16

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
"ZARAGOZA"

FALLA DE ORIGEN

ESTUDIO DEL ANALISIS ESTADISTICO DE CORRESPONDENCIAS Y SU APLICACION A LAS CIENCIAS BIOLOGICAS.

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 :

MARTIN DANIEL GUTIERREZ LEYVA

DIRECTOR DE TESIS:

M. EN C. ARMANDO CERVANTES SANDOVAL

MEXICO, D. F. 1995

U N A M
FES
ZARAGOZA

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

A MIS PADRES

DANIEL GUTIERREZ MEDINA

GLORIA LEYVA DE GUTIERREZ

A MIS HERMANOS

ROGELIO

ERNESTO

JULIETA

GABRIELA

A MIS SOBRINOS

JORGE

FERNANDO

GLORIA

KAREN

JESSICA

ERNESTO

MONICA

CON EL DESEO DE QUE ELLOS LOGREN MAYORES METAS

AGRADEZCO AL M. EN C. ARMANDO CERVANTES SANDOVAL SU IMPORTANTE DIRECCION Y COLABORACION EN EL DESARROLLO DE ESTA TESIS. ASI COMO SU PACIENTE DEDICACION PARA CONMIGO.

AGRADEZCO AL BIOL. GERMAN CALVA VAZQUEZ, BIOL. J. SALVADOR HERNANDEZ AVILES, BIOL. DAVID N. ESPINOSA ORGANISTA Y M. EN C. MA. JOSE MARQUEZ DOS SANTO POR LA REVISION DE LA TESIS, POR SUS OBSERVACIONES Y CONSEJOS.

I N D I C E

CAPITULO I.

INTRODUCCION.

En el análisis estadístico multivariado existe un conjunto de técnicas llamadas de ordenación que se encargan de resumir las principales tendencias de variación de un conjunto grande de datos, ordenándolos en uno o varios ejes de forma que sea posible describir su estructura interna. Entre estos métodos se encuentra el análisis de correspondencias.

El análisis de correspondencias ordena la información que contiene una tabla de contingencia de tamaño n*m, donde las hileras representan muestras y las columnas señalan la medición de varias características en esas muestras, en gráficas donde hileras y columnas se representan como puntos. Esta gráfica conserva gran cantidad de información relativa a los datos originales. De esta manera es posible describir la estructura interna de los datos y obtener información útil acerca del fenómeno que se está estudiando.

En la investigación biológica se realizan estudios donde se manejan un gran número de variables medidas sobre un extenso grupo de entidades de muestreo. Las observaciones que se obtienen de estos estudios generan un gran volúmen de datos que deben analizarse en forma conjunta para explicar el comportamiento de un fenómeno. La aplicación del análisis de correspondencias resulta ser, entonces, una técnica adecuada para estas situaciones, ya que permite analizar todas las variables de manera simultánea, requisito importante en Biología y que, además, garantiza la objetividad de los estudios.

Dado el gran potencial de aplicación que presenta el análisis de correspondencias para el estudio y descripción de los fenómenos biológicos, este trabajo es de gran interés para el

Biólogo en la aplicación de esta técnica, y en un futuro no muy lejano pueda servir como una base sólida para expresar los resultados mucho mejor que un informe general. Con este fin se proponen los siguientes objetivos.

OBJETIVO GENERAL

Revisar y analizar el fundamento teórico de la técnica del análisis de correspondencias describiendo cuales son sus propiedades más importantes, como herramienta de análisis estadístico multivariado, y mostrar su aplicación en las diferentes áreas de la investigación biológica.

OBJETIVOS PARTICULARES

- Conocer la técnica del análisis de correspondencias para analizar y describir su fundamento teórico.

- Proponer una estrategia para efectuar el análisis y descripción de los fenómenos que se presentan en la investigación biológica, aplicando el análisis de correspondencias a un estudio de caso, con la ayuda del programa CORAN.

- Describir sus propiedades más importantes y su limitación para su aplicación a las diferentes áreas de investigación biológica.

Para cumplir estos objetivos se presentan siete capítulos, además del presente, que muestran los diferentes aspectos que rodean a esta técnica de correspondencias. De manera que su contenido queda estructurado de la siguiente forma.

En el capítulo dos se da una visión general de la importancia del análisis de correspondencia como parte de la estadística descriptiva multidimensional en la investigación biológica.

En el capítulo tres se presenta el análisis teórico de la técnica del análisis de correspondencias.

El capítulo cuatro está dedicado a la descripción de las ayudas para la interpretación del análisis de correspondencias.

El capítulo cinco se dedica a la presentación del programa CORAN y a la descripción de los parámetros necesarios para su manejo.

El capítulo seis muestra un estudio de caso en el que se analiza la aplicación del análisis de correspondencias.

En el capítulo siete se discute la importancia del análisis de correspondencias en el análisis y descripción de los datos que se obtienen en una investigación biológica.

Por último, en el capítulo ocho se presentan las conclusiones del trabajo.

CAPITULO II.

LA BIOLOGIA Y EL ANALISIS ESTADISTICO MULTIVARIADO.

En la ciencia los conceptos matemáticos ocupan un sitio relevante, ya que para la descripción y entendimiento de los fenómenos naturales y sociales se recurre cada vez, y con mayor frecuencia, a la matemática y su forma de conceptualizar los fenómenos. El camino no es nuevo, nace con las antiguas civilizaciones; en Grecia, con Platón, durante una exposición entre sus contemporáneos ya se discutía la visión de un universo organizado sobre la base de principios matemáticos.

El desarrollo de las matemáticas ha sido multifacético; abarcando la estadística descriptiva multivariada, rama de la matemática que se ocupa de examinar numerosas variables simultáneamente. La aplicación del análisis estadístico multivariado a la Biología ha revolucionado no sólo las estrategias de investigación, sino la propia interpretación de los fenómenos bajo estudio.

En Biología, la mayor parte de los fenómenos tienen la influencia de muchos factores causales, incontrolables en su variación y muy a menudo vinculados estrechamente. La estadística multivariada permite medir tales fenómenos y averiguar la realidad de las diferencias existentes. La información que se obtiene en estos estudios es el resultado de múltiples observaciones y mediciones realizadas en campo, que se traducen en grandes volúmenes de datos. Probablemente los investigadores estén familiarizados con la presentación de los datos en forma de tablas, conocidas en término más general como matrices de datos, compuestas de hileras y columnas. Por ejemplo, la matriz de datos que comúnmente se presenta en los estudios de ecología de comunidades son tablas que muestran la frecuencia de varias

especies en cada una de las entidades de muestreo. De este modo, existen dos formas para representar la matriz de datos: cada hilera indica las diferentes especies y cada columna una unidad de muestreo, o viceversa las hileras indican las unidades de muestreo y las columnas las diferentes especies.

La interpretación de estas tablas de datos conduce al análisis de un fenómeno de manera más objetiva. En estas matrices de datos, que generalmente son muy extensas, no es fácil percibir su estructura. Pielou (1984), explica que la estructura interna de una matriz de datos puede ser cualquier modelo sistemático que indique, por ejemplo, ciertos grupos de especies que se presenten juntos, o que las unidades de muestreo, cuando son ordenadas apropiadamente, pueden exhibir una tendencia continua en su composición de especies.

En la estadística descriptiva multivariada se encuentra un conjunto de técnicas de ordenación que permiten agrupar un gran conjunto de datos a lo largo de uno o varios ejes y con ello dar origen al establecimiento de la estructura de la matriz de datos. El término *ordenación* fué introducido y definido por Goodall en 1954 como un *arreglo de unidades en un orden uni o multi-dimensional*. Entre estas técnicas de ordenación se encuentra el análisis de correspondencias, que aunque se ha utilizado muy poco en Biología, su principal aportación se presenta en el campo de la ecología de comunidades, principalmente vegetal.

El análisis de correspondencias es una técnica estadística multivariada que describe las relaciones existentes entre las hileras y columnas de una matriz de datos no negativos, que se conceptualiza, en términos generales, como una matriz con el formato de *entidades*x*atributos*; esta representación es muy útil para resumir resultados de los estudios que se llevan a cabo, en un subespacio dimensional. El producto final es una gráfica de dos dimensiones donde las entidades y columnas de la matriz de

datos se representan como puntos. En esta gráfica es posible identificar las entidades o atributos similares al encontrarse muy cerca entre sí, y entidades disimilares cuando se encuentran muy alejados entre ellos. Después de obtener tal arreglo espacial, es posible identificar grupos de entidades, o intentar reconocer comportamientos similares de diferentes atributos, e incluso caracterizar subgrupos de entidades a partir de algún atributo en particular.

Existen dos conceptos espaciales muy importantes en los que se explica el concepto y propósito fundamental de la técnica de ordenación. 1)los datos de una matriz *entidades*atributos* se pueden conceptualizar como un espacio de entidades donde dichas entidades son los ejes del espacio multidimensional y los atributos son los puntos localizados por las frecuencias de cada entidad; 2)existe un segundo espacio, a partir del cual los atributos son los ejes y las entidades son los puntos localizados a lo largo de estos ejes de acuerdo a la importancia de la entidad en cada atributo.

Ambos espacios poseen información suficiente para describir la matriz de datos; sin embargo, una representación gráfica que cuente con un espacio dimensional mayor a tres ejes, resulta imposible de interpretar. El análisis de correspondencias permite proyectar ambos espacios en un subespacio de dos dimesiones, en el que es posible resumir satisfactoriamente la dispersión de los puntos y permite una representación geométrica sencilla e interpretable.

Para conjuntar ambas nubes de puntos, el análisis de correspondencias parte de dos matrices de inercia, cada una de ellas contiene información de las hileras y columnas de la tabla de datos. Utiliza, además, la métrica *ji-cuadrada*; esta métrica permite obtener la representación simultánea de las dos nubes de puntos, ya que la traza de ambas matrices reflejan la distribución *ji-cuadrada* que prueba la hipótesis de independencia

de las hileras y columnas de la tabla de datos original. La desventaja de un subespacio en dos dimensiones es que cierto grado de fidelidad de la estructura de los datos se sacrifica; sin embargo, se cuenta con una representación de los datos muy accesible y sencilla.

Algunas de las fórmulas implicadas en el análisis de correspondencias aparecen dentro de los trabajos de Fisher (1940). El tratado de estadística de Kendall y Stuart (1961) evoca *el análisis canónico de las tablas de contingencia*, que viene a efectuar lo esencial de las operaciones del análisis de correspondencias para calcular en definitiva los parámetros destinados a probar la hipótesis de independencia de las líneas y columnas de la tabla. Pero hace tan solo diez años que Jean-Paul Benzécri pusó en evidencia las propiedades algebraicas del método, además de describir como los datos se apartan de esta hipótesis a través de la asociación existente entre las líneas y columnas de la tabla (Volle, 1985).

El análisis multivariado es una herramienta que se emplea frecuentemente en trabajos de ecología de comunidades. Sin embargo, su aplicación no se ve restringida a esta área, es posible que se involucre en el uso y administración del suelo, en los estudios de la contaminación, agricultura y pesca, siendo sólo algunos ejemplos; por lo que a continuación se describen algunas aplicaciones.

La clasificación de la tierra en categorías apropiadas para diferentes usos requiere de un análisis muy extenso, ya que existen numerosos factores que toman parte en esta tarea. El análisis multivariado es el indicado, a causa de que los numerosos parámetros individuales, considerados de manera aislada, carecen de importancia además de que muchos de ellos necesitan ser considerados en forma conjunta. La aplicación del análisis multivariado en la toma de decisiones del manejo de la

tierra es muy amplio, por ejemplo a través del análisis de la vegetación es posible clasificar el habitat de los bosques con el proposito de un manejo mejor, orientado a su aprovechamiento y preservación.

Además de la clasificación y manejo de la tierra, el análisis multivariado se puede aplicar a los estudios concernientes al medio lacustre y marino; por ejemplo, se puede utilizar para clasificar los lagos de acuerdo a su composición de organismos presentes y a los factores ambientales; la clasificación del habitat de los peces también puede ser útil para un sistema de explotación.

La ordenación de la vegetación es otra de las posibles aplicaciones del análisis multivariado ya que puede identificar los factores ambientales e históricos más importantes de una región en estudio. Tales resultados proporcionan un cuadro general para el estudio del impacto de la distribución que puede tener una especie en particular en varios tipos de vegetación. El manejo de la población de esta especie y su localización podría ser necesaria para reducir la destrucción de comunidades herbáceas y del suelo.

La contaminación atmosférica es un problema complejo que involucra numerosas variables. Por ejemplo, el análisis multivariado podría ayudar a identificar los principales factores que afectan en la contaminación atmosférica; analizar la concentración de elementos en aerosol para poder identificar los orígenes del aire contaminado.

El habitat lo constituyen la totalidad de factores que ejercen efecto en la presencia de un conjunto de especies, incluidos los bióticos y abióticos, cada especie se encuentra en un rango distintivo del habitat. Las variables físicas y bióticas con las que las especies se relacionan definen el nicho de esas especies. Los conceptos de habitat y nicho son básicos en la

teoría de la ecología de comunidades para comprender la organización de las especies. El análisis multivariado de una matriz de datos muestras*especies, en el que se compara la relación existente entre las distintas especies registradas en un gran número de muestras o localidades podría ayudar a aclarar las relaciones del habitat; el análisis de una matriz de datos muestras*variables, en el que es posible comparar la relación existente entre las muestras o localidades caracterizadas por un gran número de variables que pueden ser físicas y bióticas revelaría las características principales del nicho.

CAPITULO III.

TEORIA DEL ANALISIS DE CORRESPONDENCIAS.

El análisis de correspondencias se utiliza para estudiar tablas de contingencia de tamaño n*m, donde se comparan hileras con columnas. En el análisis de una tabla de contingencia se prueba la hipótesis nula de independencia entre sus hileras y columnas. Esto se retoma en el análisis de correspondencias al cuantificar esa independencia con ayuda de la métrica *ji-cuadrada*, y con ello dar origen a un criterio de asociación entre hileras y columnas similares. Ambas nubes de puntos: hileras y columnas, se encuentran en un espacio multidimensional $\mathbb{R}^m$ y $\mathbb{R}^n$ respectivamente. Desde el punto de vista geométrico, las proyecciones ortogonales de un punto en el espacio multidimensional, se consideran sus coordenadas. A partir de esto, es posible extraer dos de sus ejes principales y construir un subespacio de dos dimensiones. Dado que ambas nubes de puntos expresan el mismo porcentaje de variación explicada, este hecho se refleja en los ejes principales. Así, es posible seleccionar los ejes principales que expresen la mayor variación explicada y construir un solo gráfico, donde las dos nubes de puntos serán representadas. Esta gráfica es útil para obtener información de la tabla de datos, tal como la identificación de un grupo de muestras, o intentar reconocer comportamiento similar de diferentes elementos. También se puede caracterizar subgrupos de muestras a partir de una variable en particular. Al reconocer variabilidad entre los datos es posible identificar la razón de esta variabilidad, y mostrar la existencia de una o varias causas responsables de tal estructura de los datos.

Antes de exponer la teoría del análisis de correspondencias se consideran algunas propiedades geométricas de la matriz de datos en la sección 3.1; posteriormente se describe la técnica.

considerando cuatro puntos muy importantes: en la sección 3.2 se define la nube de puntos en $\mathbb{R}^m$ y $\mathbb{R}^n$ de la matriz de datos, en la sección 3.3 se define la métrica *ji-cuadrada*, la inercia total para la nube de puntos en N(i) y N(j) se expresa en la sección 3.4, el cálculo de los ejes principales y coordenadas para la nube de puntos en N(i) y N(j) se presenta en la sección 3.5. Para finalizar el capítulo, la sección 3.6 presenta los elementos suplementarios, que son útiles para enriquecer la interpretación de resultados.

3.1 Propiedades geométricas del campo de observación, matriz de datos.

La información que contiene cada una de las hileras de la matriz de datos se puede expresar vectorialmente. Un vector es la unidad algebraica y geométrica básica de un espacio multidimensional, y se refiere a un conjunto de números reales (Pichardo, 1990). Entonces, la información de cada hilera se expresa por el vector a_i, la letra minúscula a remarcada muestra que es un vector columna, si el mismo arreglo de elementos que componen este vector se escribe como una hilera, se forma un vector hilera y se denota como a^t. Este vector puede proyectarse como un punto en un espacio euclideano cuya dimensión, y por supuesto la posición de este punto, está dada por el número de componentes que presente el vector a_i. Así, en el caso de tres dimensiones, el vector a_i se representa por tres números α_{i1}, α_{i2} y α_{i3}, donde α_{i1} es la coordenada asociada a la proyección ortogonal del vector a_i sobre el primer eje; α_{i2} es la coordenada asociada con la proyección de a_i sobre el segundo eje; y α_{i3} es la coordenada asociada a la proyección de a_i sobre el tercer eje. Cuando las hileras de una tabla de datos representan entidades, éstas son caracterizadas por m variables; de tal suerte, las entidades se pueden proyectar en el espacio dimensional dado por el número de variables que estén presentes, (Figura 3.1).

Cuando se tienen tres muestras y se desea analizar las variables, se puede usar la representación geométrica que se muestra en la figura 3.2; la información de cada columna se expresa vectorialmente, b_j, de manera que las variables se proyectan en el espacio de las muestras.

De las figuras 3.1 y 3.2 es importante reconocer la forma de la nube de puntos y observar los grupos que pueden aparecer, lo cual conduce a una mejor interpretación de los datos contenidos en la tabla de datos. Esta interpretación es moderadamente fácil de realizar en diagramas de 2 ó 3 dimensiones; cuando se está en un espacio multidimensional resulta imposible llevar a cabo dicha tarea. El análisis de correspondencias reduce el número de dimensiones para cada una de esas nubes de puntos, recuperando sus características esenciales de ambas nubes, representándolos en una gráfica de dos dimensiones, donde se proyectan muestras y variables, como consecuencia de la relación de dualidad que existe entre ambos subespacios.

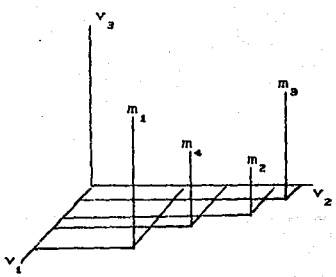


Fig 3.1. Representación de las muestras en el espacio de las variables.

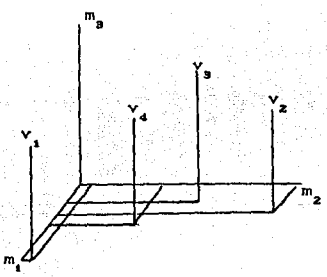


Fig 3.2. Representación de las variables en el espacio de las muestras.

3.2 Definición de la nube de puntos en $\mathbb{R}^m$ y $\mathbb{R}^n$ que contienen la información de la matriz de datos.

Las observaciones iniciales en los estudios biológicos se expresan frecuentemente en forma de una matriz de datos denotada por K, la letra mayúscula indica una matriz completa, que contiene dos descriptores i y j que poseen n y m modalidades respectivamente, donde k_{ij} representa valores no negativos. Las hileras de la matriz, por lo tanto, indican variación en la representación de la i-ésima entidad sobre las m variables examinadas, mientras que las columnas indican la composición de la j-ésima variable con respecto a las n entidades observadas.

$$K = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1m} \\ k_{21} & k_{22} & \cdots & k_{2m} \\ \vdots & \vdots & & \vdots \\ k_{n1} & k_{n2} & \cdots & k_{nm} \end{bmatrix}$$

La suma o total de los elementos de cada una de las hileras se denota como

$$k_{i\cdot} = \sum_{j=1}^{m} k_{ij}$$

y la suma de las columnas como

$$k_{\cdot j} = \sum_{i=1}^{n} k_{ij}$$

mientras que el total de los elementos de K se denota como

$$k_{\cdot\cdot} = \sum_{i=1}^{n} \sum_{j=1}^{m} k_{ij}$$

La información de la tabla contiene, por lo general, datos que pueden presentar una amplia variación, o estar contenidos dentro de una muy pequeña; así que, si los componentes de los vectores a_i en $\mathbb{R}^m$ se construyen con las frecuencias estadísticas de K, es decir los k_{ij}, sin una previa transformación, ocasiona que la distancia entre los puntos sea muy grande o pequeña, y que la nube de puntos quede muy dispersa. Lo anterior causa que la interpretación de los datos sea poco útil, ya que en la gráfica solo es posible analizar la composición de cada entidad, sin poder comparar las entidades en conjunto. Así, es necesario construir una base ortonormal en la que se puedan proyectar los puntos, sin alterar su estructura; para ello se toma en cuenta la importancia de cada una de las variables que describen a las diversas entidades, esto es, estudiar los perfiles de distribución de las variables dentro de cada entidad.

Similarmente, en el espacio $\mathbb{R}^n$ donde los vectores b_j están dados por las columnas, el interés es analizar la distribución de las muestras dentro de cada variable.

De modo que se define la matriz de frecuencias relativas como:

$$P = \left[\langle P_{ij} \rangle \right] = \frac{k_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} k_{ij}}$$

que puede expresarse como una tabla.

El caso que se presenta en la tabla de la figura 3.3 (línea i, columna j) contiene la frecuencia relativa P_{ij}. Los márgenes de la tabla contienen los totales hilera y los totales columna, la suma total de la tabla es igual a uno.

Así, el total de la i-ésima hilera de P se representa por

$$P_{i.} = \sum_{j=1}^{m} P_{ij} \qquad i = 1,2,\ldots,n$$

El caso del j-ésimo total columna de P se representa por

$$P_{.j} = \sum_{i=1}^{n} P_{ij} \qquad j = 1,2,\ldots,m$$

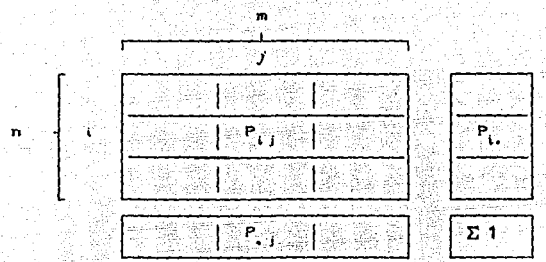


Figura 3.3. Matriz de frecuencias relativas P, de n hileras y m columnas.

Los totales hilera de P conforman el vector r, mientras que los totales columna constituyen los elementos del vector c. De estos dos vectores se definen dos matrices diagonales: Dr y Dc.

$$r' = (P_{1.} \quad P_{2.} \quad \ldots \quad P_{i.}) \qquad i = (1, 2, \ldots, n)$$

$$c' = (P_{.1} \quad P_{.2} \quad \ldots \quad P_{.j}) \qquad j = (1, 2, \ldots, m)$$

$$Dc = \begin{bmatrix} P_{\cdot 1} & & & \\ & P_{\cdot 2} & & \\ & & \ddots & \\ & & & P_{\cdot j} \end{bmatrix} \qquad Dr = \begin{bmatrix} P_{1\cdot} & & & \\ & P_{2\cdot} & & \\ & & \ddots & \\ & & & P_{i\cdot} \end{bmatrix}$$

El vector a_i representa un punto en el espacio dimensional $\mathbb{R}^m$, el cual se le llama perfil hilera. Así, los perfiles hilera constituyen una nube de n puntos, que se denota como $N(I)$, en el espacio dimensional dado por las m columnas de la matriz de datos. El perfil hilera a_i se expresa de la siguiente forma

$$a_i = \left[\frac{P_{i1}}{P_{i\cdot}}, \frac{P_{i2}}{P_{i\cdot}}, \ldots, \frac{P_{ij}}{P_{i\cdot}} \right] \qquad j = 1,2,\ldots,n$$

Significa que la i-ésima hilera de P se pondera por el reciproco de su total $P_{i\cdot}$, indicando con esto el peso de cada una de las variables j dentro de la muestra i.

Los perfiles hilera a_i pueden obtenerse a partir de la matriz A, dada por

$$A = Dr^{-1}P$$

donde Dr^{-1} es el inverso de la matriz diagonal de los totales hilera de la matriz de frecuencias relativas, de manera que sus elementos diagonales son las ponderaciones que se dan a cada una de las hileras de la matriz P. Así, las hileras de la matriz A definen los perfiles hilera a_i.

Cada entidad, representada por el vector a_i, el perfil hilera, parte del origen y para cualquiera de ellos la suma de sus coordenadas es igual a uno, lo cual implica una dependencia lineal entre sus coordenadas. Existe, entonces, una base estándar donde se representan los vectores a_i. Este conjunto de puntos se encuentra en un subespacio de m-1 dimensiones, y se le dá el nombre de distribución "*simplex*", figura 4, de modo que

$$\sum_{j=1}^{m} a_i = 1$$

$$a_{ij} \geq 0$$

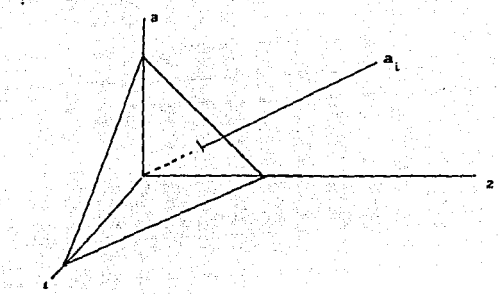


Figura 3.4. Proyección de los perfiles hilera en el "*simplex*" representado en tres ejes.

Los totales columna de la matriz de frecuencias relativas constituyen los m elementos del centro de gravedad o promedio ponderado en N_m, que se obtiene matricialmente, en función de la matriz A, como se indica a continuación

$$g = r^t D_r^{-t} P = r^t A$$

el j-ésimo elemento del vector g está dado por g_j, el cual coincide con los totales columna de la matriz P.

Los totales hilera $P_{i.}$ definen un conjunto de ponderaciones tales que $\sum_{i=1}^{n} P_{i.} = 1$. A estas ponderaciones se les da el nombre de *masas* hilera. A cada uno de los n puntos representados por a_i se les asocia una *masa* m_i, que indica la importancia relativa de cada muestra i dentro de la población.

De igual forma se define la nube de perfiles columna en $\mathbb{R}^n$, que se expresa por $N(j)$, cuya expresión es:

$$B = D_c^{-1} P^t$$

donde los elementos de la matriz diagonal D_c^{-1} ponderan a cada una de las hileras de la matriz transpuesta de P. Los perfiles columna se obtienen de las hileras de B. Los totales columna $P_{.j}$ funcionan como *masas* m_j, y se les asocia a cada uno de los perfiles b_j.

El centro de gravedad h de $N(j)$ se establece en la siguiente expresión:

$$h = c^t D_c^{-1} P^t = r^t B$$

donde el i-ésimo elemento de h está representado por h_i, que corresponde a los totales hilera de la matriz P.

3.3 Definición de la métrica ji-cuadrada.

Para calcular la distancia entre dos perfiles hilera, los valores extremos de los elementos de a_i ó a_i, afectan el cálculo de la distancia. Esta es una limitación muy importante de la distancia euclideana. Es decir, cuando el j-ésimo elemento de uno o más de los perfiles hilera a_i es grande, la distancia se verá influida por ese elemento, restando importancia a los demás elementos.

La distancia más adecuada para el análisis de correspondencias es la distancia ji-cuadrada, que permite ponderar las diferencias entre los elementos correspondientes en a_i y $a_{i'}$ por el recíproco del total de la j-ésima columna de P. Significa que reduce las diferencias entre los elementos a_i y $a_{i'}$, correspondientes a las columnas más numerosas, e incrementa las diferencias más pequeñas correspondientes a columnas menos numerosas, lográndose mantener con ello las distancias reales que existen entre ambos puntos, y que en la representación gráfica final permitan representar los puntos sin alterar su estructura.

Así que la distancia ji-cuadrada se expresa por

$$d^2(a_i, a_{i'}) = \sum_{j=1}^{m} \frac{1}{g_j} (a_{ij} - a_{i'j})^2$$

$$= (a_i - a_{i'})' D_c^{-1} (a_i - a_{i'})$$

Esta distancia se llama así por su semejanza con el estadístico χ^2 que se utiliza para contrastar la hipótesis nula de independencia entre hileras y columnas de una tabla de contingencia.

La razón principal para elegir la distancia ji-cuadrada, que permite verificar la propiedad de la equivalencia distribucional, se expresa en Lebart et al. (1984) de la siguiente manera:

1) Si dos variables tienen idénticos perfiles y se suman, entonces la distancia entre las entidades permanece inalterable.

2) Si se suman dos entidades con idénticos perfiles de distribución, entonces la distancia entre las variables permanece sin cambio.

Estas propiedades son importantes, porque garantizan la invariabilidad de los resultados, independientemente de como las variables fueron originalmente codificadas, ya que las distancias y la inercia no se modifican.

Es posible considerar dos puntos que están sobrepuestos en el espacio como un punto simple, correspondiente al total de las dos categorías sumadas.

De este modo, la representación de las entidades sólo cambiará muy escasamente si las variables con perfiles similares se suman. En general, no existe pérdida de información cuando ciertas categorías se suman. Contrariamente, no existe beneficio alguno al subdividir categorías homogéneas.

Esto se demuestra en el caso de sumar dos entidades i'' y i''' en una entidad i' cuya frecuencia relativa $P'_{i.}$ satisface la relación

$$P'_{i.} = P''_{i.} + P'''_{i.}$$

Expresando la distancia $d^2(j, j')$ entre dos variables j y j', únicamente dos términos, llamado T_1 y T_2, hacen uso de i'' y i'''

$$T_1+T_2 = \frac{1}{P_{i'.}} \left[\frac{P_{i''j}}{P_{.j}} - \frac{P_{i''j'}}{P_{.j'}} \right]^2 + \frac{1}{P_{i'''.}} \left[\frac{P_{i'''j}}{P_{.j}} - \frac{P_{i'''j'}}{P_{.j'}} \right]^2$$

Después de sumarlos se reemplazan por T_0, tal como sigue

$$T_0 = \frac{1}{P_{i'.}} \left[\frac{P_{i'j}}{P_{.j'}} - \frac{P_{i'j'}}{P_{.j'}} \right]^2$$

y se demuestra que

$$T_o = T_1 + T_2$$

T_o se escribe como

$$T_o = P_{i'.} \left[\frac{P_{i'j}}{P_{i'.} P_{.j}} - \frac{P_{i'j'}}{P_{i'.} P_{.j'}} \right]$$

T_1 y T_2 se escriben similarmente; las tres cantidades son iguales, de manera que los perfiles de i'', i''' y i' son idénticos.

Pichardo (1990) maneja la variabilidad o dispersión de un conjunto de vectores con respecto a su centro de gravedad como un promedio de distancias ji-cuadrada ponderada por las masas asignadas a dichos puntos

$$\sum_{i=1}^{n} P_{i.} (a_i - g)^t D_c^{-1} (a_i - g)$$

En Ruiz (1989) se demuestra que el estadístico χ^2 es proporcional al promedio anterior, con constante de proporcionalidad X..

$$\chi^2 = X.. \sum_{i=1}^{n} P_{i.} (a_i - g)^t D_c^{-1} (a_i - g)$$

3.4 Inercia total para la nube de puntos en $N_{(I)}$ y $N_{(J)}$.

La inercia total se define en Ruiz (1989) como la variación total explicada por las nubes de puntos $N_{(I)}$ y $N_{(J)}$ expresadas por $In_{(I)}$ y $In_{(J)}$ respectivamente.

La inercia total $In_{(I)}$ de $N_{(I)}$ se cuantifica por el promedio ponderado de las distancias ji-cuadrada de los perfiles hilera a_i con su centro de gravedad g.

$$In(i) = \sum_{i=1}^{n} P_{i.} (a_i - g)^t D_c^{-1} (a_i - g)$$

De manera análoga, In(j) de N(j) está dada por el promedio ponderado de las distancias ji-cuadrada de los perfiles columna b_j con su centro de gravedad h

$$In(j) = \sum_{j=1}^{m} P_{.j} (b_j - h)^t D_r^{-1} (b_j - h)$$

Al comparar χ^2 definida anteriormente, con la expresión de la inercia total se tiene que la inercia total de los perfiles hilera en N(i) se representa por

$$In(i) = \chi^2 / X..$$

En tanto que la inercia total de los perfiles columna en N(j) se expresa como

$$In(j) = \chi^2 / X..$$

Donde se observa que

$$In(i) = \chi^2 / X.. = In(j)$$

lo que significa igualdad en la variación total explicada tanto para la nube de puntos en N(i) como para N(j).

3.5 Cálculo de los ejes principales y coordenadas para las nubes de puntos en N(i) y N(j).

Hasta el momento se tienen nubes de puntos en N(i) y N(j), en las que se conoce su variación total explicada a través de la inercia total. Estas dos nubes de puntos se encuentran en espacios dimensionales (m-1) y (n-1), de los cuales se desea obtener un subespacio de dos dimensiones, donde puedan representarse de manera accesible para facilitar su interpretación.

Es posible establecer el conjunto de resultados del análisis de correspondencias a partir de las nubes de puntos construidas anteriormente, con la métrica ji-cuadrada asociada a cada nube. Sin embargo, cuando Volle (1985) determina los ejes principales de una nube de puntos en un espacio euclideano $\mathbb{R}^K$ hace uso de la métrica euclideana canónica, estos ejes principales son ortogonales al origen del espacio $\mathbb{R}^K$, construidos a partir de la matriz de inercia S. Los resultados que de este análisis se obtienen pueden ser usados por el análisis de correspondencias, siempre que la distancia ji-cuadrada y las coordenadas de las dos nubes de puntos sean modificadas.

Se debe tener presente que al modificar las coordenadas, la distancia $d^2(a_i, a_{i'})$ y $d^2(b_j, b_{j'})$ y la inercia total no se alteran.

Las nuevas coordenadas para los n puntos de $N(i)$, acompañados cada uno de una masa m_i y situados dentro del espacio $\mathbb{R}^m$, están dados por el vector a_i.

$$a_i = \frac{P_{ij}}{P_{i\cdot}} \frac{1}{\sqrt{P_{\cdot j}}}$$

El centro de gravedad de $N(i)$ es g, de manera que g_j es el j-ésimo elemento de ese vector.

$$g_j = \sum_{i=1}^{n} P_{i\cdot} \frac{P_{ij}}{P_{i\cdot}} \frac{1}{\sqrt{P_{\cdot j}}}$$

$$g_j = \sum_{i=1}^{n} \frac{P_{ij}}{\sqrt{P_{\cdot j}}}$$

$$g_j = \sqrt{P_{\cdot j}}$$

Si se considera un ejemplo en el que las entidades se caracterizan por tres variables, la nube de puntos N_0 en $\mathbb{R}^m$ se reduce en su dimensionalidad y su nuevo espacio es de m-1 dimensiones. De tal suerte que la nube de puntos a_i, que representa las entidades, queda contenida en un subespacio de dos dimensiones, Figura 3.5.

De esta nube de puntos se obtiene información útil para analizar las entidades, tomando en cuenta la distribución de las variables en cada una de las entidades. Para este ejemplo no es difícil lograrlo, ya que se encuentra en un espacio de dos dimensiones; pero cuando se trabaja con tablas de datos más grandes, este análisis de datos no es tan fácil.

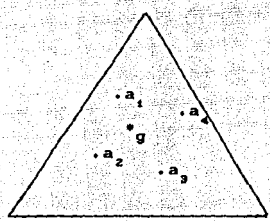


Figura 3.5. Proyección de los perfiles hilera a_i ,y su centro de gravedad g ,en el simplex.

La métrica que emplea análisis de correspondencias es la ji-cuadrada que se expresa a continuación

$$d^2(a_i, a_{i'}) = \sum_{j=1}^{m} \frac{1}{P_{\cdot j}} \left[\frac{P_{ij}}{P_{i\cdot}} - \frac{P_{i'j}}{P_{i'\cdot}} \right]^2$$

Al modificar la distancia ji-cuadrada, la nueva distancia es la siguiente:

$$d^2(a_i, a_{i'}) = \sum_{j=1}^{m} \left[\frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}} - \frac{P_{i'j}}{P_{i'.}\sqrt{P_{.j}}} \right]^2$$

Siendo ésta la distancia euclideana canónica, ya que el cuadrado de la distancia entre dos puntos es igual a la suma de los cuadrados de las diferencias de sus coordenadas. Los resultados que se obtienen de la distancia ji-cuadrada y euclideana son iguales.

La inercia total de N(i) es

$$In(i) = \sum_{i=1}^{n} P_{i.} \left| a_i - g \right|^2$$

$$\left| a_i - g \right|^2 = \sum_{j=1}^{m} \left[\frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}} - \sqrt{P_{.j}} \right]^2$$

$$In(i) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{P_{i.}\,(P_{ij} - P_{i.}\,P_{.j})^2}{P_{i.}^2\,P_{.j}}$$

$$In(i) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(P_{ij} - P_{i.}\,P_{.j})^2}{P_{i.}\,P_{.j}}$$

Para N(j) las nuevas coordenadas de los m puntos, acompañados de una masa m_j y dentro del espacio $\mathbb{R}^n$, se expresan de la siguiente manera:

$$b_j = \frac{P_{ij}}{P_{\cdot j}} \frac{1}{\sqrt{P_{i\cdot}}}$$

El centro de gravedad de $N(j)$ es el vector h de coordenadas

$$h_j = \sqrt{P_{i\cdot}}$$

$I_{N(i)}$ es simétrica, por consiguiente $I_{N(i)} = I_{N(j)}$. La inercia de las nubes $N(i)$ y $N(j)$ se interpretan como una medida aproximada de la información aportada por la matriz de datos.

Así que para obtener el subespacio óptimo, donde es posible analizar los datos de $N(i)$, se trazan ejes ortogonales que cruzan por el centro de gravedad. Sobre ellos se proyectan los puntos de los perfiles hilera, de modo que en cada uno de estos ejes estarán representados los puntos hilera, figura 3.6. A cada eje se le da el nombre de *eje principal*, y sobre él se proyectan los perfiles hilera, dando origen así a sus coordenadas. Figura 3.7.

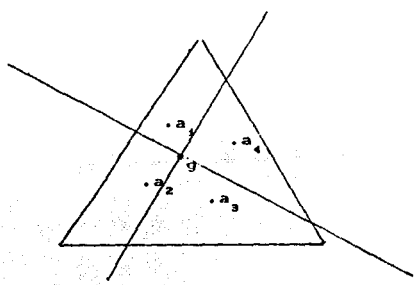


Figura 3.6. Trazo de los ejes ortogonales que cruzan por el centro de gravedad g en el simplex.

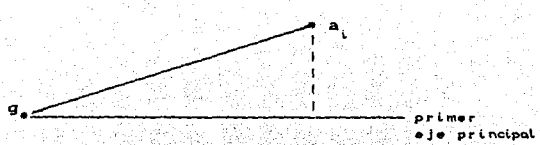


Figura 3.7 Primer eje ortogonal donde se proyecta el perfil hilera a_i

Para obtener los ejes principales de la nube de puntos N(i) es necesario partir de la matriz de inercia V que corresponde al centro de gravedad g. de término general

$$V_{jj'} = \sum_{i=1}^{n} P_{i.} \left(a_{ij} - g_j \right) \left(a_{ij'} - g_{j'} \right)$$

$$V_{jj'} = \sum_{i=1}^{n} P_{i.} \left[\frac{P_{ij}}{P_{i.} \sqrt{P_{.j}}} - \sqrt{g_j} \right] \left[\frac{P_{ij'}}{P_{i.} \sqrt{P_{.j'}}} - \sqrt{g_{j'}} \right]$$

$$V_{jj'} = \sum_{i=1}^{n} \left[\frac{P_{ij} - P_{i.} P_{.j}}{\sqrt{P_{i.} P_{.j}}} \right] \left[\frac{P_{ij'} - P_{i.} P_{.j'}}{\sqrt{P_{i.} P_{.j'}}} \right]$$

V es una matriz de inercia simétrica, porque $V_{jj'} = V_{j'j}$. La traza de esta matriz es la cantidad de inercia total I(N), que se calcula para conocer la variación total explicada por la nube de puntos en el espacio multidimensional.

Volle (1985) considera el análisis factorial de una nube de puntos en $\mathbb{R}^m$ con n puntos x_i, representados por las coordenadas x_{ij} $(i=1,2,\ldots,n)$ y $(j=1,2,\ldots,m)$ con una masa m_i, y la métrica

euclideana canónica.

La inercia total de N(u) con respecto al punto *P* se expresa de la siguiente manera

$$In(u) = \sum_{i=1}^{n} m_i \mid x_i - P \mid^2$$

Esta misma inercia de N(u) se puede expresar por las proyecciones ortogonales de x_i sobre el vector u, que es un vector unitario de coordenadas u_j, que pasa por *P*, figura 3.8. Así, la inercia total de N(u) se obtiene a partir de la siguiente expresión

$$In(u) = \sum_{i=1}^{n} m_i \mid z_i - P \mid^2$$

Al considerar que *P* es el origen del sistema de ejes, la inercia ahora se expresa como

$$In(u) = \sum_{i=1}^{n} m_i \mid z_i \mid^2$$

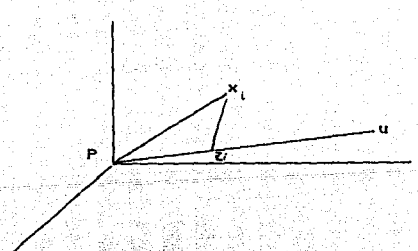


Figura 3.8. Inercia de los puntos z_i acompañados de la *masa* m_i, proyecciones ortogonales de x_i en el vector u que pasa por P.

con $$|Z_i|^2 = (x_i^t u)^2$$

como $$x_i^t u = \sum_{j=1}^{m} x_{ij}\, u_j$$

$$|Z_i|^2 = \sum_{j=1}^{m}\sum_{j'=1}^{m} x_{ij}\, x_{ij'}\, u_j\, u_{j'}$$

Se puede apreciar que $x_i x_i^t$ es la matriz (m,m) de término general $X_{ij}X_{ij'}$. De manera que $(x_i^t u)^2$ puede ser representada como

$$U^t X_i X_i^t U$$

$$In(U) = \sum_{i=1}^{n} m_i\, U^t X_i X_i^t U$$

$$In(U) = U^t \left[\sum_{i=1}^{n} m_i X_i X_i^t \right] U$$

De esta expresión se observa que se está calculando la nueva distancia euclideana canónica, de Z_i a el origen del sistema de ejes. Las coordenadas del vector unitario u tienen un papel importante ya que son éstos los que dan la posición de Z_i, y por lo tanto, expresan la inercia de $N(u)$.

El término entre corchetes de $In(U)$ es la matriz de inercia S, que se obtiene al premultiplicar la matriz B, de término general $b_{ij} = a_{ij}\sqrt{m_i}$ por su transpuesta.

$$In(U) = U^t B^t B\, U$$

$$In(U) = U^t S\, U$$

La expresión de $I_n(U)$ pertenece a una forma cuadrática donde S es una matriz cuadrática ($S = S^t$). La matriz simétrica S y la forma cuadrática ($U^tS\ U$) se llaman positiva semidefinida (psd) ya que ($U^tSU \geq 0$) para todo U (Zarate y Alvarez, 1985).

Para encontrar la dirección del vector u que presenta la máxima expresión de $I_n(U)$, llamesele *primer eje principal*, Lebart (1984) emplea los multiplicadores de Lagrange.

Para encontrar la dirección de U que maximize U^tSU se debe contar con la restricción de $U^tU=1$, ya que esto es lo que permite que los ejes sean ortogonales y que además den lugar a una base ortonormal para $\mathbb{R}^m$; así, se obtiene la siguiente relación

$$Su_\alpha = \lambda_\alpha u_\alpha$$

premultiplicando por u_α

$$u_\alpha^t Su_\alpha = \lambda_\alpha = I_n(u_\alpha)$$

El primer eje principal es el eigenvector u_1 que corresponde a λ_1. Los eigenvalores de S se clasifican en orden decreciente. El eigenvalor más grande se denota por λ_1 y el menor por λ_m; el eigenvalor común es señalado por el índice α ($\alpha = 1,2,\ldots,m$). La inercia explicada para el primer *eje principal* es λ_1.

Al considerar la matriz de inercia V de N(), que se construyó para el análisis de correspondencias, con respecto a su centro de gravedad g, de término general

$$V_{jj'} = \sum_{i=1}^{n} P_{i\cdot} (a_{ij} - g_j)(a_{ij'} - g_{j'})$$

y la matriz S, que se construyó para el análisis factorial de una nube de puntos, de término general

$$S_{jj'} = \sum_{i=1}^{n} P_{i\cdot} x_{ij} x_{ij'}$$

Volle (1985) establece los siguientes resultados

$$VG = 0$$

$$SG = G$$

$$\text{si } U'G = 0,\ VU = SU$$

Al diagonalizar VG=λG se observa que G es el eigenvector de V, donde el eigenvalor es cero. Este mismo eigenvector G pertenece a la matriz de inercia S, al diagonalizar SG=λG en cuyo caso el eigenvalor es uno. Las matrices V y S tienen los mismos eigenvectores ortogonales a G, y sus correspondientes eigenvalores son los mismos.

Al diagonalizar S se obtienen los siguientes resultados:

- El vector G, asociado al eigenvalor 1, que se le llama *eigenvector trivial* no aporta nada al análisis de N(i).
- u_1 es el primer eigenvector de V, asociada a λ_1.
- u_2 se encuentra asociado a λ_2.
- u_α es el α-ésimo eigenvector de V asociado a λ_α.

Para obtener los resultados de N(i) en el análisis de correspondencias, se diagonaliza la matriz S, donde las coordenadas se han modificado:

$$S_{jj'} = \sum_{i=1}^{n} P_{i.} \left[\frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}} \right] \left[\frac{P_{ij'}}{P_{i.}\sqrt{P_{.j'}}} \right]$$

$$S_{jj'} = \sum_{i=1}^{n} \frac{P_{ij}\ P_{ij'}}{P_{i.}\sqrt{P_{.j}\ P_{.j'}}}$$

Recordando que se puede escribir como

$$S = B^t B$$

donde B es la matriz de tamaño n*m, de término general

$$b_{ij} = a_{ij} \sqrt{m_i}$$

$$b_{ij} = \frac{P_{ij}}{P_{i.} \sqrt{P_{.j}}} \sqrt{P_{i.}}$$

$$b_{ij} = \frac{P_{ij}}{\sqrt{P_{i.} P_{.j}}}$$

Obtener los eigenvectores y eigenvalores para el análisis de $N(j)$ resulta de una permutación de las cordenadas de $N(i)$ por los de $N(j)$ en la matriz de inercia. Así se obtiene la matriz de inercia C, de término general

$$c_{ii'} = \sum_{j=1}^{m} \frac{P_{ij} P_{i'j}}{P_{.j} \sqrt{P_{i.} P_{i'.}}}$$

que se puede expresar por

$$C = BB^t$$

Al diagonalizar C, se obtiene el *eigenvector trivial* H asociado a el eigenvalor 1, y w_α es el α-ésimo eigenvector de C asociado a $\lambda\alpha$.

Al tomar en cuenta los resultados de Volle (1985), al realizar un análisis factorial de $N(i)$ y $N(j)$, es posible aplicarlos al análisis de correspondencias.

Al suponer que w_α es el eigenvector de C asociado a $\lambda_\alpha=0$,

$$Cw_\alpha = \lambda_\alpha w_\alpha$$

$$BB^t w_\alpha = \lambda_\alpha w_\alpha$$

Premultiplicando esta expresión por B^t se obtiene

$$(B^tB)B^tW = \lambda_\alpha w_\alpha$$

$$S(B^tw_\alpha) = \lambda_\alpha(B^tw_\alpha)$$

$B^tw_\alpha=0$. Si esto no ocurre entonces $BB^tw_\alpha=0$, lo que estaría en contradición con la hipótesis $\lambda_\alpha=0$. De manera que B^tw_α es un eigenvector de S asociado a el eigenvalor λ_α.

Sin embargo, B^tw_α no es unitario. Así que se debe encontrar un vector del tipo KB^tw_α que sea unitario:

$$(KB^tw_\alpha)^t\ KB^tw_\alpha = 1$$

$$K^2w_\alpha BB^tw_\alpha = K^2\lambda_\alpha w_\alpha^2 = K^2\lambda_\alpha = 1$$

de lo cual $$K = \frac{1}{\sqrt{\lambda_\alpha}}$$

$\frac{1}{\sqrt{\lambda_\alpha}}\ B^tw_\alpha$ es el eigenvector unitario de S asociado a λ_α.

Así que se obtienen las fórmulas de transición, importantes, ya que con ellas solo es necesario diagonalizar una matriz de inercia para obtener los eigenvectores de la otra matriz de inercia:

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} B^t w_\alpha$$

$$w_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} B u_\alpha$$

Estas expresiones son posibles por razones de simetría.

A partir de las dos fórmulas de transición, que juegan un papel importante en la interpretación del análisis de correspondencias, es posible establecer una segunda relación.

Para la i-ésima coordenada de w_α:

$$w_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{k} b_{ij} u_{\alpha j}$$

como $b_{ij} = \dfrac{P_{ij}}{\sqrt{P_{i.} P_{.j}}}$

$$w_{\alpha i} = \sqrt{\frac{P_{i.}}{\lambda_\alpha}} \sum_{j=1}^{k} \frac{P_{ij}}{P_{.j}} u_{\alpha j}$$

al considerar $F_\alpha(i)$ la coordenada de a_i sobre el eje u_α se obtiene la siguiente expresión:

$$w_{\alpha i} = F_{\alpha}(i) \sqrt{\frac{P_{i.}}{\lambda_{\alpha}}}$$

$$u_{\alpha j} = G_{\alpha}(j) \sqrt{\frac{P_{.j}}{\lambda_{\alpha}}}$$

La segunda relación se establece al intercambiar los papeles de las letras i y j, siendo $G_{\alpha}(j)$ la coordenada de b_j sobre el eje w_{α}.

Si

$$G\alpha(j) = \sum_{i=1}^{k} w_{\alpha i} \frac{P_{ij}}{P_{.j}\sqrt{P_{i.}}}$$

$$G\alpha(j) = F\alpha(i) \sqrt{P_{i.}/\lambda_{\alpha}} \frac{P_{ij}}{P_{.j}\sqrt{P_{i.}}}$$

Entonces

$$G\alpha(j) = \frac{1}{\lambda_{\alpha}} \sum_{i=1} \frac{P_{ij}}{P_{.j}} F\alpha(i)$$

$$F\alpha(i) = \frac{1}{\lambda_{\alpha}} \sum_{j=1} \frac{P_{ij}}{P_{i.}} G\alpha(j)$$

Estas expresiones proporcionan las nuevas coordenadas para las nubes de puntos hilera y columna, en los nuevos espacios multidimensionales donde los ejes principales estan dados por los eigenvectores que se obtienen de la matriz de inercia S y C respectivamente. Así, el siguiente paso es elegir los ejes principales que expliquen el mayor porcentaje de variación, y con ellos construir la gráfica de dos dimensiones, donde finalmente se representan las dos nubes de puntos.

3.6 Elementos suplementarios.

Cuando se desea que un perfil hilera o columna se represente en el análisis, sin que tenga un papel activo en la determinación de los ejes principales, se puede considerar como elemento suplementario, estos puntos son útiles para enriquecer la representación gráfica.

La matriz de datos P puede, entonces, incrementarse por m_s columnas suplementarias, el caso de las n_s hileras se deduce por simple permutación de índices, figura 3.9.

Así, se procede a ubicar los perfiles de los m_s nuevos puntos con respecto a los m puntos analizados en $\mathbb{R}^n$. Sea P^+_{ij} la i-ésima coordenada de la j-ésima columna suplementaria. El perfil de este elemento es el vector cuyo j-ésimo componente es

$$b^+_j = \frac{P^+_{ij}}{P^+_{\cdot j}\sqrt{P_{i\cdot}}}$$

donde

$$P^+_{\cdot j} = \sum_{i=1}^{n} P^+_{ij}$$

La proyección del j-esimo punto en el α-ésimo eje principal, esta dado por

$$G_{\alpha}(j) = \frac{1}{\sqrt{\lambda\alpha}} \sum_{i=1}^{n} \frac{P^{+}_{ij}}{P^{+}_{.j}} F_{\alpha}(i)$$

donde $F_{\alpha}(i)$ es la i-ésima coordenada de los perfiles hilera, con respecto a el α-ésimo eje principal.

Para una hilera suplementaria a_i, se tiene, en forma análoga

$$G_{\alpha}(j) = \frac{1}{\sqrt{\lambda\alpha}} \sum_{i=1}^{n} \frac{P^{+}_{ij}}{P^{+}_{.j}} F_{\alpha}(i)$$

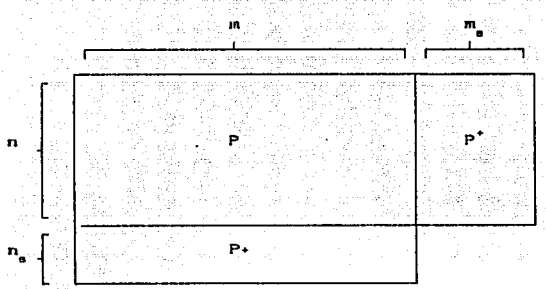


Figura 9. Representación de hileras y columnas suplementarias.

Estas dos expresiones proporcionan las coordenadas de los elementos suplementarios en el espacio multidimensional, sin que contribuyan a la inercia total y por lo tanto a la construcción de los ejes principales. Estos elementos ayudan a obtener una mejor descripción de los datos.

CAPITULO IV.

INTERPRETACION DEL ANALISIS DE CORRESPONDENCIAS.

Después del manejo algebraico y numérico, del análisis de correspondencias, falta interpretar correctamente los resultados que se obtienen, etapa muy importante, que además puede requerir de nuevos cálculos. Cuando se busca llevar la información original, que se encuentra como un conjunto de puntos en el espacio multidimensional, a un gráfico de tan sólo dos dimensiones ocurre una pérdida de inercia, o variabilidad explicada, para ser más preciso. Sin embargo, la pérdida relativa de ciertos aspectos de la información se ve compensada con su mayor simplicidad. Así, es importante interpretar tanto los eigenvectores como los eigenvalores; decidir que eigenvectores serán consideradados para la construcción del gráfico, tomando en cuenta el valor de su respectivo eigenvalor.

Cuando las nubes de perfiles hilera y columna se representan simultáneamente en una gráfica, con respecto a dos ejes principales, es necesario averiguar que perfil determina la orientación de los ejes e interpretar la dispersión de los distintos perfiles. De este modo, se consideran las siguientes medidas que serán útiles en la correcta interpretación de la gráfica que se construye:

1. Contribuciones absolutas.
2. Contribuciones relativas.
3. Calidad de la representación.

4.1 Contribuciones absolutas.

Las contribuciones absolutas indican la proporción de variabilidad explicada para cada variable, en relación al α-ésimo eje principal; esta proporción se calcula con respecto al grupo completo de variables.

La variabilidad explicada para el α-ésimo eje principal se expresa por $\lambda\alpha$, de modo que

$$\lambda\alpha = \sum_{i=1}^{n} m_i F^2_{\alpha(i)}$$

donde $F^2_{\alpha(i)}$ es la coordenada de la i-ésima variable a_i, y m_i es su respectiva masa; $\lambda\alpha$ es el eigenvalor asociado al α-ésimo eigenvector, llamado eje principal; ambos valores se obtienen al diagonalizar la matriz de inercia.

De lo anterior se desprende que es posible cuantificar la aportación de la variable a_i en la construcción del α-ésimo eje principal. Así, se define la contribución absoluta, $CA_{\alpha(i)}$, para la i-ésima variable

$$CA_{\alpha(i)} = \frac{m_i F^2_{\alpha(i)}}{\lambda\alpha}$$

Para la j-ésima variable, b_j, la contribución absoluta está dada por

$$CA_{\alpha(j)} = \frac{m_j G^2_{\alpha(j)}}{\lambda\alpha}$$

Cuando la variable a_i o b_j presenta una masa grande, tiene una contribución grande en la determinación y dirección del α-ésimo eje principal; sin embargo, también puede verse que la variable a_i o b_j contribuirá grandemente a la variación explicada para el α-ésimo eje principal si se ubica muy alejado de su centro de gravedad.

4.2 Correlaciones al cuadrado.

La correlación cuadrada es una más de las ayudas que proporciona el análisis de correspondencias para interpretar la representación gráfica de un conjunto de datos, en un plano de dos dimensiones, ya que permite conocer que tan cerca se encuentra una variable en el α-ésimo eje principal.

Si se considera el primer eje principal, figura 4.1, es posible cuantificar la cercanía de la variable a_i a este eje, a través de la siguiente expresión

$$\cos \theta = \frac{f_{i1}}{d_i}$$

donde f_{i1} es la coordenada de a_i en el primer eje principal, y d_i es la distancia de a_i a su centro de gravedad g

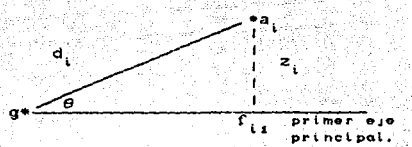


Figura 4.1. Coordenada f_{i1} del i-ésimo perfil hilera a_i al primer eje principal.

Dado que $m_i d_i^2$ es la variabilidad explicada de a_i con respecto a la inercia total I_{nn}, y que parte de ésta variabilidad explicada para el primer eje principal es $m_i f_{i1}^2$, se obtiene una nueva expresión

$$\cos \theta = \frac{f_{i1}}{d_i} = \frac{m_i f_{i1}^2}{m_i d_i^2}$$

Esta expresión da origen a la definición de la correlación cuadrada $CC_\alpha(i)$, y al extenderlo a los demás ejes principales se obtiene la siguiente expresión:

$$CC_\alpha(i) = \cos^2\theta_\alpha = \frac{f^2_{i\alpha}}{d^2(a_i,g)}$$

donde $d^2(a_i,g)$ es la distancia cuadrada del perfil hilera a_i y su centro de gravedad g, θ es el ángulo entre a_i y el α-ésimo eje principal.

En cuanto a la correlación al cuadrado de los perfiles columna b_j se define como

$$CC_\psi(j) = \cos^2\psi_\alpha = \frac{g^2_{j\alpha}}{d^2(b_i,r)}$$

donde ψ es el ángulo entre el α-ésimo eje principal y b_j, $g_{j\alpha}$ es la coordenada del perfil columna con respecto al eje principal, y r es su centro de gravedad.

Si $CC_\alpha(i)$ esta cercano a 1, significa que el perfil hilera a_i se encuentra muy cerca al α-ésimo eje principal; es decir, está bién representado y por lo tanto este perfil se explica ampliamente en este eje principal; esto mismo se presenta para la correlación al cuadrado de los perfiles columna. Por el contrario, si $CC_\alpha(i)$ esta cercano a cero, a_i se encuentra muy alejado del α-ésimo eje principal, es decir, es casi ortogonal a este eje; su participación en la explicación de este eje es pobre, por lo que es más probable que este mejor representado en otro eje principal.

Se observa que la suma de las correlaciones al cuadrado de los perfiles hilera a_i proporcionados para cada eje principal, que compone el espacio multidimensional, es igual a uno.

$$\sum_{\alpha=1}^{k} CC_{\alpha}(i) = 1$$

4.3 Calidad de la representación.

La calidad de la representación es una más de las ayudas que proporciona el análisis de correspondencias para interpretar la gráfica de dos dimensiones; para el caso del perfil hilera a_i, la calidad denotado por $CALD_2(i)$, está dado por

$$CALD_2(i) = \sum_{\alpha=1}^{2} CC_{\alpha}(i)$$

Cuando un punto está bién representado en el plano de dos dimensiones, $CALD_2(i)$ esta próximo a 0.90; cuando el valor de la calidad de la representación esta cercana a 0.20 entonces la representación gráfica de este punto no está bién representado en estos dos ejes.

De manera análoga se define la calidad de la representación para los perfiles columnas b_j en el subespacio formado por los dos primeros ejes principales y está dada por

$$CALD_p(j) = \sum_{\alpha=1}^{p} CC_{\alpha}(j)$$

4.4 Edición de datos.

Analizar los datos que se obtienen de una investigación biológica requiere de la interacción entre las condiciones en las que se efectúan las observaciones y los resultados que se obtienen al emplear el análisis de correspondencias. Este es un aspecto muy importante, y que muchas veces se descuida. Se tiene la idea errónea de que cuando se utiliza esta técnica estadística, será capaz de describir los datos, sin la necesidad de averiguar, por parte del investigador, como fueron codificados los datos y si se pueden manejar indistintamente. Si se desea describir ampliamente la información recabada, es necesario seleccionar las variables que serán utilizadas, se debe tener una idea definida y clara del objetivo de la investigación; es fundamental, entonces, contar con una gran experiencia y juicio para la selección de variables que serán involucradas en el proceso descriptivo del caso de estudio. En otras palabras, es necesario retener únicamente aquellos hechos que están relacionados a un punto de vista en particular. Esta condición hace que la interpretación sea más fácil y clara. En la práctica, este requisito conduce a identificar grupos de variables; algunos de ellos tendrán un papel activo en la construcción de tipologías, mientras que otros un papel de variables ilustrativas.

Las variables utilizadas se definen con base a características físicas, químicas, ecológicas o biológicas presentes en los objetos de interés para el estudio que se esté llevando a cabo. Se debe tener especial cuidado en la forma como se reportan los datos. El análisis de correspondencias requiere que estos datos sean expresados como números enteros positivos; esto es, deben ser presentados como frecuencias estadísticas.

La diferencia que existe entre una variable activa y una ilustrativa es que esta última no participa en el análisis; esto es, en la construcción del eje principal, la interpretación de su

correlación cuadrada es lo importante. De acuerdo con lo anterior, las variables ilustrativas pueden convertirse en variables activas, o por el contrario. Este proceso algunas veces da una interpretación más completa.

Para obtener tanto la gráfica de dos dimensiones, como las ayudas para la interpretación de la tabla de datos, análisis de correspondencias requiere de un programa de computadora que sea eficiente en la presentación de tales resultados. En el siguiente capítulo se presenta el programa CORAN y se mecionan sus características. Entre ellas está la de poder analizar las variables de forma activa, suplementaria e incluso eliminarlas del análisis con solo modificar el identificador de columnas.

CAPITULO V.

PROGRAMA CORAN PARA EL ANALISIS DE CORRESPONDENCIAS.

Es evidente que el analizar un gran volumen de datos, sin contar con el auxilio de lenguajes de programación y equipo de cómputo adecuado ocasionaría un caos en el momento de realizar extensos manejos algebraicos y numéricos. El análisis de correspondecias podría ser poco utilizado, a pesar de ser una técnica que permite analizar tablas de contingencia de tamaño n*m, cuyos elementos son números enteros positivos, ya que requiere de grandes cálculos numéricos, que realizarlos con calculadora de mano resultaría poco práctico, si se considera que lo importante es contar con una rápida interpretación de la información de diversas investigaciones. Afortunadamente existe tecnología en computadoras que permite desarrollar programas que se encargan de realizar las operaciones numéricas necesarias.

Para el análisis de correspondencias se cuenta con el programa "CORAN", que se basa en el programa publicado por Lebart et al. (1984), que permite obtener representaciones geométricas de las hileras y columnas de una tabla de contingencia (n*m), como puntos en un mismo subespacio de dos dimensiones.

5.1 Principales características del programa CORAN.

Está escrito en FORTRAN IV, puede ejecutarse en computadoras de tamaño medio o grande.

Consta de dos subrutinas principales: SELEC y CORAN. La primera de ellas se encarga de la lectura de datos y selección de los parámetros necesarios para la ejecución de la segunda subrutina, que se encarga de realizar todos los cálculos numéricos.

Puede manejar grandes matrices de datos, sin limitación práctica en el número de hileras, la matriz de datos nunca se almacena en memoria central. Cuenta con una subrutina de diagonalización, bajo la cual el número de columnas a manejar puede ser bastante grande.

El programa esta autocontenido, y cuenta con los procedimientos numéricos y gráficos necesarios.

Puede usarse para procesar tablas binarias así como tablas de contingencia. Su ejecución en el caso de grandes tablas binarias no puede competir con el programa de análisis de correspondecias múltiple, trabajando directamente sobre una matriz codificada reducida.

5.2 Parámetros utilizados por CORAN.

Para utilizar CORAN es necesario crear un archivo de datos, usando cualquier editor de texto, y definir en él los parámetros que permiten elegir, entre otras cosas, el número de gráficas que se desean analizar, el número de coordenadas y ejes principales que pueden ser calculadas, así como el hecho de elegir las variables y muestras que pueden participar como elementos activos o suplementarios. El cuadro 5.1 presenta una tabla, donde se muestra la posición de los diferentes parámetros para su fácil ubicación.

I. Titulo.
Se escribe el nombre del estudio de caso en un máximo de 80 carácteres.

II. Cinco parámetros escritos en un formato 5I4, es decir son cinco números enteros de un máximo de cuatro cifras:
1. IEXA Indica el número de hilera que contiene la matriz de datos.

2. NQEXA Número de columnas con que cuenta la matriz de datos, el identificador de hilera no se cuenta como una columna, esto es el nombre que se le da a cada una de las hileras.

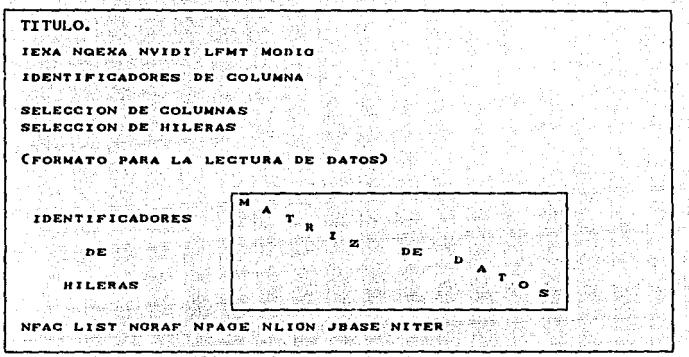


Figura. 5.1. Archivo de datos donde se especifican los parámetros necesarios para usar CORAN.

3. NVIDI Longitud del identificador de hilera, es decir, el nombre que presenta cada una de las muestras. Puede ser un multiplo de 4 carácteres. Máximo NVIDI=15, correspondiendo a 60 carácteres. Los primeros 4, 8 o 12 carácteres aparecen en la gráfica. El identificador es necesario y debe de estar en el inicio de cada hilera; por ejemplo, si el nombre de la muestra cuenta con 12 caracteres entonces el valor NVIDI=3.

4. LFMT Número de registros de formato, es decir, es la linea en la que se especifica el formato para la lectura de la matriz de datos y del identificador de hileras. Cada registro (linea) acepta hasta 20 caracteres. Si un registro no es suficiente para escribir el formato de lectura se puede continuar en el siguiente registro, para ello es necesario indicarlo con este parámetro. Si solo se requiere un registro de formato, LFMT=1.

5. MODIG Es el modo que define la selección de las hileras. Existe dos formatos para seleccionar y definir las hilera para el análisis. Si MODIG=0, todas las hilera de la matriz de datos son activas, es decir participan todas las muestras en el análisis. Si MODIG=1, existen dos formas de selección. Algunas hileras pueden ser suplementarias o ignoradas.

III. Identificadores de columnas.

Para identificar una columna se requiere de cuatro carácteres que se leen, antes de seleccionar las columnas, en un formato fijo (20A4) para las NQEXA columnas originales; es decir, el nombre de cada variable consta de cuatro carácteres, y se pueden escribir hasta 20 nombres por cada registro (una hilera).

IV. Selección de columnas.

Con este parámetro es posible seleccionar el número de columnas a trabajar. Su lectura es en formato 80I1 para las NQEXA columnas de la matriz de datos; significa que es posible asignar por cada registro (una hilera) 80 indicadores que señalen la selección de la columna, de acuerdo al siguiente código:

0 = columnas eliminadas para el análisis
1 = columnas que participan en el análisis
2 = columnas ilustrativas

V. Selección de hileras.

Cuando MODIG=0, no hay selección de hileras para el análisis de la matriz de datos, todas las muestras participan. Cuando MODIG=1 es posible seleccionar las hileras, y establecer cual de ellas toman parte en el análisis en forma activa o suplementaria, o simplemente son eliminados de éste, por lo que se leen los indicadores con un formato 80I1, para cada una de las IEXA hileras. Los indicadores se registran con el siguiente código:

0 = hileras con esta selección son eliminadas del análisis.
1 = hileras que participan en el análisis.
2 = hileras ilustrativas.

VI. Formato para la lectura de datos.
El formato se escribe en paréntesis en LFMT registros. El formato inicia con NVIDI*A4, es decir la longitud del identificador de la hilera (nombre de la muestra). El resto se lee en un formato real (F), aún para valores enteros, se debe expresar por: (NQEXA F longitud máxima del dato mayor de la matriz de datos).

VII. Los datos.
Los datos se leen de acuerdo con el formato anterior. Existen IEXA hileras, cada una de las hileras tiene una longitud de NVIDI+NQEXA.

VIII. Siete parámetros en un formato 7I4, son siete números enteros de una longitud máxima de 4 cifras.
1. NFAC Número de coordenadas principales que se calculan y que se presentan en la tabla de resultados.

2. LIST3 Este parámetro es útil para elegir la impresión de los resultados de la información de las hileras. Su código está dado de la siguiente manera:
   0 = no se imprime
   1 = se imprimen las coordenadas y contribuciones de las hileras.
La información de las columnas siempre se imprime.

3. NGRAF Se indica el número de gráficas que deben imprimirse (NGRAF≤10). En esta versión, los planos principales son sucesivamente: (1,2), (2,3), (3,4), etc., generalmente se imprimen las dos primeras gráficas por ser las que contienen la máxima variación explicada. Todos los puntos hilera, columnas, activas o suplementarias se muestran en las gráficas.

4. NPAGE Número de páginas para el ancho de cada gráfica (NPAGE≤8). Se recomienda que NPAGE=1.

5. NLIGN Número de líneas por gráfica. Se recomienda que NLIGN=60*NPAGE-2. Si NLING=0, ambos ejes son escalados idénticamente. La selección más común NLIGN=59, ya que permite que la gráfica sea representada en una página, considerando NPAGE=1.

6. JBASE Dimensión del espacio para la aproximación en el caso de que sea lectura directa. Si JBASE=0 la diagonalización toma lugar en memoria principal (elección usual). Si JBASE=1 el valor por omisión es JBASE=NFAC+3. En otro caso JBASE=dimensión del espacio por aproximación.

7. NITER (Unicamente si JBASE≠0), es el número de iteracciones en lectura directa. Si NITER=0 el valor por omisión es NITER=8.

A continuación se presenta un ejemplo concreto, en el que se puede apreciar como se distribuyen los distintos parámetros que se definieron anteriormente.

El ejemplo corresponde al estudio de las posibles relaciones que existen entre varias ventajas con un número espcecífico de trabajos, tomado de Lebart (1984). Esta información se encuentra en el archivo de datos del cuadro 5.2, y que a continuación se describe.

El archivo de datos indica que son 20 trabajos (IEXA=20) y 15 ventajas (NQEXA=15). NVIDI=6 indica que la longuitud máxima es de 24 caracteres, en la que se puede escribir el nombre de las muestras. El formato de lectura para los datos ocupa tan sólo un registro, LFMT=1. El parámetro MODIG=1 indica que las hileras pueden ser activas, suplementarias o sencillamente ignoradas. En la selección de columnas se observa que la variable 14 es eliminada del análisis, la variable 2 y 5 son suplementarias, las demás participan en el análisis. En la selección de hileras se observa que la hilera 16 participa como elemento suplementario y los demás empleos toman parte en el análisis.

El número de coordenadas principales que se calculan es de 6 (NFAC=6). Las coordenadas y contribuciones de los empleos (hileras) también se imprimen (LIST3=1). Finalmente, sólo se imprime una gráfica (NGRAF=1), el ancho de ésta es de una página (NPAGE=1) y presenta 58 líneas (NLIGN=58).

5.3 Instructivo para el manejo del programa CORAN.

Después de que se ha creado el archivo de datos es necesario poner en marcha el programa, para ello se escribe CORAN y se pulsa (RETURN). A continuación aparece en pantalla la palabra CORAN y se pregunta: DAME EL NOMBRE DEL ARCHIVO DE DATOS?. Se debe escribir el nombre del archivo que contiene la información, el archivo de los datos debe estar en el mismo disco que el programa, o en el mismo directorio.

Cuando inicia el análisis aparece en pantalla un texto en el que se lee:

C O R A N
ANALISIS DE
CORRESPONDENCIAS
EL NOMBRE DEL ARCHIVO DE SALIDA CORRESPO.SAL

.Al finalizar el análisis aparece el texto:

ANALISIS DE
CORRESPONDENCIAS
T E R M I N A D O

La información proveniente del análisis de correspondencias se almacena en el archivo CORRESPO.SAL, en el mismo disco o directorio en que se encuentra el programa CORAN. Para tener acceso a la información es necesario imprimir los resultados.

5.4 Resultados obtenidos del programa CORAN.

La lectura de datos y selección de los parámetros del archivo de datos (cuadro 5.2) se lleva a cabo a través de la subrutina SELEC, esta información da origen a los primeros resultados que se obtienen del análisis de correspondecias, cuadro 5.3.

Los siguientes resultados que se obtienen, pertenecen a la segunda subrutina, CORAN. El cuadro 5.4 muestra las tarjetas de parámetros para realizar el análisis de correspondencias. Posteriormente, se presenta un histograma para los primeros eigenvalores, en el que se puede identificar, de forma clara y precisa, el porcentaje de la variación explicada para cada eje principal, cuadro 5.5. Los cuadros 5.6 y 5.7 muestra las coordenadas, contribuciones absolutas y correlaciones al cuadrado de los primeros seis ejes principales tanto para las ventajes y empleos respectivamente. La primera columna muestra sus masas; la segunda, la distancia ji-cuadrada entre los puntos y el centro de gravedad. Finalmente, en la gráfica 5.1 se muestra la representación de las ventajas y de las muestras como puntos.

```
ANALYSIS OCCUPATIONS / JOB ADVANTAGES
   20  15   6   1   1
VARIFREEHUMASCHESALASECUCOMPINTENEARATMOSOCIAUTOLIKEOTHENONE
12112111111101
11111111111111121111
(1X,6A4,F1.0,14F4.0)
 FARM1FARMING/FISHING      4 159   0   3   2   2   9   3  12   2   1   4  11  15  12
 FAR2?FARM/FOOD INDUSTRY 1  13   3  10  17  12   4   1   8   3   5   1   9   5  11
 ENER1ENERGY/MINES         1   9   1   0   4  13   0   2   2   0   2   1   4   3   6
 STEE1STEEL                5   5   2   9  18   5   3   2   6   5   5   0   2   3  22
 CHEM1CHEMICAL/GLASS/OIL 2   7   1   4  15   5   2   1   6   1   2   2   3   0   5
 WOOD1WOOD/PAPER           2   5   0   4   1   0   3   0   2   1   1   1   1   0   3
 AUT 1AUTO/AVIATION/SHIP 2   3   1   8  16  17   1   8   7   2   4   3   6   1  24
 TEXT1TEXTILE/LEATHER      3  18   0   6  16   5   4   4  13   4   2   3   6   2  26
 PHAR1PHARMACY INDUSTRY    3   7   3   6   6   0   0   2   6   3   3   0   2   1   8
 MANU1MANUFACTURING        0  18   1  12  31   7   0   8  19  11   3   2  10   4  26
 CONS1CONSTRUCTION         7  63   2   9  31   9   4   6   9  10   3   4  14   8  35
 FOOD1FOOD/GROCERY         2  43  16   7   6   4   7   1   8   2   0   1   6   1   7
 SBUS1SMALL BUSINESS       8  95  23  15  15   2  13   7   9   5   2   3  13   4  18
 MBUS1MISCEL.BUSINESS      5  32   9   9  17   4   5   4   7   4   3   0   8   3  18
 ADMI1ADMINISTRATIVE SER.9  26  10  24  24  80  10  17  11   3   8   2   6   9  16
 TELE1TELECOMMUNICATION    1   7   2  11   3  14   2   6   3   1   1   2   1   3   5
 SO.S1SOCIAL SERVICES      4  10  10   8   2   1   6   4   2   3   1   0   3   2   1
 HE.S1HEALTH SERVICES      3  31  16  15  11  19   5  19  10   2   3   7  24   1   5
 TEAC1TEACHING/RESEARCH    2  33  27  21   9  18  27  24   3   4  43   8  18   3  11
 TRAN1TRANSPORTATION       2  19   2  12  12  21   0   1   4   5   5   1   3   3  13
   6   1   1   1  58   0   0
```

Cuadro 5.2 Archivo de datos en el que se encuentran los parámetros necesarios para iniciar el programa CORAN.

```
1                                  PASO  1: SELEC  1:
0-------------------------------------------------------------------------------
      TITULO=ANALYSIS OCCUPATIONS / JOB ADVANTAGES
      IEXA= 20    NOEXA= 15    NVIDI=  6    LFMT =  1    MODIG=  1
0NOMBRE DE LAS COLUMNAS
  VARI  FREE  HUMA  SCHE  SALA  SECU  COMP  INTE  NEAR  ATMO  SOCI  AUTO  LIKE  OTHE  NONE
0RESUMEN DE SELECCION
0     TIPO  1     NUMERO DE VARIABLES    12
0     TIPO  2     NUMERO DE VARIABLES     2
0VECTOR INDICADOR DE  ,    15  ELEMENTOS EN GRUPOS DE 10/
 12112111111  11101
0RESUMEN DE SELECCION
0     TIPO  1     NUMERO DE VARIABLES    19
0     TIPO  2     NUMERO DE VARIABLES     1
0VECTOR INDICADOR DE  ,    20  ELEMENTOS EN GRUPOS DE 10/
 11111111111  11111121111
0FORMATO
(1X,6A4,F1.0,14F4.0)
0
0-------------------------------------------------------------------------------
1                                  FIN DE LECTURA Y SELECCION
```

Cuadro 5.3 Fin de lectura y selección de los parámetros del archivo de datos.

```
:                                        PASD   ** CORAN **
0---------------------------------------------------------------------------------------------------------------------------
0TARJETAS DE PARAMETROS PARA CORAN
 --------------------------------
0         NFAC =   6    LIST3=   1    NGRAF=   1    NPAGE=   1    NLIGN=  58    JBASE=   0    NITER=   0
  0 ARCH. LECT.=   13  (NLEG)    ANALYSIS OCCUPATIONS / JOB ADVANTAGES
0ARCHIVO DE SALIDA  =  14  (NSAV)
0ARCHIVO DE SALIDA  =  11  (NGUS)
0PRE-LECTURA DE PARAMETROS SOBRE LOS ARCHIVOS          ICARD=  19    ITOT =  20    NACT =  12    NVAR =  14    NVIDI=   6
0MEMORIA USADA                    UD. TIENE RESERVADO 10000           UD. NECESITA        464
```

Cuadro 5.4 Tarjeta de parámetros para el an lisis de corfespondencias.

```
EIGENVALORES
--------------------------------------------------------------------------------------
SUMA DE EIGENVALORES              .46907420
HISTOGRAMA DE LOS PRIMEROS EIGENVALORES
       EIGENVALOR  PORCENTAJE   PORCENTAJE
                                  ACUM.
  1     .15251670     32.51        32.51     **************************************
  2     .14810440     31.57        64.09     *****************************
  3     .06014945     12.82        76.91     ***********************
  4     .04179156      8.91        85.82     ****************
  5     .02319160      4.92        90.75     ***********
  6     .01550103      3.30        94.05     ********
  7     .01107482      2.36        96.41     ******
  8     .00779358      1.58        97.99     ****
  9     .00424813       .91        98.89     **
 10     .00225270       .69        99.59     **
 11     .00194005       .41       100.00     *
```

Cuadro 5.5 Histograma de los primeros 11 eigenvalores que indican el porciento de variación explicada para los ejes principales.

1 COORDENADAS Y CONTRIBUCIONES DE LAS COLUMNAS

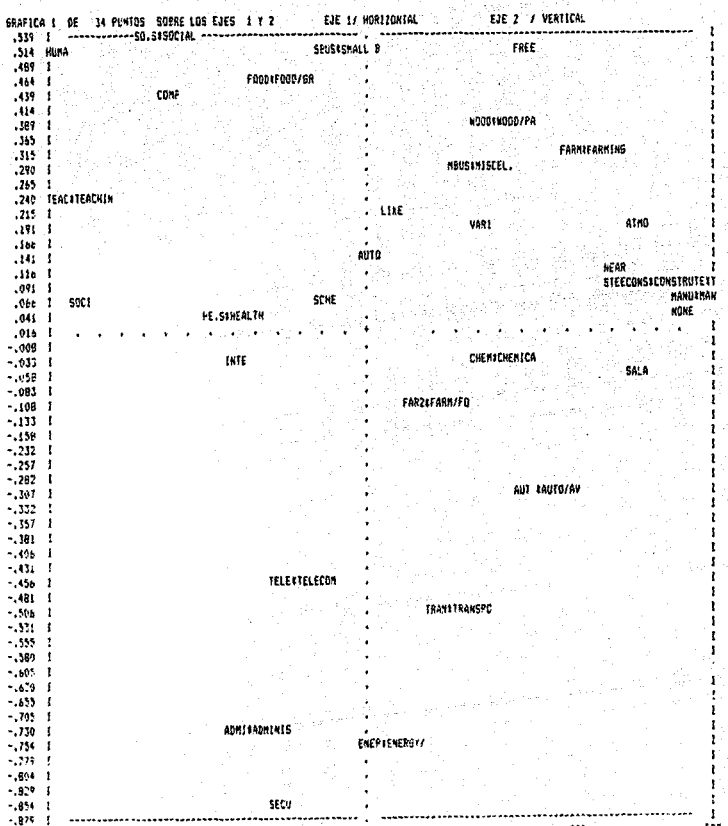
ONOMBRES	MASAS	DIST.²	COORDENADAS					CONTRIBUCIONES ABSOLUTAS					CORRELACIONES AL CUADRADO				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
VARI	.040	.42	.17	.17	-.25	-.29	.28	.9	.8	4.1	8.2	13.4	.08	.07	.14	.20	.18
HUNA	.080	.83	-.59	.49	-.39	-.19	-.17	18.1	13.0	20.5	7.0	9.6	.42	.29	.18	.04	.03
SOME	.121	.09	-.12	.04	.04	-.16	-.07	1.1	.1	.3	7.1	2.9	.16	.02	.02	.28	.06
SECU	.141	.83	-.19	-.88	-.13	-.03	.04	3.5	73.3	3.8	.4	.9	.05	.93	.02	.00	.00
CORP	.065	.52	-.40	.42	.01	-.04	.40	6.6	7.8	.0	.3	45.2	.30	.34	.00	.00	.31
INTE	.072	.28	-.28	-.04	.04	.28	-.19	3.8	.1	.2	13.3	11.6	.29	.00	.01	.28	.13
NEAR	.090	.35	.45	.11	-.12	.13	.04	12.1	.7	2.1	3.9	.7	.59	.03	.04	.05	.00
ATMO	.044	.46	.48	.18	.08	-.21	-.24	6.7	1.0	.4	4.7	10.9	.51	.07	.01	.10	.12
SOCI	.060	.94	-.54	.06	.79	-.06	-.01	11.6	.2	62.3	.5	.0	.31	.00	.66	.00	.00
AUTO	.027	.46	-.03	.13	.09	.56	.19	.0	.3	.4	20.4	4.2	.00	.04	.02	.69	.08
LIKE	.094	.23	.03	.21	-.12	.37	-.04	.1	2.7	2.2	30.7	.6	.00	.16	.06	.59	.01
NONE	.168	.37	.57	.02	.11	-.09	.00	35.4	.0	3.7	3.5	.0	.86	.00	.04	.02	.00
ELEMENTOS SUPLEMENTARIOS																	
FREE	.393	2.10	.29	.49	-.35	.45	.83	.0	.0	.0	.0	.0	.04	.12	.06	.09	.33
SALA	.159	.37	.46	-.08	.04	-.09	-.15	.0	.0	.0	.0	.0	.57	.02	.01	.02	.06

Cuadro 5.6 Coordenadas y contribuciones para las columnas (ventajas) del análisis de los datos.

COORDENADAS Y CONTRIBUCIONES DE LOS RENGLONES

ONOMBRES	MASAS	DIST.²	COORDENADAS					CONTRIBUCIONES ABSOLUTAS					CORRELACIONES AL CUADRADO				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
FARM¤FARMING	.040	.59	.24	.31	-.06	.39	.45	3.1	2.5	.3	14.7	34.7	.20	.16	.01	.26	.34
FAR2¤FARM/FO	.043	.13	.07	-.11	.06	.02	.03	.1	.4	.3	.0	.1	.04	.10	.03	.00	.01
ENER¤ENERGY/	.020	.79	-.01	-.76	-.07	.20	.06	.0	7.9	.2	1.8	.4	.00	.74	.01	.05	.01
STEE¤STEEL	.041	.78	.42	.07	.23	-.36	.02	4.9	.1	3.6	12.9	.1	.47	.01	.14	.34	.00
CHEM¤CHEMICA	.021	.19	.17	-.05	.01	.12	.22	.4	.0	.0	.8	4.4	.15	.01	.00	.08	.25
WOOD¤WOOD/PA	.011	.72	.18	.38	.15	-.19	.52	.2	1.1	.4	.5	13.4	.05	.28	.03	.05	.38
AUT¤AUTO/AV	.052	.26	.26	-.32	.12	.09	-.05	2.4	3.6	1.2	1.0	.5	.27	.39	.05	.03	.01
TEXTI¤TEXTILE	.049	.42	.60	.08	.10	.08	.11	11.2	.2	.9	.7	2.3	.85	.01	.02	.01	.03
PHAR¤PHARMA.	.023	.44	.33	.30	.11	-.25	-.20	1.6	1.4	.5	3.4	4.1	.25	.20	.03	.14	.09
MANU¤MANUFAC	.362	.49	.06	.04	.16	.12	-.29	12.8	.1	1.1	2.6	22.0	.64	.00	.02	.03	.17
CONS¤CONSTRU	.070	.32	.49	.07	.06	.02	-.01	11.1	.2	.5	.1	.0	.75	.02	.01	.00	.00
FOOD¤FOOD/BP	.038	.64	-.21	.45	-.50	-.17	-.01	1.1	5.3	16.1	2.7	.0	.07	.32	.39	.05	.00
SBUS¤SMALL B	.074	.22	-.11	.47	-.30	-.12	.02	.7	12.0	10.8	2.6	.1	.04	.65	.24	.04	.00
MBUS¤MIDSEL.	.148	.16	.16	.27	-.09	-.17	-.04	.8	2.2	.7	3.2	.2	.15	.44	.05	.19	.01
ADMI¤ADMINIS	.122	.67	-.27	-.74	-.16	-.10	.06	5.6	44.9	5.5	5.1	3.4	.11	.81	.04	.02	.01
SO.S¤SOCIAL	.027	.81	-.44	.54	-.28	-.33	-.04	3.4	5.3	6.3	7.1	.2	.24	.36	.17	.14	.00
HE.S¤HEALTH	.080	.39	-.29	.03	-.26	.42	-.18	4.4	.0	9.4	25.6	11.7	.21	.00	.18	.49	.09
TEAC¤TEACHIN	.136	.64	-.64	.22	.43	.02	.00	35.9	4.5	41.8	.3	.0	.63	.08	.29	.00	.00
TRAN¤TRANSPO	.040	.43	.09	-.53	.10	-.26	-.11	.2	8.1	.7	6.8	2.1	.02	.65	.02	.15	.03
ELEMENTOS SUPLEMENTARIOS																	
TELE¤TELECOM	.011	.45	-.15	-.47	-.05	-.01	-.07	.0	.0	.0	.0	.0	.07	.59	.01	.00	.01

Cuadro 5.7 Coordenadas y contribuciones de las filas (empleos) para el análisis de los datos.


GRAFICA 1 DE 14 PUNTOS SOBRE LOS EJES 1 Y 2
EJE 1/ HORIZONTAL
EJE 2 / VERTICAL
.539
.514
.489
.464
.439
.414
.389
.365
.315
.290
.265
.240
.215
.191
.166
.141
.116
.091
.066
.041
.016
-.008
-.033
-.058
-.083
-.108
-.133
-.158
-.232
-.257
-.282
-.307
-.332
-.357
-.381
-.406
-.431
-.456
-.481
-.506
-.531
-.555
-.580
-.605
-.629
-.655
-.705
-.730
-.754
-.779
-.804
-.829
-.854
-.879
-.788
-.142
.105
.566
SO.S#SOCIAL
HUMA
SEUS#SMALL B
FREE
FOOD#FOOD/BR
COMP
WOOD#WOOD/PA
FARM#FARMING
MBUSI#MISCEL.
TEAC#TEACHIN
LINE
VARI
ATMO
AUTO
NEAR
STEECONS#CONSTRUTEXT
MANUFMAN
MONE
SOCI
SCHE
HE.S#HEALTH
INTE
CHEM#CHEMICA
SALA
FAR2#FARM/FO
AUT #AUTO/AV
TELE#TELECOM
TRAN#TRANSPC
ADMI#ADMINIS
ENER#ENERGY/
SECU

CAPITULO VI.

APLICACION DEL ANALISIS DE CORRESPONDENCIAS EN UN ESTUDIO SOBRE LA DISTRIBUCION DE 12 ESPECIES DE ARAÑAS CAZADORAS EN UN AREA DE DUNAS.

Para mostrar la aplicación de la técnica en un estudio de caso se emplearon los datos de Van der Art (1979), que pertenecen a la distribución de 12 especies de arañas cazadoras en un área de dunas. Los datos indican el número de individuos de cada especie, capturadas en un periodo de 60 semanas. En las 28 localidades donde se capturaron estas especies se midieron seis variables ambientales. Los datos de abundancia de cada especie se transformaron al obtener su raíz cuadrada, para ponderar hacia abajo la alta abundancia; los datos de las variables ambientales se transformaron al obtener su logaritmo.

En este estudio se desea reconocer las especies con distribución similar en localidades específicas, y localidades que presenten una composición semejante de dichas especies. Finalmente, explicar la variabilidad en la distribución de las especies entre las localidades, con el apoyo de información sobre el medio ambiente de esa área. Esto significa que el estudio de la distribución de 12 especies de arañas cazadoras en 28 localidades comprende dos etapas muy importantes: la primera de ellas se ocupa de representar las posibles relaciones que existan entre las especies y localidades de forma sencilla para su análisis, lo cual se consigue a través de la técnica de análisis de correspondencias ya que permite representar las especies y localidades como puntos en un espacio de dos dimensiones, y además proporcionar elementos de ayuda para asegurar un análisis confiable, esta etapa se le conoce como descriptiva, dado que sólo permite obtener la estructura interna de los datos. La siguiente etapa, que toma parte en este estudio, se encarga de interpretar los resultados que se obtienen de la etapa

descriptiva. Esta interpretación surge de un trabajo externo al análisis de correspondencias que requiere del conocimiento del investigador.

Los datos se arreglaron en una matriz colocando a las 12 especies en las hileras y las localidades en las columnas, cuadro 6.1. Las especies de arañas estudiadas son: *Alop acce = Alopecosa accentuata, Alop cune = Alopecosa cuneata, Alop fabr = Alopecosa fabrilis, Arct lute = Arctosa lutetiana, Arct peri = Arctosa perita, Aulo albi = Aulonia albimana, Pard lugu = Pardosa lugubris, Pard mont= Pardosa monticola, Pard nigr = Pardosa nigriceps, Pard pull = Pardosa pullata, Troc terr = Trochosa terricola, Zora spin = Zora spinimana.* Con este cuadro de datos se construyen los perfiles hilera y perfiles columna dando origen así a los puntos hilera y columna en los dos espacios $\mathbb{R}^n$ y $\mathbb{R}^m$ donde n=12 y m=28.

	NUMERO DE LOCALIDADES																											
ESPECIES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Arct lute	0	0	1	1	1	2	3	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pard lugu	0	1	1	1	1	0	1	7	0	0	0	0	1	1	2	2	1	2	3	3	4	0	1	0	1	0	0	0
Zora spin	2	3	1	5	4	5	4	1	0	0	0	0	4	5	1	2	1	0	1	1	1	0	0	0	2	0	0	0
Pard nigr	3	3	4	5	9	5	9	1	1	0	0	1	7	3	0	1	0	0	1	0	0	0	0	0	2	0	0	0
Pard pull	6	6	6	9	8	4	9	1	1	0	1	2	8	8	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Aulo albi	2	5	3	4	3	2	4	2	0	0	1	0	4	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Troc terr	7	8	8	9	9	7	9	5	1	1	2	3	9	9	5	5	4	5	4	4	4	1	1	0	4	1	0	1
Alop cune	3	1	4	2	4	2	2	3	1	0	1	3	6	1	0	1	0	1	1	1	1	0	0	0	1	0	0	0
Para mont	7	1	5	2	1	3	5	1	5	4	9	9	4	3	0	0	0	0	0	0	1	1	2	1	3	0	1	2
Alop acce	5	0	3	1	1	0	1	0	1	1	3	4	1	0	0	0	0	0	0	0	0	0	2	4	3	3	1	5	3
Alop fabr	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	4	3	4	3	2	3	3
Arct peri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	0	4	2	1

Cuadro 6.1 Datos de abundancia de 12 arañas cazadoras, con las especies como hileras y las las localidades como columnas.

El cuadro 6.2 contiene información adicional de las localidades que corresponden a las variables ambientales: porcentaje de agua en el suelo (AGUA), porcentaje de arena suelta (ARENA), porcentaje de cobertura de capa de musgo (COBMUSG), reflección de la superficie del suelo con el cielo poco nublado (LUZ), porcentaje de troncos caidos (TRONC), y porcentaje de cobertura de capas herbáceas (COBHERB). Estas seis nuevas hileras definen seis nuevos perfiles hilera que se proyectan posteriormente en la gráfica, su propósito es enriquecer la interpretación.

	NUMERO DE LOCALIDADES.																										
VARIABLES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
AMBIENTALES																												
AGUA	5	8	6	6	8	9	8	6	5	4	4	5	9	8	9	8	9	8	7	8	7	1	0	2	3	0	0	0
ARENA	0	0	0	0	0	5	0	0	0	2	0	0	3	0	0	0	0	0	0	0	0	7	6	7	7	9	5	7
COBMUSG	7	2	5	5	0	5	1	2	9	7	9	8	1	4	1	1	1	0	3	1	1	9	9	9	2	4	8	8
LUZ	8	3	8	6	5	1	5	1	7	8	8	3	7	2	1	0	2	2	0	0	0	8	9	9	5	9	8	8
TRONC	0	3	0	0	0	7	0	9	0	0	0	0	3	0	9	9	9	9	9	9	9	0	0	0	0	0	0	0
COBHERB	9	9	9	9	9	6	9	6	6	5	7	8	9	9	5	0	5	5	2	0	2	9	6	5	8	2	6	6

Cuadro 6.2 Medición de seis variables ambientales en 28 localidades.

Es importante que en el inicio del análisis todos los elementos se consideren como elementos activos, porque esto permite obtener una visión general del estudio. Con el análisis de los primeros resultados se puede averiguar que elementos son descritos ampliamente y en cuales se tiene dificultad para su descripción. Así, el análisis de correspondencias se inicia con las 12 especies y 28 localidades como elementos activos. Los primeros resultados que se deben comentar son los eigenvalores y porcentajes de variación explicada que se muestran en el cuadro 6.3. Es necesario recordar que la representación gráfica de las especies como puntos se encuentran en un espacio multidimensional con los ejes como localidades, y al contrario las localidades como puntos en un espacio multidimensional donde los ejes

representan las especies. Para poder explicar la variabilidad de las especies se recurre a una reducción del espacio multidimensional a un espacio de dos dimensiones, donde las especies y localidades se proyectan como puntos, y de esta manera conseguir una representación geométrica sencilla y fácil de interpretar. Para saber que ejes son los que presentan un porcentaje de variación más alto, y con ellos construir la gráfica, es necesario analizar el histograma de los primeros eigenvalores.

SUMA DE LOS EIGENVALORES 1.21130600

	EIGENVALOR	PORCENTAJE	PORCENTAJE ACUM.	
1	.58953800	48.67	48.67	¶¶
2	.25332360	20.91	69.58	¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶
3	.17852020	14.74	84.32	¶¶¶¶¶¶¶¶¶¶¶¶¶¶¶
4	.06463546	5.34	89.66	¶¶¶¶¶¶¶
5	.03809249	3.14	92.80	¶¶¶¶
6	.02596870	2.14	94.95	¶¶¶
7	.02307692	1.91	96.85	¶¶¶
8	.01591375	1.31	98.16	¶¶
9	.00935082	.77	98.94	¶¶
10	.00830361	.69	99.62	¶

Cuadro 5.3 Histograma de los primeros diez eigenvalores, con 12 especies y 28 localidades.

En este histograma, se observa que los primeros 3 ejes presentan 84.32% de la variación explicada total; así que es conveniente construir una gráfica con los dos primeros ejes principales donde se cuenta con 69.58% de la variación explicada total, si se construye una gráfica con el primer y tercer eje principal se tiene 63.41% de la variación, por lo que la diferencia con la primera gráfica es escasa, de modo que es conveniente analizar la primera gráfica, Gráfica 5.1.

El cuadro 6.4 presenta las coordenadas para las 28 localidades en los primeros cuatro ejes principales; las contribuciones absolutas, y las correlaciones al cuadrado. La primer columna muestra las masas; la segunda, la distancia ji-cuadrada entre los puntos y el centro de gravedad.

LOCALIDADES	MASAS	DIST.	COORDENADAS				CONTRIBUCIONES ABSOLUTAS				CORRELACIONES AL CUADRADO			
			F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
1	.063	.28	.02	.39	-.19	.08	.0	3.8	1.2	.7	.00	.55	.12	.02
2	.050	.58	.47	.03	.35	-.05	1.9	.0	3.4	.2	.37	.00	.21	.00
3	.066	.13	.12	.22	-.02	.01	.2	1.3	.0	.0	.12	.38	.00	.00
4	.070	.34	.41	.13	.34	-.05	2.0	.5	4.5	.3	.50	.05	.34	.01
5	.073	.41	.43	.13	.37	-.05	2.3	.5	5.7	.3	.44	.04	.34	.01
6	.054	.47	.42	.22	.33	-.03	1.6	1.9	3.2	.1	.36	.09	.22	.00
7	.084	.40	.37	.28	.39	.00	2.0	2.6	4.4	.0	.35	.19	.23	.00
8	.038	1.67	.47	-.95	-.46	.10	1.4	13.3	4.4	.6	.13	.54	.12	.01
9	.018	1.50	-.15	.77	-.78	.45	.1	4.1	6.0	5.5	.01	.39	.40	.13
10	.013	2.41	-.75	.66	-1.02	.04	1.2	2.2	7.2	.0	.23	.12	.43	.00
11	.032	1.61	-.46	.71	-.87	.22	1.2	6.4	13.8	2.4	.13	.31	.48	.03
12	.039	1.21	-.25	.63	-.77	.33	.4	6.3	13.0	6.6	.05	.33	.49	.09
13	.082	.23	.31	.17	.19	-.01	1.3	1.0	1.7	.0	.42	.13	.16	.00
14	.059	.39	.43	.10	.26	.00	1.8	.2	2.2	.0	.47	.02	.17	.00
15	.014	1.99	.49	-1.14	-.34	-.08	.6	7.3	.9	.1	.12	.65	.06	.00
16	.020	1.11	.49	-.79	-.13	-.08	.8	4.9	.2	.2	.22	.57	.02	.01
17	.011	1.80	.48	-.93	-.19	-.12	.4	3.7	.2	.2	.13	.48	.02	.01
18	.014	2.02	.46	-1.11	-.48	.07	.5	6.9	1.9	.1	.11	.61	.12	.00
19	.018	1.61	.50	-1.07	-.36	-.01	.8	8.2	1.3	.0	.16	.72	.08	.00
20	.016	2.10	.50	-1.26	-.50	.02	.7	10.0	2.3	.0	.12	.75	.12	.00
21	.021	1.68	.44	-1.01	-.57	.12	.7	8.7	3.8	.5	.12	.61	.19	.01
22	.016	4.63	-1.92	-.20	.09	-.79	10.1	.3	.1	15.7	.80	.01	.00	.14
23	.023	2.82	-1.63	-.28	-.10	-.02	10.5	.7	.1	.0	.94	.03	.00	.00
24	.018	5.58	-2.20	-.24	.30	-.34	16.4	.4	.9	3.1	.94	.01	.02	.02
25	.036	.49	-.41	.00	-.22	-.45	1.0	.0	1.0	11.3	.34	.00	.10	.42
26	.014	12.27	-2.66	-1.01	1.41	1.44	17.2	5.8	15.9	46.1	.58	.03	.16	.17
27	.020	4.96	-2.27	-.08	.11	-.19	15.1	.0	.1	1.1	.91	.00	.00	.01
28	.018	2.96	-1.66	.05	-.19	-.42	8.7	.0	.4	4.8	.93	.00	.01	.06

Cuadro 6.4 Coordenadas y contribuciones de las 28 localidades.

La contribución absoluta más alta se presenta en la localidad 26, significa que esta localidad tiene una gran importancia en la construcción del primer eje, y por lo tanto es

la que más explica este eje, aún por encima de las contribuciones que se presentan en el segundo eje. Es importante observar que la distancia que existe entre este punto con el centro de gravedad es muy alta, y su masa es baja, comparándola con las demás masas; las masas son un buen indicador para saber que localidad tiene mayor relevancia en el estudio. Esta localidad 26 no es muy importante para describir el patrón de distribución de las demás especies; sin embargo, su distancia, que es mayor a los demás, ocasiona que su contribución sea alta. Esta situación puede ocasionar que no se utilice con claridad el patrón de distribución de la comunidad de especies. Las contribuciones absolutas para las localidades 22, 23, 24, 27 y 28 son altas, comparándolas con las demás contribuciones; sin embargo su masa es baja, así que su alta contribución se ve influida por sus respectivas distancias con el centro de gravedad. No obstante, estas cinco localidades son similares en cuanto a su composición de especies. Las localidades 1, 2, 3, 4, 5, 6, 7, 8, 13 y 14 tienen una masa alta, esto significa que tienen gran importancia en este estudio, pero esto no se ve reflejado en sus contribuciones; esto se debe a que las seis localidades mencionadas anteriormente tienen contribuciones altas ocasionadas por sus distancias.

En cuanto a las correlaciones al cuadrado de las 28 localidades (cuadro 6.4), las localidades 22, 23, 24, 26, 27 y 28 tienen valores altos, esto indica que se encuentran muy cerca del primer eje. La correlación al cuadrado más alta, considerando el segundo eje, se presenta en las localidades 15, 18, 19, 20 y 21, lo cual indica que se encuentran más cerca al segundo eje. Aún cuando esta información es útil, porque ayuda a conocer que puntos no pueden ser confiables para ser interpretados por estar tan alejados al eje, es mejor identificar los puntos que estén mejor representados en los dos ejes al mismo tiempo. La calidad de la representación indica que las localidades 22, 23, 24, 26, 27 y 28 son las más confiables para su interpretación en la gráfica 6.1.

En el cuadro 6.5 se observan las coordenadas para las 12 especies en los primeros cuatro ejes principales, las contribuciones absolutas y las correlaciones al cuadrado. Además en la primer columna se muestran las masas; la segunda, la distancia entre los puntos y el centro de gravedad. También se muestran las coordenadas y correlaciones al cuadrado para los elementos suplementarios, las variables ambientales, además de sus masas y sus distancias ji-cuadradas entre estos puntos y el centro de gravedad.

ESPECIES	MASAS	DIST.	COORDENADAS				CONTRIBUCIONES ABSOLUTAS				CORRELACIONES AL CUADRADO			
			F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
Arct lute	.018	1.54	.48	.40	.64	-.06	.7	1.1	4.1	.1	.15	.10	.27	.00
Pard lugu	.059	2.80	.47	-1.44	-.62	.03	2.2	49.3	12.7	.1	.08	.74	.14	.00
Zora spin	.077	.50	.47	-.11	.31	-.15	2.8	.3	4.1	2.6	.43	.02	.19	.04
Pard pull	.125	.48	.40	.25	.35	.04	3.4	6.1	8.5	.4	.33	.26	.25	.00
Aulo albi	.059	.60	.49	.23	.35	-.06	1.6	1.2	4.1	.4	.27	.08	.21	.01
Troc terr	.225	.28	.32	-.32	-.04	-.02	3.8	9.0	.2	.1	.36	.37	.01	.00
Alop cune	.070	.40	.22	.01	-.19	.15	1.2	.0	1.3	2.3	.25	.00	.09	.05
Pard mont	.125	1.03	-.28	.65	-.69	.24	1.7	20.9	33.6	11.4	.07	.39	.44	.05
Alop acce	.075	1.80	-1.16	.28	-.33	-.19	17.7	2.3	4.6	4.1	.77	.04	.06	.02
Alop fabr	.047	4.78	-2.03	-.23	.13	-.70	52.4	1.0	.5	35.3	.86	.01	.00	.10
Arct peri	.021	11.74	-2.86	-.90	1.22	1.13	29.7	6.8	17.8	42.6	.70	.07	.13	.11
ELEMENTOS SUPLEMENTARIOS.														
AGUA	.274	.64	.33	-.52	-.34	.02	.0	.0	.0	.0	.13	.33	.14	.00
ARENA	.114	4.87	-1.86	-.22	.24	-.12	.0	.0	.0	.0	.71	.01	.01	.00
COBMUSG	.218	1.48	-1.00	.14	-.40	-.04	.0	.0	.0	.0	.68	.01	.11	.00
LUZ	.247	1.17	-.98	.13	-.14	.11	.0	.0	.0	.0	.85	.01	.02	.01
TRONC	.152	4.49	.61	-1.69	-.65	-.01	.0	.0	.0	.0	.08	.64	.09	.00
COBHERB	.290	.28	-.20	.02	-.20	.01	.0	.0	.0	.0	.14	.00	.14	.00

Cuadro 6.5 Coordenadas y contribuciones de 12 especies y 6 variables ambientales.

Las especies que tienen una mayor participación en la construcción del primer eje, y por lo tanto las que tienen más importancia en la variación explicada para ese eje son: *Alop acce*, *Alop fabr* y *Arct peri*. Las especies *Pard lugu* y *Pard mont* son muy importantes en la construcción del segundo eje y son las que tienen una mayor aportación en la explicación de este eje.

La especie que tiene la mayor masa es *Troc terr*; sin embargo, encuentra muy cerca del centro de gravedad. Su contribución es baja en ambos ejes. La especie con la menor masa es *Arct lute*, es una especie que se encuentra alejada al centro de gravedad, su contribución absoluta es muy pobre en ambos ejes, además de que no está bien representada.

Las especies que se encuentran más cerca al primer eje son: *Alop acce*, *Alop fabr* y *Arct peri* ya que su correlación al cuadrado es alta. La especie más cercana al segundo eje es *Pard lugu*. Las especies mejor representadas en ambos ejes son *Pard lugu*, *Alop acce*, *Alop faur* y *Arct peri*.

En cuanto a los elementos suplementarios sus masas son muy semejantes. Las variables AGUA y COBHERB se encuentran muy cerca al centro de gravedad; por otro lado, las variables ARENA y TRONC se encuentra más alejadas del centro de gravedad; las variables COBMUSG y LUZ se encuentran en punto medio de los dos grupos de varibles. Su correlación cuadrada es alta para el primer eje y casi nula en el segundo eje. Las variables mejor representadas en ambos ejes, y por lo tanto más confiables en la interpretación son ARENA, COBMUSG, LUZ y TRONC.

La gráfica 6.1, que cuenta con el 69.58% de la variación explicada, muestra que la especies *Arct peri* y la localidad 26 se encuentran demasiado alejadas de las localidades y especies. La contribución absoluta de estos dos puntos es muy alta. La especie *Alop fabr* tiene una contribución absoluta alta y las localidades que la rodean también tienen una contribución alta, además de que estos puntos están bien representados, ya que tienen una correlación alta, se puede asegurar que está especie presenta su máxima abundancia en las localidades 22, 23, 24, 27 y 28. La especie *Pard mont* presenta su máxima abundancia en las localidades 9, 10, 11 y 12, aún cuando la contribución de esta especie no es alta, se encuentra bien representada. La especie *Alop acce* se ubica en medio de estos dos grupos, es una especie

difícil de establecer cual es su distribución. Se ha mencionado insistentemente del papel de un punto que tienen una contribución alta ya que evita que otros puntos (localidades y especies) sean mejor explicados por un eje. Así que es recomendable eliminar del análisis la especie *Arct peri* y la localidad 26, para poder llevar a cabo un análisis en que se pueda explicar lo que ocurre con los demás puntos ubicados del lado derecho de la gráfica.

Al realizar el análisis de correspondencias con la especie *Arct peri* y la localidad 26 como elementos no activos, es decir que no participan en el análisis, se observa en el cuadro 6.8 que los 3 primeros eigenvalores tienen el 87.36% de la variación total explicada, más alta que en el anáisis anterior. Así que para los dos primeros dos ejes la variación explicada es del 73.48%, lo que indica mayor confiabilidad en los resultados que se obtengan.

	SUMA DE EIGENVALORES		1.0156e+00	
	EIGENVALOR	PORCENTAJE	PORCENTAJE ACUM.	
1	.5[illegible]	49.84	49.84	################################
2	.2[illegible]	23.62	73.46	###############
3	.1[illegible]	13.90	87.36	#########
4	.03[illegible]	3.79	91.15	###
5	.02[illegible]	2.28	93.43	##
6	.02[illegible]	1.97	95.40	##
7	.01[illegible]	1.55	97.[illegible]	#
8	.00[illegible]	.55	98.[illegible]	#
9	.00[illegible]	.84	99.85	#

Cuadro 6.8 Histograma de los primeros 9 eigenvalores, con 11 especies y 27 localidades.

Gráfica 6.1 46 puntos sobre el eje Horizontal y el eje Vertical. Presenta el 89.5% de la variación explicada. Los puntos Arct pers, TRONC y 26 han sido sacados de la periferia de la gráfica. Se presentaron 3 puntos multiples (Pard nigr-7), (COEMUSS-LUZ) y (4-5)

En el cuadro 6.7 se observa que las localidades 22, 23, 24, 27 y 28 tienen una contribución absoluta alta, comparándolas con las demás localidades. Su correlación cuadrada es mayor del 60%, esto indica que se encuentran cerca al primer eje, y por lo tanto tienen una buena calidad en su representación. Las localidades 9, 10, 11 y 12 contribuyen muy poco en la construcción de los dos ejes, pero esto se debe a que su masa es pequeña; sin embargo, su correlación es alta lo que significa que su representación es buena.

LOCALIDADES	MASAS	DIST.	COORDENADAS				CONTRIBUCIONES ABSOLUTAS				CORRELACIONES AL CUADRADO			
			F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
1	.064	.25	-.11	.31	.25	.05	.2	2.5	2.9	.4	.05	.27	.25	.01
2	.052	.54	.46	.16	-.27	.00	2.2	.5	2.7	.0	.39	.04	.13	.00
3	.068	.10	.02	.19	.05	.15	.0	1.0	.1	4.1	.09	.37	.02	.24
4	.072	.30	.38	.24	-.25	-.02	2.1	1.8	3.2	.1	.49	.29	.21	.00
5	.076	.38	.41	.25	-.28	.14	2.5	2.0	4.2	4.1	.44	.16	.21	.06
6	.055	.45	.39	.32	-.20	-.21	1.6	2.4	1.6	6.5	.33	.23	.09	.10
7	.087	.36	.33	.37	-.17	.09	1.8	4.9	1.6	.0	.29	.36	.09	.00
8	.039	1.59	.49	-.99	.27	.49	1.9	15.2	1.9	24.3	.15	.60	.04	.15
9	.018	1.43	-.38	.51	.97	-.05	.5	2.0	12.4	.1	.10	.16	.66	.00
10	.013	2.35	-1.09	.23	.37	-.44	3.0	.3	6.9	6.6	.50	.02	.32	.08
11	.033	1.55	-.74	.36	.89	-.13	3.6	1.8	18.5	1.4	.35	.09	.51	.01
12	.041	1.16	-.46	.36	.87	.06	1.8	2.2	21.5	.4	.20	.11	.65	.00
13	.085	.20	.25	.23	-.12	.15	1.1	1.7	.9	5.3	.32	.27	.07	.12
14	.061	.35	.40	.19	-.16	-.24	1.9	1.0	1.1	9.7	.46	.11	.07	.17
15	.015	1.92	.54	-1.13	.07	-.52	.9	7.9	.1	9.4	.15	.67	.00	.14
16	.020	1.06	.53	-.75	-.03	-.28	1.1	4.8	.0	4.1	.27	.54	.00	.07
17	.011	1.74	.53	-.90	-.02	-.73	.6	3.8	.0	15.2	.16	.47	.00	.33
18	.015	1.94	.51	-1.14	.25	-.12	.7	7.9	.7	.6	.13	.66	.03	.01
19	.018	1.54	.55	-1.1?	.12	.06	1.1	6.8	.2	.2	.20	.75	.01	.00
20	.017	2.01	.56	-1.28	.22	.00	1.0	11.3	.5	.0	.15	.81	.02	.00
21	.022	1.60	.46	-1.07	.35	.09	.7	10.6	1.9	.5	.13	.72	.08	.01
22	.015	5.67	-2.10	-.47	-.81	-.18	12.9	1.4	6.6	1.3	.78	.04	.12	.01
23	.020	2.86	-1.60	-.45	-.21	.19	10.2	1.7	.2	.6	.89	.07	.00	.00
24	.015	6.64	-2.40	-.39	-.79	.05	16.5	1.0	6.6	.1	.87	.02	.10	.00
25	.022	.50	-.59	-.19	-.13	-.10	2.6	.5	.4	1.0	.71	.07	.03	.02
27	.017	5.79	-2.22	-.24	-.47	.26	16.2	.4	2.2	3.0	.87	.01	.05	.01
28	.017	3.42	-1.82	-.22	-.24	-.10	10.8	.3	.7	.5	.97	.01	.02	.00

Cuadro 6.7 Coordenadas y contribuciones de las 27 localidades.

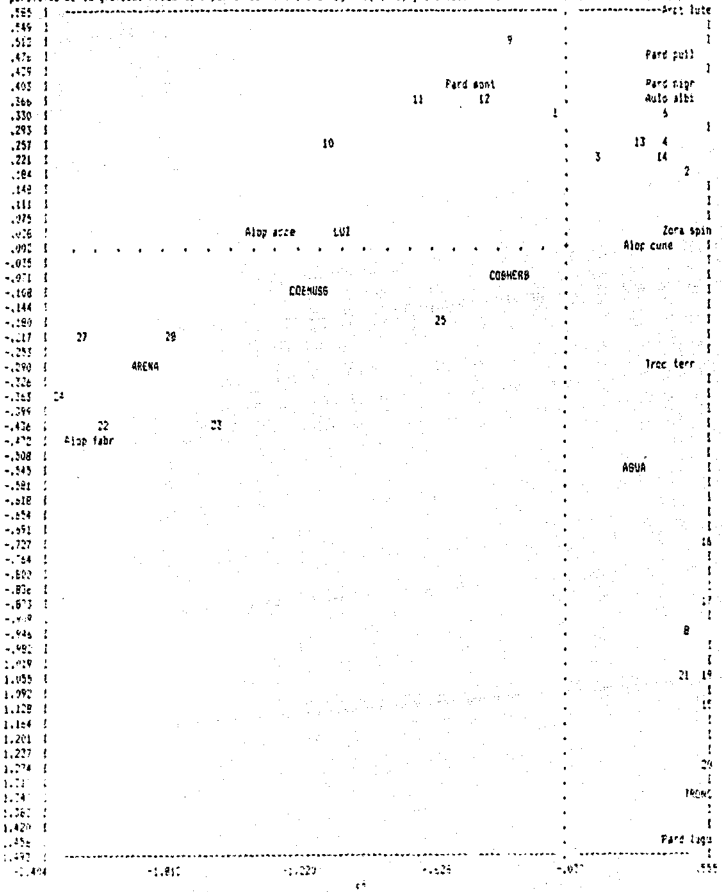
En el cuadro 6.8 la especie *Alop fabr* presenta una contribución alta, así que tiene una amplia participación en la construcción y explicación del primer eje. La especie *Pard mont*, que en el análisis anterior se veía agrupada con cuatro localidades, en este análsis su contribucción es alta y su calidad en la representación es baja, a pesar de esto se puede describir. En cuanto a la especie *Alop acce* tiene una contribución alta, pero esto se debe a su distancia que ji-cuadrada que es grande; sin embargo, su correlación al cuadrado indica que se encuentra muy cerca al primer eje y que en el segundo eje es casi ortogonal.

ESPECIES	MASAS	DIST.	COORDENADAS				CONTRIBUCIONES ABSOLUTAS				CORRELACIONES AL CUADRADO			
			F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
Arct lute	.018	1.46	.45	.58	-.45	-.11	.7	2.6	2.6	.6	.14	.23	.14	.01
Pard lugu	.061	2.70	.53	-1.49	.27	.25	3.3	56.4	3.2	9.8	.10	.82	.03	.02
Zora spin	.079	.46	.46	.01	-.30	-.27	3.3	.0	5.2	14.9	.46	.00	.20	.16
Pard nigr	.101	.46	.36	.39	-.27	.13	2.6	6.4	5.2	4.4	.29	.33	.16	.04
Pard pull	.129	.44	.35	.46	-.19	.09	3.2	11.2	2.9	2.7	.28	.48	.07	.02
Auto albi	.061	.56	.26	.33	-.25	.22	1.6	2.9	2.9	7.4	.23	.29	.12	.08
Troc terr	.230	.25	.32	-.29	-.01	-.21	4.6	8.2	.0	26.2	.39	.33	.00	.17
Alop cune	.072	.76	.25	-.91	.24	.35	.5	.0	2.8	23.5	.17	.90	.16	.35
Pard mont	.129	1.04	-.51	.39	.77	-.12	6.7	8.9	53.5	5.0	.25	.14	.56	.01
Alop acce	.076	2.17	-1.40	.01	.05	.16	29.3	.0	.2	4.9	.90	.00	.00	.01
Alop fabr	.044	6.08	-2.24	-.49	-.83	-.07	43.9	4.4	21.5	.5	.83	.04	.11	.00
ELEMENTOS SUPLEMENTARIOS														
AGUA	.282	.79	.30	-.57	.21	-.44	.0	.0	.0	.0	.11	.41	.06	.25
ARENA	.101	5.33	-1.92	-.32	-.54	-.43	.0	.0	.0	.0	.67	.02	.05	.03
COBMUSG	.217	1.75	-1.17	-.11	.19	-.19	.0	.0	.0	.0	.79	.01	.02	.02
LUZ	.238	1.29	-1.05	.01	.09	-.11	.0	.0	.0	.0	.86	.00	.01	.01
TRONC	.157	4.33	.70	-1.72	.26	-.61	.0	.0	.0	.0	.11	.68	.02	.08
COPHEPE	.295	.31	-.29	-.08	.13	-.14	.0	.0	.0	.0	.27	.02	.06	.07

Cuadro 6.8 Coordenadas y contribuciones de 11 especies y 6 variables ambientales.

En la gráfica 6.2, que explica el 73.46% de la variación explicada total, puede afirmarse que las especies *Alop fabr* y *Pard mont* estan plenamente definidas; aún más que en la gráfica 6.1 ya que es posible identificar con mayor claridad las localidades que presentan una mayor abundancia; para el caso de

Gráfica 6.2 Se cuenta con 44 puntos sobre el eje 1/horizontal y el eje 2/vertical. El punto TRONC ha sido sacado de la periferia de la gráfica. Presenta 3 puntos ocultos (Pard nigr-7), (4-5) y (15-15). Presenta el 73.4_% de la variación explicada



Pard mont, su mayor abundancia se localiza en la localidad 11 y 12; *Alop fabr* presenta su máxima abundancia en las localidades 22 y 24. En el caso de *Alop acce*, es una especie que su distribución no está bien definida.

Es importante considerar las variables ambientales, que son los elementos suplementarios que enriquecen la interpretación. Esta información es posible integrarla a la gráfica como elementos suplementarios ya que no participan en la formación de los ejes principlaes y tampoco en la inercia total. Las variables que cuentan con una correlación alta y que por lo tanto su representación en la gráfica es buena son: ARENA, COBMUSG Y LUZ. La información con que se cuenta hasta este momento permite esclarecer la distribución de 4 de las 12 especies. Así, se observa que la localidad 26 cuenta con la mayor abundancia de la especie *Arct peri*; *Alop fabr* tiene una mayor abundancia en las localidades 23, 27 y 28. Estas dos especies se ubican en áreas con un alto contenido de arena suelta. La especie *Alop acce* es un caso en que su distribución no está restringuida a un área en particular; se puede decir que se presenta en las localidades 10, 25 y 28; el área en que es posible encontrar esta especie tiene un alto porcentaje de cobertura de musgo y donde la reflexión de la luz en el suelo es grande. La especie *Pard mont* se ubica en las localidades 1, 9, 11 y 12, y un poco menos en la localidad 10.

El resto de las 8 especies, que se ubican del lado derecho de la gráfica 6.2, resulta un poco difícil describir su distribución. Sólo se puede mencionar que la especie *Pard lugu* está presente en las localidades 8, 15, 16, 17, 19, 20 y 21. Así que se realizó un nuevo análisis en el que sólo participaron estas 8 especies y las 17 localidades cercanas a ellas.

El primer resultado a analizar son los eigenvalores del cuadro 6.9; el valor de los 2 primeros cuentan con el 80.50% de la variación; así que con ellos se construyó la gráfica 6.3.

SUMA DE EIGENVALORES .36528010

	EIGENVALOR	PORCENTAJE	PORCENTAJE ACUM.	
1	.25068070	68.63	68.63	**********
2	.04337483	11.87	80.50	**********
3	.02792078	7.64	88.15	******
4	.02060049	5.64	93.79	****
5	.00992027	2.69	96.47	**
6	.00796887	2.18	98.65	*

Cuadro 6.9 Histograaa de los priaaros 6 eigenvalores, con 8 especies y 17 localidades.

El cuadro 6.10 muestra que las localidades 8, 14, 15 y 17 son las que contribuyen más en la construcción del primer y segundo eje principal. Las localidades que son importantes en este estudio, considerando sus masas, son 5, 7 y 13. Las localidades mejor representadas en la gráfica 6.3 son la 19, 20 y 21.

LOCALIDADES	MASAS	DIST.	COORDENADAS				CONTRIBUCIONES ABSOLUTAS				CORRELACIONES AL CUADRADO			
			F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
1	.061	.18	.25	-.06	-.24	-.19	1.6	.6	12.7	11.3	.36	.02	.33	.22
2	.072	.22	.19	.03	-.34	.18	1.1	.1	30.7	11.7	.17	.00	.53	.15
3	.075	.13	.18	-.20	-.07	-.16	1.0	7.1	1.4	8.9	.25	.32	.04	.19
4	.096	.10	.27	.07	-.09	.12	2.8	1.0	3.0	7.1	.70	.04	.08	.15
5	.104	.13	.28	-.09	.10	-.08	3.3	2.0	3.5	3.5	.63	.06	.07	.05
6	.072	.25	.32	.20	.28	.00	3.0	6.5	20.5	.0	.42	.16	.32	.00
7	.110	.22	.37	.01	.22	.11	5.1	.0	18.5	6.6	.62	.00	.21	.05
8	.052	1.12	-.92	-.46	.02	.21	15.2	26.6	.1	11.5	.76	.19	.00	.04
13	.107	.12	.25	-.18	.00	-.13	2.6	8.4	.0	8.3	.51	.28	.00	.13
14	.080	.15	.20	.25	-.08	.12	1.2	11.7	2.1	5.5	.25	.41	.05	.09
15	.021	1.21	-.96	.51	-.07	-.05	7.9	12.6	.4	.2	.76	.21	.00	.00
16	.029	.55	-.62	.27	.12	-.16	4.5	5.1	1.5	3.9	.69	.14	.03	.05
17	.016	1.09	-.75	.68	-.12	-.17	7.6	17.2	.6	2.2	.52	.43	.01	.03
18	.021	1.25	-1.00	.09	-.12	-.38	8.5	.4	1.0	14.7	.80	.01	.01	.11
19	.027	.90	-.92	-.03	.19	-.03	9.0	.0	3.3	.1	.94	.00	.04	.00
20	.024	1.26	-1.12	.03	.06	.00	12.0	.0	.4	.0	.99	.00	.00	.00
21	.029	1.26	-1.08	-.09	.03	.18	13.7	.5	.1	4.6	.93	.01	.00	.03

Cuadro 6.10 Coordenadas y contribuciones de 17 localidades.

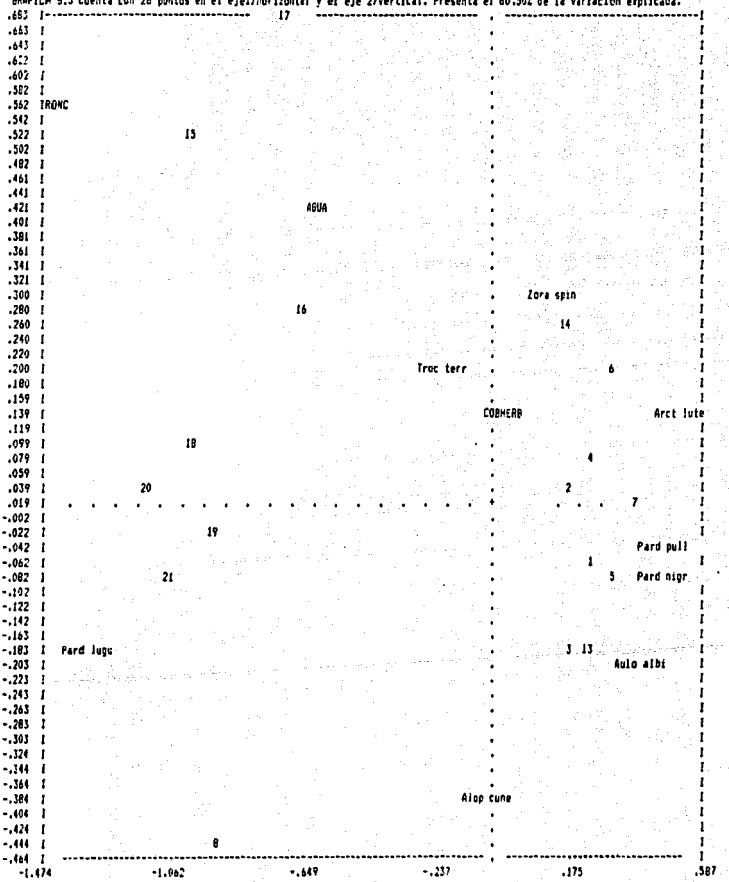
En el cuadro 6.11 se observa que la especie *Arct lute* tiene la masa más baja que las demás especies, su contribución absoluta es también la más baja; en cuanto a su correlación cuadrada indica que es casi ortogonal al segundo eje. Es una especie que no está bien representada en la gráfica, y por lo tanto no es muy importante en este estudio. Las especies *Pard lugu* y *Pard pull* presentan la contribución absoluta más alta, lo que significa que estas especies participan más que las demás en la construcción y explicación del primer eje. En cuanto al segundo eje, las especies *Zora spin* y *Troc terr* son las que tienen una mayor participación en la construcción y explicación de este.

ESPECIES	MASAS	DIST.	COORDENADAS				CONTRIBUCIONES ABSOLUTAS				CORRELACIONES AL CUADRADO			
			F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
Arct lute	.027	.93	.59	.12	.64	.15	3.7	1.0	39.0	2.8	.37	.02	.44	.02
Pard lugu	.083	1.94	-1.36	-.20	.10	.21	60.9	7.8	2.8	17.2	.95	.02	.00	.02
Zora spin	.110	.16	.09	.29	.08	.11	.3	21.0	2.5	6.1	.05	.52	.04	.07
Pard nigr	.136	.26	.42	-.10	.21	-.04	5.7	3.3	22.4	1.2	.70	.04	.18	.01
Pard pull	.176	.25	.44	-.05	-.14	.08	13.8	1.0	12.0	5.6	.80	.01	.08	.03
Aelo albi	.065	.31	.36	-.21	-.23	.29	3.3	8.0	15.3	16.4	.42	.14	.17	.13
Troc terr	.297	.12	-.25	.19	-.07	-.11	7.1	25.2	5.9	18.9	.50	.31	.05	.11
Alop cune	.088	.26	-.07	-.40	.03	-.27	.2	32.7	.2	21.9	.02	.62	.00	.29
ELEMENTOS SUPLEMENTARIOS														
ABLA	.345	.61	-.61	.40	.00	-.20	.0	.0	.0	.0	.60	.26	.00	.06
TRONC	.227	2.73	-1.47	.56	.14	-.28	.0	.0	.0	.0	.80	.12	.01	.03
COBHERB	.275	.23	-.06	.14	-.23	-.11	.0	.0	.0	.0	.02	.08	.23	.05

Cuadro 6.11 Coordenadas y contribuciones de 8 especies y 3 variables ambientales.

En la gráfica 6.3 se observa que la especie *Pard lugu* se ubica principalmente en las localidades 8, 18, 19, 20 y 21; sin embargo, es posible indicar que su mayor abundancia se localiza en la localidad 8, y no en la localidad 21, como se podría suponer ya que se encuentra más cerca de esta especie. Esto se debe a que la localidad 8 tiene una masa mayor que la localidad 21, lo que indica que es más importante, también se ve reforzado

GRAFICA 6.3 Cuenta con 26 puntos en el ejeI/horizontal y el eje 2/vertical. Presenta el 80.59% de la variación explicada.



porque la localidad 8 cuenta con una contribución más alta que la otra localidad; además de que está mejor representada que ésta última.

Las especies *Pard pull*, *Pard nigr* y *Aulo albi* se localizan en los mismas localidades: 1, 3, 5, 7 y 13; aunque, de acuerdo con las masas, las localidades 5, 7 y 13 indican que son más importantes que las otras localidades; además que tienen una mejor representación en la gráfica: También se observa que tienen una buena representación en la gráfica, sus contribuciones son más altas que de las localidades 1 y 3. El análisis anterior conduce a una descripción más detallada; la especie *Pard pull* tiene su mayor abundancia en la localidad 5 y 7; la especie *Pard nigr* en la localidad 5; y la especie *Aulo albi* en la localidad 13.

La especie *Zora spin* se distribuye principalmente en las localidad 14. La especie *Troc terr* tiene la masa más grande que cualquier otra especie, y su distancia es la más pequeña; esto indica que es una especie muy común, es decir que su distribución no está restringida a un área en especial, es una especie cosmopolita.

Finalmente, al tomar en cuenta las seis variables ambientales y el conjunto de especies se puede establecer que la especie *Arctosa perita* y *Alopecosa fabrilis* se localizan en habitats con un alto porcentaje de arena. *Alopecosa accentuata* y *Pardosa monticola* en habitats con un porcentaje medio de arena, el resto de las especies se localizan en habitat donde el porcentaje de arena es muy bajo. Para *Arctosa perita*, *Alopecosa accentuata*, *Alopecosa fabrilis* y *Pardosa monticola* se presenta el mismo rango de cobertura de musgo en el habitat. Con respecto al rango del contenido de agua se presenta en forma inversa. *Arctosa lutetiana*, *Pardosa pullata*, *Pardosa nigriceps*, *Aulonia albimana* y *Pardosa monticola* están presentes en habitats con un alto porcentaje de cobertura herbacea. *Pardosa lugubris* presenta una

posición poco comprendida, su habitat se caracteriza por un alto porcentaje de troncos y ramas. *Trochosa terricola, Zora spinimana y alopecosa cuneata* ocupan un sitio intermedio entre habitats con un alto porcentaje de pasto y de árboles.

CAPITULO VII.

DISCUSION.

En un gran número de estudios que se realizan en Biología es necesario reconocer las interrelaciones que existen en grupos de muestras, caracterizados por más de una variable, para poder comprenderlos con mayor claridad; este hecho se basa en agrupar muestras que comparten el mayor número de características semejantes. El caso de estudio que se presentó en el capítulo seis es sólo un ejemplo; en éste se describe el patrón de distribución de 12 especies de arañas en 28 sitios de una región específica. Para lograr lo anterior resulta necesario reconocer las especies con una distribución similar en los 28 sitios, y a la vez, los sitios que son similares en su composición de especies. Existen aun más ejemplos en los que es importante reconocer las interrelaciones entre entidades-variables; tal es el caso de un estudio en agronomía del que se desea conocer la relación que existe entre un grupo de especies cultivadas con diferentes tratamientos en un área específica; este mismo enfoque se puede presentar en estudios de taxonomía, en los que se busca identificar grupos de organismos relacionados por cierto número de cacarterísticas semejantes.

El análisis de correspondencias, una técnica estadística multivariada, es una herramienta muy útil en la descripción y análisis de la estructura interna de un conjunto de datos que en Biología suelen ser muy extensos, además de complejos. Para llevar a cabo esta tarea es importante conocer el fundamento estadístico y matemático de la técnica, ya que esto permite obtener el máximo aprovechamiento de la técnica, y con esto describir y analizar el conjunto de datos con mayor precisión. El análisis de correspondencias tiene la capacidad para representar a las muestras y variables como puntos en un subespacio de dos dimensiones, y el de conocer que cantidad de información, proporcionada por los datos se explica en esta gráfica. Con esta

representación no solo es posible analizar a las muestras en función de las variables, sino también a las variables en función de las muestras de forma conjunta. Es importante resaltar dos aspectos muy importantes que permiten definir la estructura de los datos. El primero de ellos, que los datos multivariados pueden ser expresados en combinaciones lineales a partir de las variables originales. El segundo, dirigido principalmente al hecho de que las muestras y variables están representadas en un subespacio de dos dimensiones, es que la inercia total para ambas nubes de puntos es igual, lo que significa que los ejes principales reflejan la misma cantidad de dispersión en ambas nubes de puntos; además de la correspondencia que existe entre los ejes principales de las dos nubes, dada por las fórmulas de transición.

Los datos que se requieren para este tipo de estudios se arreglan en una tabla, llamada comunmente matriz de datos. La matriz de datos cuenta con una doble entrada, con especies por un lado y sitios por el otro; es decir, cada hilera representa una especie y cada columna un sitio, en el caso del estudio de la distribución de la comunidad de especies de arañas, figurando en la intersección el número de veces que aparece una especie en cada sitio. En términos generales, la matriz de datos, que se obtiene en estos estudio, cuenta con hileras que representa a cada una de las muestras y columnas que representa a las diferentes variables, indicando en la intersección de éstas un valor positivo que expresa una frecuencia.

Para describir las posibles interrelaciones que existen en las especies y sitio, es posible conceptualizar la tabla de datos en dos espacios multidimensionales: espacio de especies y espacio de sitios. Estos dos espacios contienen toda la información de los datos originales, sin embargo, analizar dichos espacios resulta poco menos que imposible para el investigador, aún así la utilidad de ellos se refleja en dos aspectos muy importantes. 1) La relación de las especies, o sitios, se manifiesta en estos espacios, donde las especies similares se encuentran cerca una de

otra, mientras que las disimilares se encuentran alejadas; 2) La distribución de los sitios-punto están concentrados en un lugar específico, del mismo modo, las especies-punto muestran una estructura definida.

Tanto el espacio de las especies, como el de los sitios, cuentan con un gran número de dimensiones lo que hace imposible vizualizar la estructura interna de los datos. Sin embargo, el análisis de correspondencias puede llevar la información original a un subespacio tal que disminuyendo la dimensionalidad recoja la mayor inercia posible de los datos originales y en el que se representan las especies y los sitios de forma conjunta. La pérdida relativa de ciertos aspectos de la información, al llevar a cabo la reducción de la dimensionalidad, se recupera con una mayor simplicidad en la presentación de la información que contiene la tabla de datos.

Para una clara interpretación del conjunto de los datos, es importante que su contenido sea homogéneo, esto es, que se retengan los datos que estén relacionados únicamente a un solo punto de vista o un objetivo en particular. En todas las investigaciones que se llevan a cabo se registra una gran cantidad de información, no sólo referente a un objetivo en particular como es el caso del estudio de la distribución de las 12 especies de arañas, en el que se registran los valores de abundancia, sino también que se obtiene información adicional útil, como es el caso de las mediciones de variables ambientales. Esta información adicional no toma parte activa en el análisis de correspondencias, sin embargo, esta selección hace que la interpretación de los datos sea más fácil y clara. Una consecuencia inmediata de lo anterior es que se pueden identificar dos grupos de elementos; el primero de ellos tiene un papel activo en el análisis, el otro un papel suplementario, estos suelen dar información adicional que ayuda a comprender mejor la estructura de los datos, al ser representados en la gráfica con los elementos activos. La diferencia entre estos dos tipos de elementos es muy importante, ya que los elementos

suplementarios no contribuyen a la inercia total y por consiguiente a la construcción de los ejes principales. En algunas ocasiones las variables activas pueden tomar un papel suplementario en el análisis, facilitando aún más la interpretación ya que el porcentaje de la variación explicada aumenta y por lo tanto los resultados de este análisis sería más confiables, pero la decisión de que elemento sea activo o suplementario depende del investigador.

El análisis de correspondencias no sólo proporciona una gráfica de dos dimensiones en la que resume la información de la tabla de datos original, sino que también proporciona ayuda en la interpretación de tales resultados.

Cada eje principal tiene un porciento de variación explicada, expresada por el eigenvalor, de la información contenida en los datos; con la contribución absoluta se puede saber de que manera contribuye cada una de las variables a esa explicación. Cuando el valor es alto para una variable en particular, con respecto a las demás variables, se establece que su máxima explicación se encuentra en este eje. Se debe tener cuidado en la interpretación de las contribuciones ya que se podría pensar que sólo las conribuciones con un valor superior a un 60% pueden ser tomadas en cuenta; esto no necesariamente ocurre así, se deben considerar los valores de las contribuciones de forma conjunta y analizarlos sólo con respecto a ese estudio.

Un apoyo muy importante para analizar las contribuciones absolutas son las correlaciones al cuadrado, por que estos valores pueden indicar que tan cerca o alejado se encuentra una variable con respecto a cada eje principal. Esto se traduce en establecer si una variable cualquiera se encuentra bien representada o no.

Cada muestra y variable cuenta con una masa, con esta información se puede identificar que elemento tiene una mayor participación, con respecto al grupo de muestras y variables, en el estudio. Esto es útil para el análisis, porque aquellos

elementos que cuentan con una masa muy pequeña es un claro indicio de que su participación en el estudio es mínimo y que si se elimina de éste puede ocasionar que el porciento de la variación explicada por la gráfica aumente, y por lo tanto los demás elementos estén mejor representados en la gráfica. Cuando la masa de un elemento cualquiera es grande, implica que es muy importante en el estudio; estos elementos son facilmente interpretados y analizados, por lo que es posible analizarlos como elementos suplementarios, ya que este tipo de elementos tiene una fuerte contribución en la inercia total y ocasiona que otros elementos no sean fácilmente interpretados.

No sólo se consideran las masas para tomar la decisión de que un elemento en particular actue en forma suplementaria, activa e inclusive sea eliminado del análisis, también se toma en cuenta una más de las ayudas del análisis de correspondencias que es la distancia ji-cuadrada. Cuando la distancia es grande, la nube de puntos está muy dispersa y los puntos más alejados del centro de gravedad pueden considerarse como elementos aberrantes, es decir que su participación no es indispensable en el estudio; lo que puede ocasionar es que sea difícil determinar las principales tendencias de variación. Si los puntos están muy cerca al centro de gravedad su contribución a la variabilidad es menor. Así, con la masa y distancia ji-cuadrada el análisis de los datos y de la gráfica se enriquece.

El apoyo del programa CORAN es muy valioso, realiza con rapidez gran cantidad de cálculos numéricos que se requieren en el análisis de correspondencias. La presentación de los resultados es un punto que debe sobresalir. Paquetes estadísticos como el NT-SYS y BIOMECO, que cuentan con el análisis de correspondencias, presentan los resultados en diferentes archivos, las ayudas como son las contribuciones absolutas y correlaciones al cuadrado se obtienen en distintos archivos tanto para las hileras como para las columnas de la tabla de datos, esto mismo ocurre con los eigenvalores y eigenvectores, y sin una presentación fácil de visualizar, y no sólo eso, para obtener los

gráficos es necesario entrar en otra rutina del programa y realizar una serie de instrucciones, no muy sencillas, para su obtención. CORAN presenta un solo archivo donde se cuenta con toda la información necesaria para la interpretación de la estructura interna de los datos, inclusive las gráfica. Estos datos se pueden imprimir rápidamente y analizarse. En primer término aparece la selección de los parámetros descritos en el capítulo cinco, y la tarjeta de parámetros para CORAN; después aparece el histograma de los eigenvalores en el que es muy fácil averiguar el porcentaje de variación explicada para cada eje principal. Las coordenadas y ayudas de las hileras y columnas se presentan en dos cuadros, donde la información se puede visualizar y analizar con suma rapidez. Las gráficas las realiza el propio programa, solo es necesario indicar en el archivo de entrada de datos el número de gráficas que se desean analizar; aún cuando solo presenta gráficas con los ejes (1,2) (2,3) (3,4)..., la rapidez y sencillez de este programa sobresale sobre otros programas. Se ha manifestado que en un estudio no sólo se requiere de uno solo análisis, sino en ocasiones es necesario llevar a cabo más de uno, en los que se modifica la forma en que participan las variables, para obtener resultados que describan ampliamente el conjunto de datos, con CORAN sólo es necesario modificar los identificadores de hileras y columnas

CAPITULO VIII.

CONCLUSIONES.

Describir y analizar las posibles interrelaciones que existen en grupos de muestras, presentes en gran número de estudios realizados en Biología, es sólo uno de los pasos que conducen a una comprensión más amplia y bien definida de tales estudios. El gran número de variables que caracterizan a cada una de las muestras y la compleja relación que existe entre ellas dificulta la comprensión de estos estudios. Por este motivo es importante contar con una herramienta capaz de realizar esta tarea de forma objetiva y sobre fundamentos teóricos bien establecidos. El análisis de correspondencias es una herramienta estadística que puede extraer la máxima información de un conjunto de datos, arreglados en una tabla de contingencia, resultado de múltiples observaciones y mediciones realizadas en campo, al proyectar las hileras y columnas como puntos en un subespacio de dos dimensiones donde hileras y columnas similares están juntas, representando muestras y variables respectivamente, y las disimilares se encuentran alejadas.

Conocer el fundamento estadístico y matemático de la técnica análisis de correspondencias no sólo permite realizar una interpretación más confiable de los resultados que ésta genera, sino que garantiza el buen manejo de la técnica, estos es, saber que tipo de datos maneja, que es lo que se espera en los resultados y como interpretarlos. El analisis de correspondencias analiza la interdependencia entre un grupo de variables que puede abarcar desde la independencia total hasta la colinealidad, cuando una variable es combinación lineal de otra variable. En términos más generales, se analiza la posible similitud que existe en un grupo de variables, identificando aquellas que son más semejantes y agrupándolas.

El aspecto más relevante del análisis de correspondencias es el de presentar las interrelaciones de las hileras y columnas de la tabla de contingencia en una sola gráfica, lo cual se debe a la relación de dualidad que existe entre ambas nubes de puntos y no solo eso, cuenta con un conjunto de ayudas muy importantes que garantizan la interpretación de la gráfica con mayor precisión y así obtener un análisis objetivo de los datos.

El apoyo que proporciona el programa CORAN resulta de gran valor, ya que es muy sencillo de manejar, a diferencia de otros paquetes estadísticos en los que se requiere de un aprendizaje prolongado para su manejo. CORAN proporciona todos los datos necesarios para la interpretación de la gráfica en cuadros que son sencillos de visualizar; además, el hecho de que proporcione la gráfica sin la necesidad de efectuar más pasos, reduce el tiempo del manejo del programa y permite realizar un mayor número de análisis para obtener mejores resultados. Si se desea modificar la forma en que participan las variables, sólo es necesario modificar los identificadores de hileras y columnas permaneciendo todo la demás información igual.

La aplicación del análsis de correspondencias no debe limitarse a estudios que se llevan a cabo en ecología de comunidades, donde esta técnica se ha aplicado con mayor frecuencia; su potencial de aplicación debe extenderse a las diferentes área de la Biología ya que se ha observado que esta técnica no requiere de ningún supuesto, excepto el que se deba contar con una tabla de contingencia que contenga los datos. Este hecho no limita la técnica sino por el contrario, existe una gran cantidad de estudios en los que en algun momento se cuenta con este tipo de tablas que deben ser analizadas.

Una técnica estadística como lo es el análisis de correspondencias no puede responder por sí sola a las preguntas que se plantean en el estudio de un fenómeno. Es una herramienta que facilita la investigación, al manejar gran cantidad de

información. El investigador es quien interpreta y encuentra soluciones al problema, que de otra forma sería difícil de llevar a cabo, ya que el gran número de variables involucradas en el estudio y la complejidad de sus interrelaciones haría imposible de comprender. Es una herramienta que ayuda a generar nuevos estudios o a comprobar hipótesis que se plantean en el estudio inicial ya que revela las relaciones existentes entre las variables.

B I B L I O G R A F I A.

BIBLIOGRAFIA CITADA.

Fisher C. (1940). On non linear species responses models in ordnation. Academics Press, U.S.A. pp. 204

Kendall D. G. and Stuart A. (1961). The advanced theory of statistics. Griffin London. pp. 260

Lebart L., Morineau A. and Warwich K.M. (1984). Multivariate Descriptive Statistical Analisys, Correspondence Analisys and Related Techniques for Large Matrices. John Wiley & Sons, New York. pp. 231.

Pichardo G.L.M.G. (1990). Eliminación de dos Problemas Resultantes da la Aplicación del Análisis de Correspondencias a la Fitogeografía. Tésis de Maestría en Estadística Aplicada. Colegio de Postgraduados. Chapingo. México. pp. 154.

Pielou E.C. (1984). The interpretation of Ecological Data, a Primer on Classification and Ordination. John Wiley & Sons.Inc. New York. pp. 263.

Ruiz R.J. (1989). Análisis de Correspondencias Simple y Múltiple y sus Aplicaciones a un Estudio de Caso. Tésis de Maestría en Estadística Aplicada. Colegio de Postgraduados. Chapingo, México. pp. 98.

Volle M. (1985). Analyse des Donnés. Ed. Economica. Paris, France.

Van der Art, P.J.M. and N. smeek-Enserink. (1975). Correlations Between Distributions of Hunting Spiders (Lycosidae, Ctenidae) and environmental Characteristics in a Dune Area. Netherlands Journal of Zoology. 25(1) pp 1-45.

Zárate L.G.P. y Alvarez S.M.O. (1985). Aplicaciones de las Descomposiciones Singular y Espectral de una Matriz. Colegio de Postgraduados. Chapingo, México. Agrociencia 61(103-126)

BIBLIOGRAFIA CONSULTADA.

Cajo J.F. and Ter Braak. (1986) Canonical Analysis, A New Eigenvector Technique For Multivariate Direct Gradient Analysis. Ecology 67(5) pp 1167-1179.

Dewit C.T. and Gourdiaan J. (1978). Simulation of Ecological Processes. Pudoc, Netherlands. pp. 175.

Digby P.G.N. and Kempton R.A. (1987). Multivariate Analysis of Ecological Communities. Chapman and Hall, London. pp. 206.

Gauch H.G. (1985). Multivariate Analysis in Community Ecology. Cambridge University Press. New York. pp.298.

Gittins R. (1969). The Application of Ordination Techniques. British Ecological Society 9(37-66)

Green H.R. (1979). Sampling Design and Statistical Methods for Enviromental Biological. Jhon Wiley & Sons, U.S.A.

Hair J.F. and Anderson R.E. (1979). Multivariate Data Analysis. Petroleum Publishing Company. Oklahoma, U.S.A. pp. 360.

Halfon E. (1979). Theoretical Systems Ecology Advanced and Case Studies. Academic Press, U.S.A. pp. 516.

Jongman R.H.G. and TerBraak C.J.F. (1987). Data Analysis in Community and Landscape Ecology. Pudoc Wageningen, Netherlands. pp. 299.

Leduw J. and Heigden G.M.P. (1988). Correspondence Analysis of Incomplete Contingency Tables. Psychometrika 53:2(223-233).

Legendre L. and Legendre P. (1983). Numerical Ecology. Elsevier Scientific Publishing Company, New York. pp. 419.

Legendre P. and Legendre L. (1987). Developments in Numerical Ecology. Springer-Verlag Nato ASI Series, New York. pp. 585.

Lomnicki A. (1988). The Place of Modelling in Ecology. OIKOS. 52(139-142)

Ludwig J.A. and Reynolds J.F. (1988). Statistical Ecology a Primer on Methods. John Wiley & Sons, New York. pp. 337.

Michel D. (1977). Statistical Analysis in Geology, Correspondence Analysis. Quarterly of the Colorado School of Mines 72(1-59).

Muller-Dombis D. and Ellenberg H. (1974). Aims and Methods of Vegetation Ecology. John Wiley & Sons, New York. pp. 547.

Mutis G.H.E. (1989). La Economía Mexicana en el Periodo 1939-1979, Una Aplicación de los Métodos Multivariados. IV Foro Nacional de Estadística Aplicada. I.I.M.A.S. - U.N.A.M.

Nash J.C. and Shlien S. (1987). Simple Algorithms for the Partial Singular Value Decomposition. The Computer Journal 30:3(268-275)

Pielou E.C. (1976). Mathematical Ecology. A Wiley-Interscience Publication. John Wiley & Sons, Inc. New York. pp. 385.

Shirley D. and Stanley W. (1983). Statistics for Research. John Wiley & Sons, New York. pp 210.

Sokal R.R. y Rohlf F.J. (1979). Biometria. H. Blume , Madrid España.

Steel G.D.R. and Torrie H.J. (1981). Principles and Procedures of Statistics, a Biometrical approach. McGraw-Hill. pp 240.

Williams W.T. and Lambert J.M. (1960). Multivariate Methods in Plant Ecology. J. Ecol. 47(83-101)

Zar J.H. (1984). Biostatistical Analisys. Prentice-Hill Inc. Englewood Cliffz, N.J. pp. 718.