
465.79						94.57 %
0.00	17.40					3.53 %
0.00	0.00	5.83				1.18 %
0.00	0.00	0.00	3.17			0.64 %
0.00	0.00	0.00	0.00	0.18		0.04 %
0.00	0.00	0.00	0.00	0.00	0.00	0.00 %

Cuadro 5.2: Porcentaje de variación explicada para cada eje

4.22	1.71	1.81	0.37	0.07	0.00
-18.54	-0.97	-0.21	0.33	-0.01	0.00
-0.35	-1.05	1.93	-0.01	0.22	0.00
7.82	-0.77	0.76	1.14	-0.17	0.00
0.29	0.98	0.14	-1.20	-0.26	0.00
6.56	-2.87	-0.21	-0.62	0.15	0.00

Cuadro 5.3: Matriz de coordenadas principales **C**

c'_1	=	4.22	-18.54	-0.35	7.82	0.29	6.56
c'_2		1.71	-0.97	-1.05	-0.77	0.98	-2.87
c'_3		-1.81	-0.21	1.93	0.76	0.14	-0.21
c'_4		0.37	0.33	-0.01	1.14	-1.20	-0.62
c'_5		0.07	-0.01	0.22	-0.17	-0.26	0.15
c'_6		0.00	0.00	0.00	0.00	0.00	0.00

Cuadro 5.4: Vectores de coordenadas principales

Si se utilizan los vectores c'_1 y c'_2 del cuadro 5.4 y se observa en el cuadro 5.3 cual es la contribución en cuanto a variación de dichos ejes,con lo cual se logra explicar el **98.10 %** de variación en los datos, porcentaje aceptable ya que al realizar el gráfico 5.1 no se sacrifica mucha información, se encuentra registrado casi la totalidad de la información

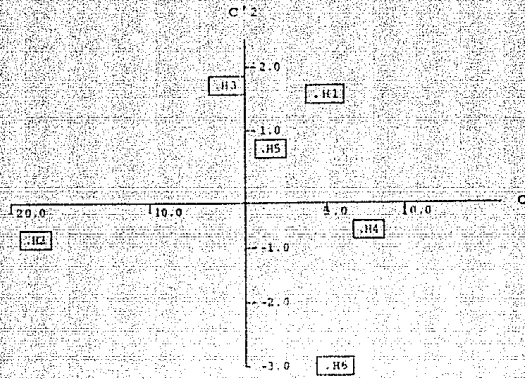
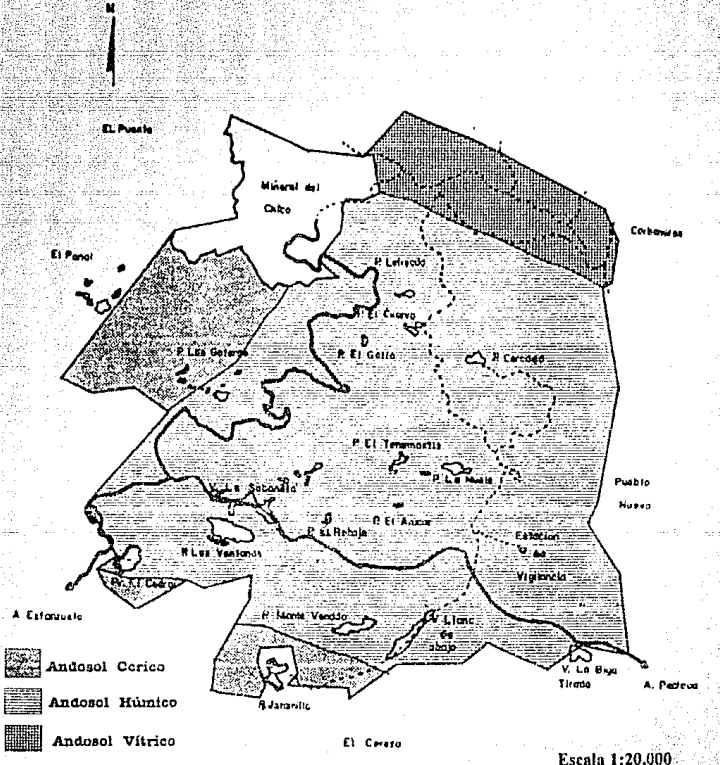


Gráfico 5.1: Coordenadas principales c'_1 y c'_2 para seis perfiles del Parque Nacional "El Chico"

Gráfico con el cual se corroboran las hipótesis planteadas en campo ya que se pueden agrupar los perfiles **1**, **3** y **5** como un grupo, los perfiles **2** y **4** como otro y finalmente el perfil **6**. Con esta información aunada a la recabada en estudios anteriores se obtuvo el mapa 5.2, que indica la distribución de suelos del parque nacional " El Chico ".

Mapa 5.2. Distribución de los Tipos de Suelo en El Parque Nacional "El Chico".



Escala 1:20,000

CAPITULO VI: INSTRUCCIONES DE MANEJO DE NTSYS-PC

NTSYS-PC es un programa de análisis estadístico mediante el cual se pueden realizar las operaciones y cálculos de matrices que el análisis multivariado requiere, además de la obtención de gráficos que son de gran ayuda para la interpretación de los resultados.

En el caso del análisis de coordenadas facilita la obtención de cálculos correspondientes a cada uno de los pasos de la técnica y proporciona los gráficos correspondientes en dos o tres dimensiones, éste último tiene la ventaja de poder rotar para poder observar mejor la estructura de los datos.

El manejo de NTSYS-PC no es de gran complejidad pero sí requiere de conocimientos básicos sobre el análisis multivariado, tales como tipo de matriz, medidas de distancia, etc.

Se pueden realizar los cálculos en cualquiera de los pasos de la técnica siempre y cuando se posea la matriz con el formato adecuado para realizarlos.

Lo primero que debe considerarse es la captura de los datos iniciales, éstos pueden capturarse dentro de NTSYS-PC o se pueden importar desde hojas de cálculo, siempre y cuando posean el formato adecuado. Esta operación se realiza en la opción **Files-Edit**, en pantalla aparece un campo en el cual se nombra la matriz que se desea generar o modificar, NTSYS-PC acepta cualquier extensión para el nombre del archivo, una vez que se ha proporcionado el nombre se traslada al campo de creación de la matriz con **F2**.

En dicho campo los primeros renglones corresponden a líneas de comentarios que permitan identificar la matriz, por ejemplo, la zona de estudio o el tipo de datos. Dichas líneas inician con comillas dobles ("), o comillas sencillas ('). la siguiente línea es de parámetros de identificación de la matriz, la cual consta de cinco campos:

a) Código del tipo de matriz:

1 = Matriz rectangular de datos.
2 = Matriz simétrica de distancia.
3 = Matriz simétrica de similaridad.
4 = Matriz diagonal.
5 = Matriz árbol para datos de distancia.
6 = Matriz árbol para datos de similaridad.
7 = Matriz gráfico para datos de distancia.
8 = Matriz gráfico para datos de similaridad.

b) Número de renglones (se adiciona una **L** si se desea que se encuentren etiquetados). En el caso de una matriz gráfico éste corresponde al número de **nodes**. Se recomienda que las entidades sean colocadas en los renglones.

c) Número de columnas (se agrega una **L** si se desea etiquetar las variables). Para una matriz gráfico corresponde al número de **edges**.

d) Codigo para valores marcados (**0** = ninguno). Este parámetro indica el número de datos no registrados.

e) Valor para el código **d**.

Las siguientes dos líneas son opcionales y corresponden a las etiquetas de los renglones y columnas respectivamente. Después se anotan los elementos de la matriz, con un formato libre de 255 caracteres por línea.

Por ejemplo: Si se tiene una matriz rectangular que contiene ocho entidades y cuatro variables correspondientes a un estudio de abundancia, en la cual se desea etiquetar del uno al ocho las entidades y no se tienen valores marcados, el formato de la matriz es el siguiente:

```
" Estudio de abundancia
" Ocho entidades con cuatro variables
" Matriz de datos iniciales
1 8L 4 0
1 2 3 4 5 6 7 8
MATRIZ DE DATOS
```

Donde:

```
" Estudio de abundancia
" Ocho entidades con cuatro variables
" Matriz de datos iniciales
```

Corresponden a los comentarios que identifican la matriz.

1 Indica el tipo de matriz (en este caso una matriz rectangular de datos.

8L Muestra el número de hileras de la matriz (la **L** indica que se encuentran etiquetados).

4 Corresponde al número de columnas, que no tienen etiquetas.

0 código de valores no registrados en la matriz de datos.

1 2 3 4 5 6 7 8 Son las etiquetas de los renglones.

MATRIZ DE DATOS Valores $\mathbf{X}_{i,j}$

Cuando se tiene la matriz de datos iniciales se procede a realizar los cálculos siguiendo cada una de las etapas de la técnica. De acuerdo con ésta, el orden para la realización de las operaciones es el siguiente.

Obtención de:

a) Matriz de distancia (**D**)

b) Matriz de asociación (**A**)

c) Eigenanálisis

c.i) Matriz de eigenvalores

c.ii) Matriz de eigenvectores

d) Matriz de coordenadas principales (**C**)

e) Gráfico

a) Matriz de distancia: Desde el menú principal se selecciona primero el tipo de medida que se desea utilizar. Cuando se tienen datos de frecuencia se utilizan los índices que se encuentran dentro del comando **SIMGEND**, para datos cuantitativos se utilizan los índices del comando **SIMINT**, el comando **SIMQUAL** se emplea para datos cualitativos.

En los tres comandos se procede de forma similar; primero se da el nombre de la matriz de datos iniciales, posteriormente que recibirá la matriz **D** de valores calculados, se indica la dirección en la cual se desean las operaciones (renglones en caso de entidades y columnas para variables). En este paso se puede verificar si se realizó la operación en el sentido adecuado verificando las dimensiones de la matriz cuadrada resultante (**D**), que deben corresponder al número de entidades presentes.

Si en la operación a) se obtuvieron valores de similaridad en lugar de valores de distancia se obtiene una matriz de similaridad (**S**), por lo cual se debe calcular la matriz complemento para poder obtener la matriz de asociación (**A**) y continuar con la técnica. El cálculo de la matriz complemento se realiza fuera del paquete.

b) Matriz de asociación: La siguiente etapa es crear la matriz de asociación (A), para lo cual se utiliza el comando **DCENTER** del menú principal, indicando nuevamente la matriz cuadrada (D) y el nombre de la matriz en que se van a guardar los valores calculados, además se le debe indicar la realización de los cálculos con la obtención de distancias al cuadrado.

c) Eigenanálisis: Una vez que se tiene la matriz **A** se realiza el eigenanálisis, seleccionando la opción **EIGEN** del menú principal; en esta parte se debe tener cuidado de seleccionar el vector de escalamiento adecuado.

d) Matriz de coordenadas principales: también es posible obtener la matriz **C** dentro de la opción **EIGEN**, para lo cual se modifica el vector de escalamiento.

e) Gráfico: En el gráfico se realiza la agrupación de entidades similares. Dicho gráfico se puede realizar en dos o tres dimensiones (opciones **Matrix Plot** y **3-D Model**, respectivamente), para obtener el gráfico se utiliza la matriz **C**, indicando el sentido de los vectores con respecto a las columnas de forma que se representan los vectores **c'**. Cuando el gráfico es de tres dimensiones se pueden hacer rotaciones para observar las entidades desde varios ángulos. Las rotaciones son en dos sentidos utilizando **s** para hacer girar el plano y **t** para rotarlo sobre de si mismo.

Cabe mencionar que en el mismo gráfico se puede realizar el cálculo de alguna técnica de cluster. Para dichas técnicas se utilizan los vectores **c'** del gráfico, con estos vectores se construye una matriz rectangular, de esta matriz se calcula una matriz simétrica de disimilaridad como punto de partida de las técnicas de cluster. El cálculo del cluster se realiza en el comando ***SAHN* Clustering**, para poder observar la representación gráfica del cluster se usa el comando **Tree Display**.

CAPITULO VII: CONCLUSIONES

El análisis de coordenadas principales es una herramienta útil en el estudio de comunidades ecológicas, porque permite la comparación de varias entidades a la vez, observándolas en un gráfico que muestra como se encuentran estructuradas en cuanto a similaridad. Proporcionando un gráfico de fácil interpretación en el cual se pueden agrupar entidades semejantes.

Esta es una técnica de tipo explorativa, que permite la identificación de grupos de entidades que presentan características semejantes con lo cual se puede rediseñar un estudio enfocado solo a aquellas entidades o grupos de ellas que son de interes para el investigador.

El investigador, al utilizar el análisis de coordenadas principales, tiene la ventaja de poder manejar diferentes índices de distancia así como de similaridad, de tal forma que puede buscar aquel que le permita mostrar de la mejor manera la agrupación natural de las entidades. Este aspecto es muy importante, ya que la mayoría de las veces se busca que los datos se ajusten a una determinada técnica estadística, relizando un gran número de transformaciones que pueden ocultar varios aspectos importantes del objeto de estudio. En contraste con este aspecto cuando se realiza un análisis estadístico se debe utilizar la herramienta adecuada que permita observar la mayor cantidad de información con el menor número de transformaciones estadísticas. Coordenadas principales al tener el potencial de uso de diferentes medidas de distancia, como de índices de similaridad, permite encontrar la mejor agrupación natural en los casos en que es posible su utilización.

Dentro del conjunto de técnicas que permiten comparar entidades coordenadas principales tiene un gran potencial de uso, dado que es posible obtener el porcentaje de variabilidad explicada representada en el gráfico final, además de poder conjuntarse con técnicas de cluster para una mejor visualización

de los grupos de entidades.

La ventaja que presenta sobre la utilización del análisis de cluster, es que permite registrar el porcentaje de variación implicado al analizar los datos, así como un gráfico en el cual es posible visualizar las relaciones de distancia entre las entidades.

Al respecto de otras técnicas multivariadas como componentes principales presenta la ventaja de analizar entidades y no variables, situación que es requerida en varios casos de estudios sobre comunidades ecológicas. Sobre las técnicas univariadas y bivariadas presenta una gran ventaja que es el manejo de toda la información registrada a la vez, característica que permite la obtención de conlusiones globales y no parciales acerca del caso de estudio.

El problema mayor que presenta se encuentra en la comprensión del fundamento matemático de la técnica, dado que es complicado el cálculo de cada uno de los pasos si no se cuenta con una herramienta computacional, pero esta dificultad se salva dado que la técnica se puede manejar con el paquete estadístico NTSYS-PC, donde con facilidad se pueden realizar los cálculos correspondientes a cada uno de los pasos y la obtención del gráfico, así como del cluster correspondiente.

Se puede decir que el análisis de coordenadas principales es una herramienta útil en casos de estudio biológicos en áreas como: Taxonomía (de suelos), piscícultura (ejem: comparación entre bordos), comparación de comunidades ecológicas (ejem: elaboración de mapas de vegetación) y en general a la determinación de grupos de respuesta semejantes.

Finalmente, para cada caso de estudio que sea necesario la realización de análisis estadístico se debe seleccionar la técnica adecuada de acuerdo con los objetivos de trabajo y las hipótesis planteadas, técnicas que muchas de las veces no son utilizadas de forma aislada sino en combinación con otras. En este caso se recomienda utilizar el análisis de coordenadas principales en conjunto con alguna técnica de cluster.

BIBLIOGRAFIA

Adam, P. 1978. Geographical variation in British Saltmarsh vegetation. **Journal of Ecology**. 66:339-336.

Bates, J. W., 1978. The influence of metal availability on the bryophyte and macrolichen vegetation of four rocks types on skye and rhum. **Journal of Ecology**. 66:457-482.

Buckman, H. O., 1966. **Naturaleza y propiedades de los suelos**. Fontaner y Simon S. A. Barcelona.

Buol, S. W., 1981. **Genésis y clasificación de suelos**. Trillas. México.

Cuanalo de la Cerda H., 1981. **Manual para la descripción de perfiles de suelo en campo**. Colegio de Postgraduados. Chapingo. México.

Davison, L. M., 1983. **Multidimensional scaling**. John Wiley & Sons. New York. USA.

FitzPatrick, E. A., 1985. **Suelos su formación, clasificación y distribución**. Continental. México.

García, E., 1981. **Modificaciones al sistema de clasificación climática de Köpen (para adaptarlo a las condiciones de la República Mexicana)**. UNAM. México.

Gauch, H. G., 1977. A comparative study of reciprocal averaging and other ordination techniques. **Journal of Ecology**. 65:157-174.

Gauch, H. G., 1982. **Multivariate analysis in community ecology**. Cambridge University Press. USA.

Gower, J. C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. **Biometrika**. 53, (3)(4):325-338.

Gower, J. C., 1967. A comparison of some methods of cluster analysis. **Biometrics**. 23(4):623-637.

Gower, J. C., 1971. A general coefficient of similarity and some of its properties. **Biometrics**. 87: 857-874.

Graybill, A. F., 1969. **Introduction to matrices with applications to statiscs**. Wadsworth Publishing Company, California. USA.

Green R. H., 1979. **Sampling Design and Statistical Methodos for Enviromental biologist**. University of Western Ontario. Ed. John Wiley and Sons. New York. USA.

Huntley, B. and H. J. B. Birks. 1979. The past and present vegetation of the Morrone Birkwoods National Nature Reserve, Scotland. I. A primary phytosociological survey. **Journal of Ecology**. 67:417-446.

Howard, P. J. A., et all., 1988. Classification and dissection of environmental data using qualitative and mixed data types. **Journal of Environmental Management**. 26: 331-319.

Kempton, R. A. and P. G. N. Digby., 1987. **Multivariate analysis of ecological communities**. Chapman and Hall. New York. USA.

Kendall, M., 1982. **Multivariate analysis**. MacMillan Publishing Co.Inc. New York. USA.

Legrende L. and P. Legrende., 1983. **Developments in enviromental modelling 3, Numerical Ecology**. Elsevier Scientific Publishing Company. New York. USA.

Lowerre, J. M., 1983. An introduction to modern matrix methods and statistics. **The American Statistician**. 36 (2): 113-115.

Mandel J., 1982. Use of the singular descomposition in regression analysis. **The American Statistician**. 36 (1): 15-24.

Manly, F. J. B., 1986. **Multivariate statical methods**. Chapman and Hall. New York. USA.

Ocegueda, C. S., 1991. El análisis estadístico Cluster, métodos y aplicaciones en Biología. ENEP Zaragoza. UNAM, México. **Tesis de licenciatura**.

Overall, J. E. and C. J. Klett. 1972. **Applied multivariate analysis**. McGraw-Hill Book Company. USA.

Panchen, L. A., 1992. **Classification, evolution and nature biology**. Cambridge University Press. New York. USA.

Pielou, E. C., 1977. **Mathematical ecology**. A Wiley - Interscience Publication. John Wiley & Sons. New York. USA.

Pielou, E. C., 1984. **The Interpretation of ecological data. A primer on classification and ordenation**. A Wiley-Intercience Publications. USA.

Prentice I. C., 1977. Non-metric ordination methods in ecology. **Journal of Ecology**. 65:85-94.

Rivera, G. P., 1991. El análisis de distancia como una herramiente para la investigación biológica. ENEP Zaragoza. UNAM. México. **Tesis de licenciatura**.

Searle, S. R., 1976. **Matriz algebra for the Biological (including applications in statistics)**. John-Wiley & sons. New York. USA.

Williamson M. H., 1978. The ordination of incidence data. **Journal of Ecology**. 66:911-920.

Zarate de Lara, G. F. y M.O. Alvarez., 1985. Aplicaciones de las descomposiciones singular y espectral de una matriz. **Agrociencia**. 61:103-126.