UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

LA SEGURIDAD PÚBLICA EN MÉXICO: UN ANÁLISIS DE CONGLOMERADOS Y ESCALAS MULTIDIMENSIONALES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

ADRIANA NÓHPAL DE LA ROSA

DIRECTOR DE TESIS:
FRANCISCO SÁNCHEZ VILLARREAL

2007

Hoja de Datos del Jurado

1. Datos del alumno
   Nóhpal
   De la Rosa
   Adriana
   55 54462953
   Universidad Nacional Autónoma de Mexico
   Facultad de Ciencias
   Actuaría
   Número de cuenta: 91221663

2. Datos del Tutor
   Actuario
   Francisco
   Sánchez
   Villarreal

3. Datos del Sinodal 1
   M. en I.
   Roberto
   Chávez
   Manjarrez

4. Datos del Sinodal 2
   M. en I.
   Paola Alejandra
   Pavón
   Moreno

5. Datos del Sinodal 3
   Matemática
   Margarita Elvira
   Chávez
   Cano

6. Datos del Sinodal 4
   Actuario
   Emilio
   Gutiérrez Calderón

7. Datos del Trabajo Escrito "La Seguridad Pública en México: Un Análisis Conglomerado y de Escalas Multidimensionales"
   161 páginas
   2007

Agradecimientos

Quiero agradecer primero a las personas más maravillosas que la vida me ha presentado: *mis padres*.

A **mi mami**, la mujer más fuerte y amorosa que conozco, a ella que me ha ayudado a llegar hasta donde estoy con su ejemplo, a ella que ha sabido regañarme y consolarme cuando más lo he necesitado. Gracias Sarita por no dejarme caer y sobre todo por enseñarme a levantarme.

A **mi papi**, la persona con paciencia y generosidad infinitas. Gracias por recordarme que siempre hay alguien que me espera con los brazos abiertos y sobre todo gracias por estar cuando más lo he necesitado.

A mis hermanos **Consuelo y Carlos** que han caminado junto a mi, aunque sea en mundos paralelos. Siempre los llevo en mis pensamientos y en mi corazón.

A **Paola P.M.** por compartir sus conocimientos, por ser tan paciente y sobre todo por ser tan generosa. Gracias por brindarme tu amistad incondicional. Eres una persona única.

A uno de los mejores maestros que he tenido **Cesar**, que sin sus explicaciones y cariño no habría podido terminar esta tesis.

A **Francisco S.V.** por compartir su tiempo y conocimientos conmigo.

A **Susana B.O.** por todo el apoyo que me diste.

A **David T.** por enseñarme que mi único obstaculo soy yo.

Por supuesto, gracias a **Trini**, por enseñarme que sólo basta una sonrisa para iluminar un cuarto entero.

Al *arte* que con sus infinitas formas me mostró el significado de una línea y a los claroscuros que han iluminado mi vida.

A mis amigos: *Caro C., Dafne B., Eduardo E., Marta R. Raquel B, Rocío V.*

Por último, al sueño más hermoso que he tenido a Itzolinki.

Índice general

Índice de figuras

Índice de cuadros

Resumen

El empleo creciente de los métodos del Análisis Multivariado, junto con la capacidad actual de procesamiento de datos múltiples de los paquetes estadísticos, han dado la oportunidad de abordar exitosamente un gran número de problemas que involucran múltiples variables. Así mismo, han abierto nuevas posibilidades de aplicación, que de otra forma no serían factibles de análizar o bien muy tediosas de analizar y siempre se cargaría con un error. En el presente trabajo se introducen dos de estos métodos para hacer una comparación objetiva entre índices de delito por Entidad Federativa. Uno de ellos brindará una representación dimensional entre estados y otro dará una posible clasificación. Primero se discutirán estos dos métodos de análisis y por último se llevara a cabo su aplicación con la ayuda del paquete estadístico SPSS. Estos análisis pretenden generar una nueva perspectiva de análisis de la Seguridad Pública.

Capítulo 1

Introducción

Existe un gran número de herramientas estadísticas para el análisis de datos, en particular cuando se habla de múltiples datos-varibles existe una rama específica y especializada que afronta este tipo de análisis. Esta es la llamada Estadística Multivariada o más popularmente conocida como Análisis Multivariado. Esta rama conjunta casi todas, sino bien todas, las herramientas necesarias para abordar problemas de orígenes diversos y para propósitos diferentes. Entre estos métodos existen los que se utilizan para clasificar un conjunto de variables o bien para generar una taxonomía, si ésta es la intención. Otros se enfocan a la reducción de variables para un más fácil manejo o bien para deshechar aquellas que no son muy útiles en el análisis. Algunos más se enfocan en encontrar la relación entre las variables para así generar un modelo que sirva para predecir el comportamiento futuro de la variables o solamente para conocer la relación entre dos conjuntos de variables. Existen también aquellos que sirven para probar hipótesis sobre modelos complejos de múltiples variables.

El Análisis Multivariado ha sido empleado en diferentes disciplinas por su capacidad y fiabilidad. Disciplinas como la sociales han venido utilizando estas herramientas por su fácil manejo y objetividad, en problemas concernientes con el comportamiento humano. La mercadotecnia, por igual, ha empleado éstas para entender el comportamiento de los consumidores en el mercado, para la ubicación de nichos y para mejorar estrategias de mercado, por mencionar algunas de las aplicaciones. La psiquiatría ha creado modelos para caracterizar propiamente un síndrome, crear clasificaciones para éste, así como para diferenciar un sindrome de otro; y así canalizar y atender adecuadamente a un paciente. Los biólogos crean taxonomías,los silvicultores ubican una especie y su medio ambiente. Estas son algunas de las aplicaciones de la estadística multivariada por mencionar algunas. Las aplicaciones dependen no solo de la motivación de análisis sino también del tipo de variable que se está analizando.

El tipo de variable será determinante en la selección del método multivariado, de aquí que saber exactamente que tipo de variable se esta empleando es el punto de origen para cualquier análisis. Los métodos de análisis multivariado están ligados principalmente al objetivo y al tipo de variable. En este trabajo se darán los puntos necesarios para clasificar el tipo de variable [2] y se dará una sintesis de varios métodos dependiendo de su objetivo [3].

Este trabajo se centra en la exposición de dos técnicas multivariadas que son utilizadas principalmente como métodos descriptivos para múltiples datos. Estas dos técnicas, el Análisis de Conglomerados [5] y el de Escalas Multidimensionales [6] conjuntan un gran número de métodos de proximidad [4]. Estos métodos de proximidad a su vez están divididos en coeficientes de similaridad y disimilaridad. El primero además emplea métodos de agrupación entre grupos

como vinculación simple, el método Ward, vinculación completa, vinculación media entre otros [4]. Su aplicación se verá reflejada sobre datos referentes a los tipos de delitos registrados con su empleo se pretende encontrar una clasificación natural así como la representacion bidimensional de los datos [8]. Se explicará como se realizan estos análisis, paso a paso, en el paquete SPSS [8]. Las características de este paquete son presentadas en resumen en el capítulo 7.

Capítulo 2

Medidas y Escalas de Medida

Norman Robert Campbell, en su libro The Foundations of Science, dedica más de la mitad de sus ensayos al desarrollo de los fundamentos de la *Medición*. Da una serie de argumentos que lo llevan a concretar la noción de medición. En el camino, señala la importancia de ésta, sus reglas, sus leyes y su correcto uso. Y aunque él no define todo lo que se abarca en este capítulo, sus ideas sí son un punto muy importante de partida. En particular sus ensayos hablan de la importancia de este tema en las ciencias y de lo importante que es entenderlo.

Siendo la estadística una ciencia, necesita de la compilación de datos para estimar o inferir sobre un fenómeno. El fenómeno o hecho se tiene que poder representar de alguna manera para que pueda ser medido. Lo que implica que sin un apropiado entendimiento de la medida, no se puede estudiar algo científicamente. Siguiendo esta lógica se llegará de una manera u otra a la jerarquización o categorización de la medida, que permitirá clasificar los procedimientos estadísticos permisibles y/o adecuados. De hecho el uso de categorías en la selección o en la recomendación de un método de análisis estadístico es de suma importancia, ya que sin una apropiada categorización se puede caer en errores que se reflejarán en los resultados. Especialmente, si estas categorías no describen los atributos de los datos reales, éstos no servirán para arrojar un análisis apropiado y/o confiable.

La compilación de datos, no es más que una serie de medidas u observaciones realizadas. Estas observaciones pueden involucrar medidas de diferente naturaleza dependiendo de si transfieren información *cuantitativa o cualitativa*. De igual manera, existen jerarquías entre variables. Para sugerir los procedimientos estadísticos permisibles dentro de este tipo de escalas, el psicólogo S.S. Stevens creo una taxonomía entre variables, que será explicada. Esto nos llevara a la articulación de esté capitulo, que pretende ser el inicio de un apropiado análisis estadístico.

2.1. La Noción de Medir

Uno de los nexos que existe entre la medición y las matemáticas, es el proceso de medir, que es el proceso de convertir hechos y nexos empíricos en un modelo formal (modelo que es tomado de las matemáticas). De aquí que la medida no es más que la acción y el efecto de medir. A través de los años un sin número de científicos se ha interesado en este tema y como resultado han surgido diferentes definiciones de este concepto. Sin embargo la siguiente cita (Campbell,2002), que

pertenece a Sir Kelvin William Thompson, es la más significativa.

> When you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.[1] [8]

Puesto que medir es algo que se hace en forma rutinaria sin poner mucho cuidado al hacerlo, es necesario definir apropiadamente este concepto. Entonces el concepto de *Medir* será: la comparación de una cantidad con su respectiva unidad, con el fin de averiguar cuántas veces la primera contiene a la segunda. Por esta razón es importante notar que este concepto, como actividad, implica más que la estrecha actividad de hacer mediciones numéricas de longitud, área o volumen.

2.2. La Medición

La estadística como cualquier ciencia se encuentra fuertemente ligada al concepto de medir y ésta a su vez, con la de medición. Esto es: para poder estudiar algo científicamente se tiene que poder expresar en números "ese algo", para así poder distinguir los conceptos de una manera más sencilla. Y puesto que la medida surge de la idea de comparar, su lenguaje es el de la comparación. Al llegar a este punto se podría decir que dentro de la conciencia humana, consciente o inconscientemente, siempre está comparando cosas. Campbell indica que la medición es:

> Medición es una de las varias nociones que la ciencia moderna ha tomado del sentido común. La medición no aparece ciertamente como parte del sentido común hasta que no se alcanza un estadio de civilización relativamente alto; por otra parte, ya la concepción de medición propia del sentido común ha cambiado y se ha desarrollado enormemente en el curso de los tiempos históricos.[7]

La cita anterior enmarca el concepto de medición, que procede originalmente de un pequeño libro de divulgación científica, *What is Science*, publicado en 1921 y escrito por Norman Robert Campbell. Obviamente, como muchas otras palabras, la palabra *medición* tiene muchos significados. Sin embargo en matemáticas *medición* se entiende como el proceso por medio del cual se asigna un número a una propiedad física de algún objeto o conjunto de objetos con propósitos de comparación, de acuerdo a un conjunto de reglas o propiedades predeterminadas. Por lo tanto, el nombre de medida se reserva, para denotar el número de unidades de la propiedad dada. Entonces por medio de la medición, los atributos de nuestras percepciones se transforman en entidades conocidas y manejables llamadas "números". En breve, se dirá que la medición es la atribución de valores numéricos a las propiedades de los objetos.

2.3. Tipos de Clasificación

Debido a que la manera de expresar una propiedad en número no es evidente, se recurre la mayoría de las veces a una segunda o tercera forma de medición. Los procedimientos adoptados determinarán un grado de validez así como la manera

[1] Cuando se puede medir aquello de lo que se está hablando y expresarlo en números se sabe algo; pero cuando no se puede expresar en números, el conocimiento es magro e insatisfactorio. (Trad. del Autor)

precisa de interpretarlos. No obstante la medición adoptada, se exige la introducción del lenguaje del concepto cuantitativo, con cuya ayuda se clasificarán los objetos estudiados.

Después de la formación de los conceptos cualitativos y la división de todos los objetos en conjunto, se puede dar un paso más adelante, el cual consiste en establecer determinadas relaciones entre los conjuntos de objetos semejantes, con el auxilio de conceptos comparativos. En este caso, los conceptos comparativos ordenan todos los objetos de la región investigada en una determinada secuencia, en la cual cada objeto ocupa un determinado lugar. Por ejemplo, con ayuda de los conceptos "más pesado", "más ligero", "de igual peso", se pueden distribuir todos los objetos en una secuencia de conjuntos. A su vez un par de conceptos comparativos puede servir de base para la introducción de conceptos cuantitativos, es decir, conceptos que designan la cualidad medida. Esto es: el par "más pesado"- "más ligero" nos lleva al concepto cuantitativo de peso. Sin embargo, el tránsito de los conceptos cualitativos y comparativos a los cuantitativos se realiza sólo con ayuda de muchas proposiciones teóricas. De acuerdo a lo anterior podemos ver que entre los objetos que se utilizan o se estudian, se pueden clasificar principalmente en caracteres cuantitativos –medibles-y cualitativos –no medibles– en donde los primeros reflejan una propiedad real con una mayor exactitud y los segundos se refieren al objeto idealizado de la teoría y sólo por eso refleja el objeto real de la teoría; o sea, sólo caracterizan los objetos reales.

Si se realiza un análisis cuidadoso del uso de los números en la vida diaria, se llegaría a la conclusión de que la mayoría de los números que se emplean, no poseen las propiedades aritméticas que ordinariamente se les atribuye; esto es, no tiene sentido sumarlos, restarlos, multiplicarlos o dividirlos. De aquí que: una primera clasificación de los conceptos es la clasificación en cualitativos y cuantitativos.

Al resultado de observar un carácter cualitativo se le denomina Atributo o Variable Cualitativa. Los elementos o individuos varían cualitativamente con respecto a una determinada característica y se pueden clasificar en categorías o modalidades. Los atributos pueden ser dicotómicos – caracterizados por la presencia o ausencia de una propiedad – o politómicos. Para esta clasificación es necesario que las categorías cumplan con:

1. Estar bien definidas.
2. Ser mutuamente excluyentes, ya que ningún elemento puede pertenecer a la vez a dos modalidades.
3. Ser exhaustivas, para que todo elemento pueda incluirse en alguna de las modalidades.

Cuando el carácter es cuantitativo, el resultado de su observación recibe el nombre de variable y las medidas de la misma, valores. Entre los caracteres susceptibles de medición se pueden establecer diferencias de tipo cuantitativo y se distinguen en:

1. Variables cuantitativas discretas. Son aquellas variables cuya medición sólo puede expresarse en números enteros, pero entre dos valores consecutivos no se puede dar ningún otro intermedio. Por ejemplo, las enumeraciones o conteos.
2. Variables cuantitativas continuas. Es una variable que por naturaleza puede adoptar un número finito de valores distintos. Para todo par de valores siempre es posible determinar otro intermedio, estando sólo limitada por la precisión del instrumento de medida.

En la práctica a veces se puede elegir entre medir un carácter de forma cuantitativa o cualitativa. Pero es más aconsejable optar por la medida cuantitativa, ya que ésta siempre da más información. No obstante, dependiendo del tipo de variable, es posible asociar a su vez una escala de medición. Por lo tanto para poder realizar correctamente esta asociación uno tiene que revisar cuáles son las propiedades que debe cumplir una medida.

2.4. Propiedades de la Medida

El análisis estadístico se basa, obviamente, en datos. Pero un dato no es otra cosa que un número. Lo cual significa que, para poder analizar datos, es necesario asociar números a las características que se desea estudiar. Una asociación numérica, en el campo de las matemáticas abstractas conocido como teoría de la medida, es una medida si y solo si cumple con ciertas propiedades matemáticas:

a) La medida del conjunto debe ser igual a la suma de las medidas de todas sus partes.

b) La medida de nada o ninguno, debe ser 0.

c) La medida de una parte de algo no debe ser mayor que la medida del todo.

d) Si la medida se hace de cierto modo bajo determinadas condiciones físicas, se debe obtener resultados iguales.

La teoría de la medida tiene por objeto el estudio de los diferentes modelos que permiten establecer las reglas, que se necesitan seguir, para una correcta asignación numérica. Estas propiedades matemáticas son también denominadas "reglas de medición". El resultado de la medición puede expresarse de la siguiente manera:

$$G = gU$$

donde

- G es la propiedad de medida
- U la unidad de medición
- g el valor numérico de la magnitud correspondiente

Esta ecuación se conoce como "la ecuación fundamental de la medición". Aunque para atribuir un determinado valor numérico a la propiedad medida de acuerdo con esta ecuación, se tiene que cumplir con las siguientes normas:

a) El término técnico "Aditividad finita" o "Regla de adición" expresa que: el valor numérico de la suma de dos valores físicos de las propiedades, deberá ser igual a la suma de los valores físicos de esta propiedad.

$$gU(G_1UG_2) = g_1U + g_2U$$

En la formulación de esta regla, entre G_1 y G_2 situamos el símbolo "o" que indica la operación empírica de la unión de dos grados en una misma propiedad. Es natural que esta operación pueda diferenciarse de una suma aritmética. La operación de unión de dos grados diferentes de una propiedad no siempre está sometida a la regla indicada.

En otras palabras se dice que: *la medida del conjunto debe ser igual a la suma de las medidas de todas sus partes.* Esto es: si se tuviese que pesar doce pedazos de madera, encontraríamos que el peso es el mismo si pesáramos cada trozo de madera por separado y posteriormente sumáramos todos los pesos, que si las pesáramos todos los trozos juntos.

b) Regla de equivalencia: si el valor físico de las propiedades medidas es igual, iguales deberán ser sus expresiones numéricas.

$$G_1 = G_2 \Rightarrow g_1 = g_2$$

Se podrá decir que esta propiedad realmente está implicada por la propiedad a), un caso particular es que la medida de "nada" o ninguno, debe ser 0. El concepto puede que parezca menos ridículo si pensamos en la medida como en un conteo. El número de elementos en un conjunto es una clase de medida del conjunto.

c) La propiedad de monotonía: el valor físico de la propiedad de un cuerpo es menor que el valor físico de esta misma propiedad en otro cuerpo, entonces el valor numérico del primero deberá ser menor que el del segundo.

$$G_1 < G_2 \Rightarrow g_1 < g_2$$

Expresando lo anterior de otra manera se tiene que la medida de una parte del algo no debe ser mayor que la medida del todo. Por ejemplo, el peso de medio paquete de mantequilla es menor que el de todo el paquete. Teniendo en cuenta la propiedad a) y b), la propiedad c) es equivalente a afirmar que las medidas están expresadas por números no negativos (es decir el conjunto de los reales positivos).

d) La regla de la unidad de medida dice que: *si se llega a seleccionar un determinado cuerpo o un proceso natural fácilmente reproducible y caracterizar la unidad de medida por medio de este cuerpo o proceso este deberá conservar inmutables su medida, forma, periodicidad, etc.* Pero los cuerpos y procesos reales están sometidos a modificaciones debido a la influencia de las condiciones circundantes. Por ello, como patrones reales se toman aquellos cuerpos y procesos que son más estables respecto a las condiciones externas. Más aún si en un experimento de medición, la medida se hace bajo determinadas condiciones físicas prescritas, al repetir el experimento bajo las mismas condiciones se debe obtener resultados iguales.

2.5. Escalas de Medida

Cuando se usa una regla para asignar números, a las propiedades de los objetos o de hechos, se crea una escala. Las escalas son posibles, en primer lugar, sólo porque existe un *isomorfismo*[2] entre las propiedades de la serie numeral y las operaciones empíricas que se pueden realizar con las propiedades de los objetos.

[2]El concepto matemático de isomorfismo pretende captar la idea de tener la misma forma, la misma estructura. Una aplicación $f : X \to Y$ entre dos conjuntos dotados del mismo tipo de estructura es un isomorfismo cuando cada elemento de Y proviene de un único elemento de X y f transforma las operaciones, relaciones, etc. que hay en X en las que hay en Y. Cuando entre dos estructuras hay un isomorfismo, ambas son indistinguibles, tienen las mismas propiedades, y cualquier enunciado es simultáneamente cierto o falso.

El tipo de escala obtenido cuando se envía a los numerales para servir como representantes de un estado de cosas en la naturaleza depende del carácter de las operaciones empíricas básicas realizadas en la naturaleza. Estas operaciones están normalmente limitadas por las peculiaridades de la cosa que se desea medir con una escala, y por una elección de procedimientos concretos; pero, una vez seleccionados, los procedimientos determinan que habrá uno u otro de cuatro tipos de escala: *nominal, ordinal, de intervalo, o de razón.*

2.6. La Naturaleza de la Escala

La escala de medición de una variable matemática y/o estadística es una clasificación que fue propuesta para describir la naturaleza de la información contenida en los números asignados a objetos y, por lo tanto, a las variables. Los niveles de clasificación fueron propuestos por Stanley Smith Stevens en el artículo: *On the theory of scales of measurement*, escrito en 1949. Diferentes operaciones matemáticas sobre las variables son posibles, dependiendo del nivel al cual una variable es medida. De acuerdo a este esquema de clasificación, en estadística los tipos de estadísticas y pruebas de significancia que son apropiados dependen de la escalas de medida de las variables involucradas. Son cuatros escalas de medida las que fueron propuestas por S.S. Stevens.

1. Identidad: los números pueden servir como etiquetas para designar o identificar, objetos o clases (números nominales).los números pueden servir como etiquetas para designar o identificar, objetos o clases (números nominales).
2. Orden: los números pueden servir para representar la posición de una serie o para indicar el rango ordenado de los objetos (números ordinales).
3. Intervalos: los números pueden servir para indicar las diferencias entre los objetos (números reales).
4. Razón: los números pueden servir para indicar las proporciones entre los objetos (números reales).

2.7. Escala Nominal

La medición de una escala nominal simplemente consiste en situar a cada individuo en una u otra categoría dada o el asignarles un nombre. En este caso esta escala de medición trata de agrupar objetos en clases, de modo que todos los que pertenezcan a la misma sean equivalentes respecto del atributo o propiedad en estudio. El hecho de que a veces, en lugar de denominaciones se les atribuyan números, puede ser una de las razones por las cuales se les conoce como “medidas nominales”.

Los números asignados a las escalas nominales tienen iguales propiedades que los demás, pero en ningún momento podemos pensar en el manejo de orden, tamaño y otras propiedades de las cifras, puesto que no se deben tomar como tales. Es decir estos números carecen de propiedades cuantitativas y sirven únicamente para identificar las clases. Naturalmente los datos empleados con las escalas nominales consisten en conteos de frecuencias o tabulaciones del número de sucesos en cada clase de la variable

estudiada. Tales datos reciben indistintamente los nombres de: datos de frecuencias, datos enumerativos, datos de atributos o datos de categorías. Las únicas relaciones matemáticas adecuadas a las escalas nominales son las de equivalencia = o ≠ no equivalencia. De tal manera que una persona u objeto particular tiene la característica que define la clase o no la tiene. Este tipo de medición constituye el nivel de medición mas bajo.

Esto implica que un requisito esencial en la medida nominal es que los objetos sean caracterizados en categorías mutuamente exclusivas y exhaustivas, es decir cada objeto es asignado a una y sólo una categoría, y la diferencia entre las categorías es de tipo, más que de grado.

Las variables que son únicamente nominales son también llamadas variables categóricas. En investigación social las variables medidas en un nivel nominal incluyen: género, raza, afiliación religiosa, afiliación partidista, nivel escolar y lugar de nacimiento.

2.8. Escala Ordinal

Para los casos en que se pueden detectar diversos grados de un atributo o propiedad de un objeto, la escala ordinal es la indicada, puesto que la medición de una escala ordinal supone situar a los individuos en un orden de acuerdo con algún criterio. Por lo que los datos ordinales constituyen un escalón superior en relación con los datos nominales, porque permiten decir si un individuo está antes o después que otro en una escala. De tal manera que no solo se sabe que los datos son diferentes entre sí -característica que define a las escalas nominales- sino que se mantiene alguna clase de relación entre ellos. Es de señalarse que los números pueden asumir el lugar de los objetos en estudio, puesto que los números son representaciones parciales de éstos. Lo que nos lleva a plantear que los números pueden tratarse como si fueran diferentes; es decir, se pueden ordenar.

A pesar de que no existe ley alguna que prohiba sumar, restar, multiplicar, etc., números asignados según escalas ordinales, el resultado de tales operaciones puede no indicar nada respecto del grado de atributo en cuestión que el objeto en estudio posee. Los resultados de estos cálculos aritméticos no pueden informar absolutamente nada respecto del atributo real inherente al objeto, es decir los numerales empleados en conexión con las escalas ordinales no son cuantitativos. Mas aun, indican solamente la posición en una serie ordenada y no "cuánta" diferencia existe entre posiciones sucesivas en la escala. En este caso las operaciones que se le están permitidas o que tiene algún significado para éste tipo de escala es la relación que se expresa en términos del álgebra de las desigualdades. Es decir: a es menor que b ($a < b$) o a es mayor que b ($a > b$).

Ejemplos incluyen la escala de dureza de los minerales de Friedrich Mohs, los resultados de caballos de carrera, el cual únicamente dice cual arribo primero, segundo, tercero, etc. pero no los intervalos de tiempo; y la mayoría de las medidas en psicología y otras ciencias sociales, por ejemplo actitudes como son preferencia, prejuicio y clase social.

2.9. Escala Intervalar

Cuando no solamente es posible distinguir la diferencia entre los diversos grados de propiedad de un objeto (característica de la medida ordinal)

sino que también pueden discernirse las diferencias entre objetos iguales, se recurre a la medida de intervalo. En este caso, una unidad de medida se define en términos de algún parámetro (grado, pulgada, pie, etc.). Es decir, la medición en una escala de intervalo consiste en asignar un número a un individuo para indicar su posición exacta a lo largo de una escala continua. Los datos de intervalo ocupan otro escalón superior en la jerarquía de escalas de medición; nos permiten decir qué distancia separa a un individuo de otro dentro de una escala.

Las medidas de intervalo implican la asignación de números de modo tal, que a iguales diferencias entre los grados del atributo estudiado en un objeto, correspondan iguales diferencias entre los números. Una de las características distintivas de la medida de intervalos es que el cero no necesariamente implica que el objeto carece del atributo, puesto que en una escala de intervalo, el punto cero es arbitrario.

Los valores numéricos asociados con estas escalas son efectivamente cuantitativos y por lo tanto, permiten el uso de operaciones aritméticas, tales como suma, resta, multiplicación y división, además que poseen la propiedad de distintividad y orden. Dado que en este caso, la diferencia entre los números sí es significativa.

Ejemplos de medida intervalar son la fecha del año, temperatura en escala Celsius o Fahrenheit. Una medida intervalar usada comúnmente en una investigación científica social son las mediciones construidas como lo es el coeficiente de inteligencia estandarizado (IQ).

2.10. Escala de Razón

La medida de razón o cociente se diferencia de la de intervalo únicamente en que el punto cero no es arbitrario, sino un valor absoluto y corresponde realmente a una total ausencia de la propiedad estudiada. Por ejemplo, si tenemos una longitud igual a cero significa que no hay longitud. Por lo que cuando se observa una carencia total de propiedad, se dispone de una unidad de medida para tal efecto. Más aun, a iguales diferencias entre los números asignados, corresponden iguales diferencias en el grado de atributo presente en el objeto de estudio. Por todo lo anterior podemos decir que las mediciones en una escala de razón tienen todas las características de las mediciones de intervalo, pero con el rasgo adicional de que la razón de dos valores cualesquiera, es independiente de la unidad de medición (por ejemplo, 4 metros es a 2 metros como 2 metros es a 1 metro).

Al analizar la escala de intervalo y de razón se asume que estos dos tipos de escalas se basan en números reales y este tipo de números nos indica el nivel más alto de las mediciones científicas. Los valores numéricos asociados con estas escalas son efectivamente cuantitativos y por tanto, permiten el uso de operaciones aritméticas tales como suma, resta, multiplicación y división. Por lo que, diferencias iguales entre puntos de cualquier parte de la escala son iguales entre si.

Una de las características de las escalas de orden superior es que se les puede transformar fácilmente en escalas de orden más bajo. Así, el resultado de una carrera de una milla se puede expresar en unidades de tiempo (escala de razones). Los tiempos se pueden transformar en datos de una escala ordinal, por ejemplo primero, segundo y tercer lugar en el orden de llegada. Sin embargo, no es posible efectuar la transformación inversa. Si

por ejemplo, conocemos únicamente el orden de llegada en una carrera, no podemos expresar los resultados en términos de una escala de razones (tiempo). Aunque es admisible transformar las marcas de una escala de mayor nivel a otra de menor nivel, no es recomendable hacerlo, por lo general, ya que se pierde información cuantitativa.

Las variables sociales de razón incluyen edad, tiempo de residencia en un lugar determinado, número de organizaciones, el número de asistentes a una iglesia en un periodo de tiempo, entre otros.

2.11. Clasificación de las variables por los valores que pueden adoptar

Puesto que existen confusiones en el uso y la interpretación de las reglas de clasificación que propuso Stevens, existe una clasificación alternativa. Ésta se usa de manera frecuente, porque solo divide las variables en dos grupos y crea menos dificultades al clasificar las variables. Por esto, investigadores de diferentes disciplinas, aún aquellos que no cuentan con una formación matemática, pueden hacer un uso adecuado de técnicas estadísticas multivariadas sin problema alguno. La clasificación es la siguiente fusionada con aquella de S.S. Stevens.

1. Variables no métricas o cualitativas (escala nominal u ordinal).
2. Variables métricas o cuantitativas (de intervalo o de razón).

Esta clasificación aparece frecuentemente en los libros, como ya se había mencionado, por lo que es importante conocerla y entenderla. No obstante, no hay que olvidar que la clasificación de S. S. Stevens es la que cuenta con más apoyo entre los estadísticos. En la presentación de técnicas Estadística Multivariadas (Ver Capítulo 3) se utilizará la clasificación propuesta por Stevens. Sin embargo, se anexará esta última, para que sea más fácil el manejo de las técnicas estadísticas multivariadas a cualquier forma a la que esté acostumbrado el lector.

2.12. Comentarios

En este capítulo, se aludió a la definición de medida dada por Norman Campbell y se consideró a la medida el punto de partida para el establecimiento de cualquier taxonomía. Se indicó que: no únicamente existe una taxonomía que está basada en los atributos de las variables, sino que también existe una taxonomía basada en los procedimientos matemáticos permisibles entre variables. Si bien la clasificación sugerida por S.S. Stevens no es única, sí es la más apropiada, para el manejo de variables. Aunque ambas clasificaciones, la de Stevens y la basada en atributos, son usadas por igual. Por lo que además se presenta una tercera taxonomía que conjunta las dos taxonomías, la que está basada en los atributos y la está basada en los procedimientos matemáticos permisibles. En resumen la escala de medida es sin duda una herramienta necesaria e inexcluible en cualquier disciplina científica, en particular en la estadística.

Escala	**Descripción**	**Relaciones Definidas**	**Estructura Matemáica de Grupo**	**Estadísticas Permisibles**	**Ejemplos Típicos**	**Pruebas Estadísticas Apropiadas**
Nominal	Usa números para identificar objetos, individuos, eventos o grupos.	Equivalencia o pertenencia a una categoría	Grupo de Permutación $x' = f(x)$. $f(x)$ significa caulquier sustitución biunívoca	Frecuencia, Moda, Coeficiente de Contingencia.	Enumeración de jugadores de de fútbol. Sexo: Masculino, Femenino.	Pruebas estadísticas no paramétricas.
Ordinal	Además para identificación, los números proveen infformación acerca de la catidad relativa de algunas característcas representadas por un evento, objeto, etc.	Equivalencia, determinación de mayor o menor.	Grupo isotónico $x' = f(x)$, donde $f()x$ significa cualquier función monótona creciente	Mediana, Percentiles, Sperman r, Kendall r, Kendall W.	Pruebas de inteligencia, calificaciones como NA, S, B, MB	Prueba estadísticas no paramétricas
Intervalar	Posee todas las propiedades de las escalas nominales y ordinales más los intervalos consecutivos entre puntos que son iguales.	Equivalencia, determinación de mayor o menor, proporción conocida de un intervalo a cualquier otro.	Grupo lineal general $x' = ax + b$	Media, Desviación estándar, Correlación del momento producto de Person, Correlación del múltiple momento producto.	Temperatura, posición en una línea, pruebas de inteligencia, resultados estandarizados.	Pruebas estadísticas paramétricas.
Razón	Incorpora todas las propiedades nominal, ordinal así como la escala intervalar, más la inclusión del punto cero absoluto.	Equivalencia, determinación de mayor o menor, proporción conocida de un intervalo a cualquier otro.Proporción conocida de un valor de la escala a cualquier otro.	Grupo de similitud $x' = ax$	Media geométrica, Media armónica, Coeficiente de Variación.	Longitud, enumeración, densidad, intervalos de tiempo, etc.	Pruebas estadísticas paramétricas

Cuadro 2.1: Escalas de Medida (Adaptación de S. S. Stevens)

Capítulo 3

Aspectos del Análisis Multivariado

La investigación científica se ha explicado como un proceso de aprendizaje iterativo. Los objetivos pertinentes a la explicación de un fenómeno físico o social tienen que ser especificados y después probados mediante la recolección y el análisis de datos. Como resultado, el análisis de los datos recolectados mediante experimentación u observación sugerirán una explicación modificada del fenómeno. A través de este proceso iterativo de aprendizaje, variables son descartadas o agregadas al estudio. Como las explicaciones de los fenómenos reales suelen ser complejas, hace falta un número muy grande de variables. En este capítulo se presentarán métodos estadísticos diseñados para extraer información de este tipo de conjuntos de datos. Debido a que los datos incluyen medidas simultáneas sobre muchas variables, este conjunto de metodologías se le llama Análisis Multivariado.

El Análisis Multivariado reúne un gran número de metodologías y sólo se expondrán las más usadas. La exposición de éstas será únicamente un esbozo, puesto que se intenta introducir del uso de estas técnicas únicamente. Primero se darán ejemplos del uso de éstas mediante una síntesis de publicaciones actuales, que fueron escritas por científicos de diferentes disciplinas. Luego se presentaran los métodos de forma individual, señalando su historia y su uso. También se dará una clasificación de acuerdo a los propósitos de la técnica. Posteriormente se resumirán éstas señalando: el objetivo, la técnica multivariada, el tipo de variable que se usa, así como los usos y propósitos de la técnica.

3.1. Análisis Multivariado

Los métodos estadísticos pueden ser de gran ayuda cuando se quiere asimilar u organizar resultados numéricos, pero es esencial saber que es lo que esta pasando. Cuando los datos provienen de medidas simultáneas sobre muchas variables se tiene un conjunto de datos multivaridados. Este conjunto de datos multivariados aunque complejo, se tiene que analizar en conjunto y no en forma separada usando métodos univaridados. Por lo que, se requiere el uso de técnicas estadísticas avanzadas para poder extraer información relevante y/o necesaria para dar una interpretación adecuada a los datos. Los métodos o técnicas que permiten analizar simultáneamente conjuntos amplios de variables son los métodos multivariados. Y la parte de la estadística que alberga las técnicas para analizar simultáneamente variables múltiples es el Análisis Multivariado.

Everitt y Dunn (2000) [13] mencionan que existen diversas razones para usar

técnicas estadísticas multivariadas. Estas razones son: el tratar de encontrar una estructura o patrón en los datos, el deseo de encontrar similaridades en un conjunto de variables, la posibilidad de poder o no discriminar los grupos definidos, así como la exploración de patrones y la inferencia sobre estos.

1. En un conjunto de datos multivariados encontrar una estructura o patrón en los datos enriquecerá la interpretación que se piensa dar a las observaciones y también puede ayuda a simplificar la interpretación. Estos patrones pueden reflejar perfiles multivariados o similaridad. El origen de este patrón puede obedecer a que las medidas provienen de grupos similares de objetos, pero es muy posible que no se sepa cuales son estos grupos, cuantos son o bien cuales son los objetos que pertenecen a un grupo especifico. Por lo que se usan estos datos para explorar estas posibilidades.

2. En un conjunto de variables pueden existir similaridades, altamente correlacionadas, que sugieran ser la misma variable. En este caso se pueden usar técnicas para reducción de datos. Si se pude determinar cuáles de las variables son indicadores de cuales variables, antes de hacer cualquier análisis, se podría probar si los datos son o no consistentes con un modelo de medidas.

3. Si se pueden definir grupos se puede conocer cómo los perfiles multivaridados pueden discriminarse entre ellos.

4. Posiblemente una de las más difíciles e interesantes áreas, es la exploración de patrones de asociación entre conjuntos de medidas multivariadas para inferir patrones causales.

En lo que se refiere a la Estadística Multivariada, se puede entender como una extensión de la estadística Univariada, aunque la primera estará relacionada con una distribución conjunta de varias variables. Sir Maurice Kendall dice: "Formally, then, we may define multivariate analysis as the branch of statistics which is concerned with the relationships among sets of dependent variables and the individuals which bear them[1]". [25]

Hay que tener en cuenta que un análisis estadístico encierra un conjunto de métodos y/o técnicas univariantes o multivariantes, que permiten estudiar y tratar en bloque una o varias variables. Estas variables puede ser: de orden, nominal, de intervalo o de razón. Cualquiera que sea el tipo de variable, invariablemente se involucra un método estadístico multidimensional complejo y una diversidad de enfoques teóricos y prácticos de un estudio multidimensional. Claramente esto nos lleva a procesos matemáticos elaborados, que por fuerza, han de apoyarse en el cálculo matricial y en técnicas de complejidad inherente. Es ésta la razón por la cual, hasta época muy reciente, se ha comenzado a difundir suficientemente su aplicación. De tal manera que la comunidad científica se ha beneficiado del empleo de estas técnicas avanzadas, con la ayuda de ordenadores y una amplia gama de paquetes estadísticos.

3.2. Aplicaciones del Análisis Multivariado

Los investigadores usan técnicas estadísticas avanzadas para examinar las relaciones entre variables múltiples, tales como dieta, ejercicio realizado y problemas cardiovasculares o bien para predecir información como son futuras tasas

[1] Formalmente, se define al Análisis Multivariado como la rama de la estadística que está interesada en la relación latente entre conjuntos de variables dependientes e individuos. (Trad. Aut)

de interés o desempleo. Por lo que un gran numero de personas, desde las ciencias sociales a las agencias gubernamentales y las agencias de negocios profesionales, depende de los resultados de los modelos multivariados para la toma de decisiones.

Las aplicaciones de técnicas estadísticas tienen un rango considerablemente amplio, que va en aumento cada día, por lo que conocerlo proporciona una fuente de información valiosa. Una muestra de aplicaciones reales servirá, naturalmente, para sugerir nuevas aplicaciones. Por lo tanto: se expondrán algunas aplicaciones de técnicas del Análisis Multivariado, ligadas ciertas áreas del conocimiento. Estas aplicaciones responden a las sugeridas por Sir Maurice Kendall (1980). Particularmente esta muestra de aplicaciones es una síntesis de trabajos realizados por investigadores de diferentes disciplinas, que encontraron en las técnicas del Análisis Multivariado herramientas matemáticas que apoyaran sus teorías o sus estudios.

1. Antropología.**Instrumentación de Colecciones Craniofaciales Multicéntricas**: Investigadores de Inglaterra estudiaron 200 cráneos humanos para definir la variabilidad de la abertura del piriform (entrada de la cavidad nasal). Trataron de demostrar que esta tarea puede realizarse por medio de redes neuronales. Se registraron tanto datos métricos como nométricos, se realizó un análisis descriptivo así como un *Análisis de Conglomerados* y un *Análisis de Discriminantes*. Estos análisis fueron comparados con los métodos convencionales, una red neuronal del tipo Kohonen (15 x 15). La clasificación resultante no contradijo a las técnicas multivariadas. Por lo que, la red neuronal puede considerarse un método conveniente para la investigación de muestras grandes de material biológico. También puede ser de utilidad en anatomía y antropología así como en la identificación del material desconocido (Prescher et al., 2005) [36] .

2. Arqueología.**Orígenes del Viejo Mundo y de los Primeros Habitantes Humanos del Nuevo Mundo**: Una visión Craneofacial Comparativa: Datos craneofaciales humanos fueron usados para estimar la similaridades y diferencias entre muestras recientes y prehistóricas del Viejo Mundo. Los datos fueron analizados usando la técnica de *Análisis de Conglomerados*, asistido por el análisis conocido como bootstrap y por *Análisis de Discriminante*. Estos análisis mostraron asociación entre los primeros pobladores del Nuevo Mundo, que poblaron la frontera Canadá - E.U., y los Ainu, gente nativa del archipiélago Japonés. También se concluyó que la ruta de entrada al Nuevo Mundo fue por la frontera noroeste. Además, se encontraron similaridades entre los Inuit, los Aleut y los Na-Dene, quienes se internaron hasta el suroeste de América, y la población del continente del Este de Asia. Por último, aunque ambos arribaron al Nuevo Mundo, en el Nuevo Mundo se muestra una mezcla de rasgos característicos de la frontera norte del Viejo Mundo y del centro del continente asiático, china central, la proporción del ultimo es mayor por ser la mas reciente (Brace et al. 2001) [4].

3. Economía. **La respuesta de la industria automotriz y de los consumidores a los cambios de los estándares económicos de la emisión y la extracción de combustible (1975-2003)**: Una revisión histórica de cambios en tecnología, precios y ventas de varias clases de vehículos: Se realizó un estudio para determinar la respuesta, tanto de la industria automotriz como la de los consumidores, a los cambios de los estándares económicos de la emisor y la extracción de combustible, que ocurrieron en los Estados Unidos y California en un periodo de casi treinta años

(1975-2003). Establecer una conexión lógica de estas respuestas al desarrollo de tecnología y cambios de factores económicos, tales como precios de vehículos, ingresos del consumidor, inflación y precio del combustible dentro de periodos iguales de tiempo. Así como correlacionar las ventas de vehículos de diferentes marcas a las características del vehículo y a los factores macroeconómicos utilizando *Análisis de Regresión* Múltiple. Los datos provinieron de varios modelos y tamaños de vehículos. El estudio indicó que los cambios en la emisión y las regulaciones económicas del combustible forzaron a la industria a desarrollar una secuencia impresionante de nuevas y mejoradas tecnologías que fueron introduciendo rápidamente en carros de pasajeros, vans, furgonetas y camionetas ligeras. Algunos de los resultados han sido vehículos de poca potencia, con emisiones ultra limpias y mejoras en rendimiento de combustible de 60-70 % comparado a los modelos 1975. Los precios de los modelos en diversas clases de vehículos se incrementaron por un factor del 1.5 al 2.0 basado en el índice del precio al consumidor. El aumento en precios del vehículo se acomodó de acuerdo al aumento en los salarios así como la creatividad en las ventas por financiamiento a través de períodos largos de pagos (Burke et al. 2004) [6].

4. Educación. **Comprensión Conceptual vs. Resolución de Problemas con Algoritmos: Un Análisis de Componentes Principales de una Evaluación Nacional**: Se analizaron los resultados de un examen nacional desde la perspectiva de aprendizaje conceptual vs. resolución de problemas con algoritmos. Se demostró que el *Análisis de Componentes Principales* puede servir como herramienta de escrutinio para aprobar papers sobre la educación de la química. Más aun, la evaluación a gran escala nacional, proveyó datos confiables, apropiados para este tipo de análisis. Los datos fueron estudiados por la evaluación nacional griega, para una muestra de 647 estudiantes de nivel medio superior (17 años aproximadamente) que se orientaron a las ciencias, ingeniería o medicina. El análisis de componentes principales condujo a la extracción de tres factores: un factor se centro en la memoria y el simple uso de preguntas de conocimiento; un segundo factor separo las preguntas conceptuales, el tercero incluyo todas las preguntas sobre computo. Las conclusiones fueron probadas por la *MANOVA* (Stamovlais et al. 2004) [45].

5. Física. **Análisis de Correlación Canónica No-lineal de la Variabilidad del Clima Pacifico Tropical usando una Aproximación de Redes Neuronales**: Los avances recientes en la modelación de redes neuronales ha conducido a la generalización de técnicas clásicas de análisis multivariado como son: *Análisis de Componentes Principales* y *Análisis de Correlación Canónica*. El método análisis de correlación canónica no-lineal es usado para estudiar las relaciones entre nivel de presión marina del pacifico tropical y la temperatura superficial del nivel del mar. La primera tendencia extraída es una oscilación proveniente del sur, del Niño, no-lineal que muestra una asimetría entre los períodos calurosos de El Niño y los períodos fríos de La Niña. La primera tendencia no-lineal del análisis de correlación canónica no-lineal se encontró un aumento gradual con el tiempo. Durante los años 1950-75, el nivel de presión marina del pacifico tropical, mostró no-linealidad, mientras que la temperatura superficial del mar reveló una débil no-linealidad. En el período 1976-99, el nivel de presión marina del Pacífico tropical mostró una débil no-linealidad, mientras que la débil no-linealidad en la temperatura superficial del nivel del mar se incremento. Una segunda tendencia del análisis de correlación canónica

no-lineal mostró fluctuaciones de tiempo mayores, nuevamente con débiles, pero notables, no-linealidad en el nivel de presión del marina del Pacífico tropical pero no en la temperatura superficial del mar (Hsieh, 2001) [19].

6. Medicina. **Análisis de Correspondencias de Genes y Tipos de Tejidos para encontrar Conexiones Genéticas provenientes de Micro-arreglos**: Se uso Análisis de Correspondencia y Análisis de Regresión Múltiple para analizar datos provenientes de micro-arreglos genéticos. Se utilizó el *Análisis de Correspondencias* para encontrar relaciones entre genes y tejidos en una gráfica bidimensional, respetando las distancias entre ambos. Para inferir sobre conexiones genéticas, se usaron correlaciones parciales. Al parecer ambas aproximaciones fueron una manera más natural de analizar expresiones de genes que el popular uso de conglomerados (Kishino y Waddell, 2000) [26].

7. Silvicultura. *Análisis Multivariante del clima a lo largo de la costa meridional de Alaska*: Diferentes técnicas multivariadas -incluyendo *Análisis de Componentes Principales, Análisis de Conglomerados y Análisis de Discrimínate*- fueron usadas para trazar 10 diferentes grupos significativos de climas a lo largo de la costa sureña de Alaska basadas en latitud, longitud, temperaturas de estación, así como precipitaciones, periodos libres de heladas y el numero total de días de temperaturas elevadas (Farr y Hard, 1987) [15].

8. Sociología.**Convergencia o Resiliencia? Un Análisis de Conglomerados de Regímenes de Estado de Bienestar (Welfare) en Países Avanzados**: Se estudió un conjunto de indicadores sociales cuantitativos usando *Análisis de Conglomerados*, con el fin de identificar regímenes políticos, de los cuales se derivan órdenes entre el mercado, el estado y las familias en la producción y distribución de recursos requeridos para el bienestar de la gente. Este análisis empírico reveló la existencia de los tres regímenes, originalmente identificados por Esping-Andersen, -el social-demócrata, liberal y conservador- al cual uno se tiene que agregar, como muchos autores han señalando, el régimen latino. Los datos revelan una fuerte y duradera relación de mutua causalidad entre configuraciones de programas sociales en las diversas sociedades bajo el análisis, las situaciones sociales que resultan de los programas sociales y finalmente, el nivel de participación cívica encabezada (o no) por la gente en movilizaciones colectivas que dan forma a programas sociales. El análisis permite identificar el lugar que ocupa Canadá en términos de programas de estado de bienestar (Walfer) capitalista (Saint-Arnaud y Bernard, 2003) [40].

9. Arte.**Una aplicación del Análisis de Componentes Principales al estudio de esculturas hindúes del Sur**: Un investigador utilizó el *Análisis de Componentes Principales* que aplicado a un estudio de esculturas del siglo XVIII, puede contribuir a la solución de uno de los mayores problemas del la historia del arte, como el de datar las esculturas de un templo. En este estudio de iconometría se midieron las distancias, punto por punto, de 40 esculturas del templo Kailasanatha. Tal estudio de medidas iconométricas será útil en la datación de esculturas, en trabajos de restauración, en la averiguación de la existencia o no de algún canon iconométrico seguido por el escultor, así como para hipotetizar sobre el número de escultores que realizaron el trabajo. En este enfoque, los historiadores del arte no dependen de medidas para comparar los diferentes estilos escultóricos, en su lugar dependen de gran manera de impresiones subjetivas visuales. El método objetivo de estudiar las similaridades entre

esculturas Hindúes revelara características especiales que han sido olvidadas por investigadores anteriores. En la datación de esculturas, los valores promedios de las características iconométricas de diferentes periodos y regiones serán útiles. Usando la técnica de Análisis de Conglomerados se pude determinar si una escultura es cercanamente similar a un conjunto de proporciones canónicas (Siromoney, Gift, Bagavandas, M. y Govindaraju, S. 1980) [42].

Los ejemplos anteriores, tomados más o menos aleatoriamente, ilustran el extenso campo actual de aplicaciones de los métodos multivariados. Las aplicaciones de los métodos del Análisis Multivariado encuentran en casi todas las disciplinas una aplicación, que puede ser extendida a cualquier dirección, sin olvidar el objetivo del análisis. Por lo que a continuación se presentará una primera clasificación del Análisis Multivariado que obedece a su objetivo.

3.3. Objetivo del Análisis Multivariado

Los ejemplos anteriores sugieren usos diferentes de estas técnicas multivariadas, que nos pueden indicar cuándo se puede emplear una u otra técnica. Se puede decir que: en particular la selección del método más apropiado depende del *tipo de datos*, el *tipo de problema* y el *tipo de objetivos* que están considerados en el análisis. Este último, el tipo de objetivo, establece un primer esquema de clasificación. Clasificación que nos permitirá, una vez delineado el objetivo de investigación, escoger la técnica apropiada para el análisis.

1. **Reducción de datos o simplificación estructural**. Se intenta simplificar la estructura del fenómeno estudiado, proporcionando una estructura lo mas simple posible, que permita una fácil interpretación. Alguna de estas técnicas son: análisis de componentes principales, análisis factorial, análisis de correspondencias, escalamiento multidimensional, análisis de correlación canónico, análisis conjunto, etc.

2. **Clasificación**. Se permite crear grupos de objetos o de variables similares entre sí, a partir de sus características medidas. Los esfuerzos van dirigidos a la obtención de tipologías. Alternativamente se busca analizar las relaciones entre variables para ver si se pueden separar los individuos en agrupaciones a posteriori. En este caso, la información conocida de los objetos sobre múltiples variables, se utiliza para asignar a los sujetos al grupo al que más se parecen. Algunas de las técnicas de clasificación son: análisis de conglomerados, análisis discriminante, etc.

3. **Relación entre dos conjuntos de variables**. Se busca la existencia de dependencia de un conjunto de variables en términos de otras, analizar su relación o incluso su predicción. Algunas de estas técnicas son: análisis de regresión múltiple, regresión logística, análisis de series temporales, etc.

4. **Pruebas de hipótesis**. Muchas de las técnicas del análisis multivariado tienen un carácter esencialmente descriptivo, pero con otras pueden ponerse a prueba hipótesis sobre modelos complejos basados en poblaciones multivariadas. Algunas de estas técnicas son: Manova o análisis multivariado de la varianza y Mancova o análisis multivariado de la covarianza.

Se maneja ésta como la primera clasificación de las técnicas multivariadas, y no otra, porque es primordial tener en mente qué es lo que se quiere hacer con los datos. En otras palabras a qué se quiere llegar con el análisis estadístico- el

objetivo. Una vez que se tiene idea de lo que se quiere, se pude ir al siguiente paso, que es la selección de la técnica. Esto nos conduce a la búsqueda de una pequeña descripción de las técnicas y de sus usos.

3.4. Breve Descripción de las Técnicas por Objetivo del Análisis Multivariado

El propósito de cualquier técnica de análisis multivariado es el obtener alguna idea de la estructura y de las características principales de los datos, que pueden tomar diversas formas. Podría quererse dividir los datos en grupos, o bien decidir si un grupo conocido está asociado a ciertas medidas. Alternativamente, podría ajustarse algún modelo matemático a los resultados. O bien, puede ser que las medidas sean de dos tipos y se requiriera que se asocie una con otra. Todas estas situaciones necesitan diferentes tipos de análisis, y es importante reconocer exactamente qué preguntas pueden contestarse y qué tipo de análisis puede contestarlas (Marriott, 1974) [30].

3.4.1. Técnicas Usadas para la Reducción de Dimensionalidad

Las técnicas usadas para la reducción de datos son: Análisis de Componentes Principales, Análisis de Factores, Análisis de Correspondencias, Escalamiento Multidimensional, Análisis Conjunto y el Análisis de Correlación Canónico. Su propósito es el de expresar el contenido esencial de los datos en pocas dimensiones, que haga fácil su entendimiento y su manejo matemático. Es esencial la representación preliminar gráfica para encontrar alguna redundancia en los datos originales, es decir ver si existen variables linealmente relacionadas, ya sea exactamente o casi exactamente. Si es el caso surgirán dificultades en el análisis que pueden ser eliminadas por medio de la reducción del número de variables. Las variables resultantes de la simplificación estructural, podrán dar una hipótesis ideal de la estructura de los datos.

Análisis de Componentes Principales

La técnica del Análisis de Componentes Principales (Principal Component Analysis, PCA) fue descrita por Karl Pearson (1901) por primera vez; aparentemente creyó que esta era la solución correcta para algunos problemas biométricos, aunque no propuso un método práctico para el cálculo de más de dos o tres variables. Más tarde, una descripción del método de cálculo práctico fue presentada por Hotelling (1933) (Manly, 1994:76) [29].

El Análisis de Componentes Principales (PCA) es una técnica que permite reducir la dimensión de un grupo de datos, excesivamente grande por el elevado número de variables que contiene, y quedarse con unas cuantas variables. Variables que son una combinación de las iniciales (componentes principales) perfectamente calculables y que sintetizan la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables. Pero como sólo se retienen las componentes (componentes principales) que explican un porcentaje alto de la variabilidad de las variables iniciales se reduce el número de variables de forma importante.

En el Análisis de Componentes Principales las variables tienen que ser de intervalo o de razón (métricas o cuantitativas). Las componentes deben de ser suficientes para resumir la mayor parte de la información contenida en las variables originales. Así mismo cada variable original podrá expresarse en función

de las componentes principales, de modo que la varianza de cada variable original se explica completamente por las componentes cuya combinación lineal la determinan.

Análisis Factorial

El desarrollo temprano del Análisis de Factores (Factor Analysis) fue atribuido a Charles Sperman (1904), que estudió las correlaciones entre calificaciones de varios tipos y notó que muchas correlaciones observadas podían ser analizadas mediante un simple modelo de puntuaciones (Manly, 1994:93) [29].

El Análisis de Factores esta diseñado para explicar un conjunto de datos multivariados, en términos de un conjunto pequeño de factores subyacentes de forma objetiva. Se trata de encontrar una estructura subyacente de la matriz de datos. Idealmente, las variables independientes son normales y continuas, con al menos 3 o 5 variables pesando en el factor. El tamaño de la muestra debería ser mayor a 50 observaciones, con más de 5 observaciones por variable. La Multicolinealidad es deseada entre las variables por que las correlaciones son la clave para la reducción de datos. La medida de suficiencia estadística Kaiser es una medida del grado a el cual cada variable puede ser estimada por todas las otras variables. Una medida suficiencia estadística Kayes de .80 o mayor es muy buena y una medida menor de .50 es muy pobre.

El análisis de factores extrae factores basados en la varianza compartida por los factores. El primer factor extraído explica la mayor varianza. Típicamente, los factores son extraídos siempre y cuando los eigenvalores sean mayores a 1.0 o bien que la gráfica indique cuántos factores se extraerán. Los pesos factoriales son las correlaciones entre el factor y las variables. Usualmente un peso factorial de .4 o mayor es requerido para atribuir una variable específica a un factor. La rotación ortogonal asume ninguna correlación entre los factores, mientras que una rotación oblicua se usa cuando se cree que alguna relación existe. Por último en el Análisis de Factores las variables tienen que ser de intervalo o de razón (métricas o cuantitativas). Los factores deben de ser suficientes para resumir la mayor parte de la información contenida en las variables originales.

Análisis de Correspondencias

El Análisis de Correspondencias (Correspondence Analysis) durante un largo periodo fue usado rutinariamente y exclusivamente en Francia por Jean-Paul Benzérci a principios de 1960. El Análisis de Correspondencias como método de ordenamiento se origino en los trabajos de Hirschfeld (1935), así como en los realizados por Fisher (1940). Es hoy en día el método más popular de clasificación para ecologistas. El método se explica en el contexto de la ordenación de lugares sobre la base de la abundancia de especies (Electronic Textbook Statsoft,2003) [47].

El Análisis de Correspondencia es una técnica descriptiva-exploratoria diseñada para analizar tablas de dos entradas o de entradas múltiples que contienen medidas de correspondencia entre filas y columnas. Los resultados proveen información similar en naturaleza a la técnica de Análisis Factorial que permiten explorar la estructura de las variables categóricas incluidas en la tabla. La tabla más común es de dos entradas.

Es decir se estudia conjuntamente el comportamiento de dos variables nominales u ordinales, aunque este análisis puede ser generalizado para el caso en que se dispone de un número de variables cualitativas mayor a dos. El objetivo del Análisis de Correspondencias es establecer relaciones entre variables nominales u ordinales (no métricas o cualitativas), enriqueciendo la información que ofrecen las tablas de contingencia, que sólo comprueban si existe alguna relación

entre las variables y la intensidad de dicha relación. El Análisis de Correspondencias revela además en qué grado contribuyen a esa detectada relación los distintos valores de las variables, información que suele ser proporcionada en modo gráfico.

Es decir el Análisis de Correspondencias tiene como objetivo el estudio de la asociación entre las categorías de múltiples variables no métricas, pudiendo obtenerse un mapa perceptual que ponga de manifiesto esta asociación en modo gráfico. Es usado para medir la efectividad de campañas publicitarias. El modelo puede ser evaluado por medio del valor de la Chi- cuadrada.

El Análisis de Correspondencia Múltiple puede ser considerado como una extensión del análisis simple de correspondencias de más de dos variables. El Análisis de Correspondencia Múltiple es un Análisis de Correspondencia simple que involucra una matriz-indicador (o diseño) con los casos como filas y las categorías de las variables como columnas. De hecho, se analiza el producto interno de la matriz.

Escalamiento Multidimensional

Los dos principales tipos de procedimientos del Escalamiento Multidimensional (Multidimensional Scaling) son llamados escalamiento clásico y escalamiento ordinal. El primero de los dos procedimientos es esencialmente un método algebraico de reconstrucción de puntos-coordenadas asumiendo que las disimilaridades son distancias Euclidianas, aunque el método es robusto a la situación en donde las distancias son errores distorsionados. El método fue originalmente propuesto por Torgerson (1952,1958) y es usualmente llamado escalamiento clásico o métrico por los psicólogos, fue popularizado por Coger (1966) bajo el nombre de análisis de coordenadas principales. Una alternativa técnica posterior fue desarrollada por R.N. Shepard y J.B. Kruskal a principios de 1960, en la cual solo se usaron las propiedades ordinales de las disimilaridades. Varias alternativas se han desarrollado desde entonces, como el método de Guttman- Lingoes, pero el método comúnmente usado es el propuesto por Kruskal. Un tratamiento riguroso es dado por Sibson (1981). El titulo original sugerido por Kruskal para este tipo de planteamiento fue escalamiento no-métrico multidimensional.

El Escalamiento Multidimensional (MDS) puede ser considerado como una alternativa del análisis de factores. En general, el propósito del análisis es el de detectar dimensiones subyacentes significativas que permitan explicar las similaridades o disimilaridades (distancias) observadas entre los objetos investigados. Con el escalamiento multidimensional se puede analizar cualquier tipo de matriz de similaridades o disimilaridades, en conjunto con la matriz de correlación.

La técnica de Escalamiento Multidimensional no requiere que los datos estén distribuidos normalmente, así como no requiere que las relaciones sean lineales; siempre y cuando el orden de las distancias (o similaridades) en la matriz sean significativas. El Escalamiento Multidimensional ofrece soluciones interpretables y más legibles. MDS puede ser aplicado a cualquier tipo de distancias o similaridades.

El Escalamiento Multidimensional tiene como propósito el crear una representación gráfica (mapa perceptual) que permita conocer la situación de los individuos en un conjunto de objetos por posiciones de cada uno en relación a los demás. Dicha situación será producto de las percepciones y preferencias o similaridades entre los objetos apreciadas por los sujetos. Estas similaridades son la entrada del análisis y pueden ser variables métricas o no métricas. El Escalamiento Multidimensional transforma estas variables en distancias entre los objetos en un espacio de dimensiones múltiples, de modo que objetos que aparecen situados más próximos entre sí son percibidos como más similares.

Existe una diferencia clave entre el Escalamiento Multidimensional y el Análisis de Conglomerados. En el primero se desconocen los elementos de juicio de los encuestados y no se conocen las variables que implícitamente están considerando éstos para realizar su evaluación de las preferencias por los objetos. En el último las similitudes entre los objetos se obtienen a partir de una combinación de las variables estudiadas.

Análisis de Correlación Canónica

El Análisis de Correlación Canónica (Canonical Correspondence Analysis) fue desarrollado por Hotelling (1935-1936). Sus aplicaciones son discutidas por Cooley y Lohnes (1971), Kshirsagar (1972) entre otros. Esta es una técnica para analizar las relaciones entre dos conjuntos de variables. Cada conjunto contiene varias variables. La correlación múltiple y simple son casos especiales de la correlación canónica en la cual uno o ambos conjuntos contienen una sola variable respectivamente.

Existen diversas medidas de correlación para expresar la relación entre dos o más variables. Por ejemplo el coeficiente de correlación momento producto de Pearson mide la magnitud de relación entre dos variables; existen medidas no paramétricas de relaciones que están basadas en la similaridad de rangos de dos variables. El Análisis de Regresión Múltiple permite estimar la relación entre una variable dependiente y un conjunto de variables independientes. El Análisis de Correspondencias Múltiple es útil para explorar la relación entre un conjunto de variables categóricas.

El Análisis de Correlación Canónica es un procedimiento adicional para estimar la relación entre variables. Específicamente, este análisis permite investigar la relación entre dos conjuntos de variables. El Análisis de Correlación Canónica se enfoca en la correlación entre una combinación lineal de las variables en un conjunto y una combinación lineal de las variables en otro conjunto. La idea es primero el de determinar el par de combinaciones lineales que tengan la correlación mas grande. Después, determinar el par de combinaciones lineales que tenga la correlación más grande entre todos los pares no correlacionados con el par seleccionado inicialmente. El proceso continúa. Los pares de combinaciones lineales son llamados variables canónicas, y sus correlaciones son llamadas correlaciones canónicas.

El Análisis de Correlación Canónica mide la fuerza de asociación entre los dos conjuntos de variables. El aspecto de maximización de la técnica representa un intento de concentrar una relación grande de dimensión entre dos conjuntos de variables en un conjunto pequeño de pares de variables canónicas.

Siendo la más flexible de todas las técnicas multivariadas, correlación canónica correlaciona simultáneamente varias variables independientes y varias variables dependientes. Esta poderosa técnica utiliza variables independientes de intervalo o de razón (métrica o cuantitativa), a diferencia de la MANOVA. También puede utilizar variables nominales u ordinales (no-métricas o cualitativas). Esta técnica tiene menos restricciones que cualquier técnica multivariada, por lo que debe ser interpretada con precaución.

Análisis Conjunto

El Análisis Conjunto (Conjoint Analysis) es un método que esta basado en trabajos de Luce y Tukey (1964) y en los métodos discretos de econometría de McFadeen (1974) premio novel de economía.

El Análisis Conjunto es una técnica que se utiliza para analizar la relación lineal o no lineal entre una variable, generalmente ordinal –aunque también puede

ser métrica- y varias variables independientes no métricas. La variable dependiente señala la preferencia (intención de compra, etc.) que el individuo muestra hacia el producto y las variables dependientes son los atributos distintivos del producto.

Es importante tener presente que sólo la variable dependiente recogerá información aportada por los individuos encuestados, ya que la información contenida en las variables independientes será especificada por el investigador en virtud de los productos que desee someter a evaluación por los encuestados.

El Análisis Conjunto permite generar un modelo individualizado por encuestado, de modo que le modelo general para toda la muestra resulte de la agregación de los modelos de todos los individuos que la componen. El Análisis Conjunto descompone las preferencias que el individuo manifiesta hacia el producto a fin de conocer qué valor le asigna a cada atributo, mientras que en el análisis discriminante y en el análisis de regresión las valoraciones de cada atributo que hace el sujeto se utilizan para componer su preferencia sobre el producto.

3.4.2. Técnicas usadas para Clasificación

Como se había mencionado, de acuerdo a las características de las medidas se crean grupos o variables similares, tan simples como sea posible. El Análisis Discriminante y la clasificación son técnicas enfocadas en separar distintos conjuntos de objetos (u observaciones) así como el de asignar nuevos objetos a grupos previamente definidos.

Análisis de Conglomerados

La historia del Análisis de Conglomerados (Cluster Analysis) se ha generado alrededor del desarrollo de algoritmos. Los métodos que facilitan la realización de procedimientos directos en computación, han usado algoritmos que involucran análisis de conglomerados. Estos algoritmos han probado ser muy útiles y pueden ser encontrados en la mayoría de los paquetes de software.

El termino "Análisis de Conglomerados" alberga a diferentes algoritmos y métodos para agrupar objetos de tipo similar en categorías respectivamente. Una pregunta que confronta a investigadores en muchas áreas es la de cómo organizar datos observados en estructuras significativas, esto es, el desarrollo de taxonomías. En otras palabras el Análisis de Conglomerados es una herramienta de análisis exploratorio de datos cuyo objetivo es el de clasificar diferentes objetos en grupos de tal forma que el grado de asociación entre dos objetos es máxima si ellos pertenecen al mismo grupo y mínima de lo contrario. El Análisis de Conglomerados puede ser usado para descubrir estructuras en los datos sin proveer una explicación-interpretación, en otras palabras, simplemente descubre estructuras en los datos sin explicar por qué existen.

La creación de grupos basados en similaridades exige una definición de similaridad o de su complemento (distancia). Existen muchas formas de medir estas distancias y diferentes reglas matemáticas para asignar los individuos a distintos grupos, dependiendo del fenómeno estudiado y del conocimiento previo del posible agrupamiento que se tenga.

El Análisis de Conglomerados suele comenzar estimando las similitudes entre los individuos (u objetos) a través de correlación (distancia o asociación) de las distintas variables (métricas o no métricas) de que se dispone. A continuación se establece un procedimiento que permite comparar los grupos en virtud de las similaridades. Por último se decide cuántos grupos se construyen, teniendo en cuenta que cuanto menos sea el número de grupos, menos homogéneos serán

los elementos que integran cada grupo. Se buscará formar el mínimo número de grupos lo más homogéneos posibles dentro de sí y lo más heterogéneos posibles entre sí.

En resumen el propósito del análisis de conglomerados es reducir un conjunto grande de datos en subgrupos significativos de individuos u objetos. La división se logra por medio de similaridades que existen entre objetos. Los outliers son un problema en esta técnica, causada frecuentemente por muchas variables irrelevantes. La muestra tiene que ser representativa de la población. Existen tres métodos principales de conglomerados: jerárquico, el cual es un proceso de árbol apropiado para un conjunto pequeño de datos; no-jerárquico, el cual requiere especificaciones del número de conglomerados por adelantado y el tercero es una combinación de los dos. Existen cuatro reglas principales para crear conglomerados: los conglomerados tienen que ser diferentes, tienen que ser mensurables, tienen que ser localizables y los conglomerados tienen que ser útiles y significativos.

Análisis Discriminante

El análisis de discriminante (discriminant analysis) es una versión gráfica de la MANOVA, la cual busca combinaciones de variables observadas que indiquen una diferencia significativa en la medias de grupos tratados. Esta técnica fue desarrollada en 1930 cuando se trabajaba sobre un mismo problema pero desde diferentes perspectivas. Fisher esta interesado en desarrollar una técnica para distinguir dos grupos de datos multivariados; su contribución fue la función lineal discriminante de Fisher. Hotelling desarrolló la prueba T^2 de Hotelling como un medio para probar diferencias significativas entre los centroides de dos muestras multivariadas. Mahalanobis estaba decidido a encontrar una manera de medir distancias multivariadas entre centroides de dos muestras; la distancia Mahalanobis es una extensión del teorema de Pitágoras para conjuntos de variables correlacionadas.

Cuando estas técnicas fueron reunidas, se creó el análisis de discriminante canónico. Era posible medir distancias entre las medias de dos muestras mediante la distancia de Mahalanobis, determinar si esa distancia es significativamente diferente de cero usando la prueba T^2 de Hotelling y desarrollar una ecuación de regresión (función lineal discriminante) permitiéndonos asignar nuevas especies a uno de los dos grupos.

La función del Análisis Discriminante (Discriminant Analysis) es la de determinar cuáles variables discriminar entre dos o más grupos que ocurren en forma natural. Por ejemplo, un investigador de la educación desea investigar cuáles variables discriminar entre estudiantes graduados de preparatoria quienes deciden: asistir a la universidad, ir una escuela técnica o no seguir estudiando. Para este propósito se podrían recolectar datos sobre numerosas variables. La mayoría de los estudiantes naturalmente caerán en una de las tres categorías. El Análisis Discriminante podría ser usado para determinar cuáles variables son las mejores para predecir la educación que seguirán los estudiantes.

El propósito del análisis discriminante es el de clasificar correctamente observaciones o gente en grupos homogéneos. Las variables deben ser de intervalo o de razón (métricas o cuantitativas) y tienen que tener un grado alto de normalidad. El análisis discriminante crea una función lineal discriminante, la cual puede ser usada para clasificar observaciones. El ajuste es logrado mediante el grado al cual las medias del grupo difieren. Para determinar cuáles variables tienen más impacto en la función discriminante, es posible ver los valores parciales de F. Entre más grande sea el valor parcial de F mayor impacto esa variable tendrá sobre la función discriminante. Esta técnica ayuda a categorizar grupos.

3.4.3. Técnicas usadas para la Relación entre dos conjuntos de variables

El Análisis de Regresión Múltiple es la técnica univariada utilizada para encontrar las relaciones entre una sola variable y un conjunto de variables relacionadas. Es natural el considerar que la extensión de este proceso es posible para relacionar varias variables a otro conjunto de variables (que podrían ser o no variables aleatorias).

Una clásica extensión de la Regresión Múltiple al caso multivariado involucra calcular Correlaciones Canónicas entre variables canónicas, las cuales son combinaciones lineales de las variables originales. En la práctica, las aplicaciones usuales han sido en el caso cuando el segundo conjunto de variables representa las diferencias entre grupos. Justo como un análisis de la varianza "entre" y "dentro", los grupos pueden ser estimados como un caso especial de regresión múltiple, tal que la extensión multivariada del Análisis de la Varianza puede ser tratada mediante métodos de Análisis Canónico; y son usados frecuentemente junto con el análisis discriminante para investigar las relaciones entre un conjunto de variables y una agrupación conocida.

Análisis de Regresión Múltiple

De acuerdo a publicaciones de Sir Francis Galton y Karl Person, Galton da la conceptualización inicial de regresión lineal, en un trabajo sobre las características hereditarias de los chíncharos. Esfuerzos subsecuentes de Galton y Person dan la técnica general de regresión múltiple y el coeficiente de correlación producto-momento. En la literatura moderna se presenta típicamente y se explica correlación antes de introducir problemas de predicción y la aplicación de regresión lineal [46] (Stanton, 2001).

El propósito general del Análisis de Regresión Múltiple (Multiple Regression Analysis) es analizar la relación entre una única variable dependiente y varias variables independientes. El objetivo es utilizar las variables independientes, cuyos valores son conocidos, para predecir los valores de una única variable dependiente. Para que finalmente se obtenga una combinación lineal de todas o de un conjunto de las variables independientes que correlacione máximamente la variable independiente. En este proceso, cada variable es ponderada, dando un porcentaje de contribución relativa de cada variable independiente a la predicción total. Así mismo se pretende seleccionar un subconjunto de todas las variables, que proponen suficiente información, sin tener que utilizar todas las variables, puesto que posiblemente se encuentren redundancias entre las variables independientes. Al obtener los pesos de la combinación lineal se utiliza un procedimiento de optimización, consistente en la minimización de los errores de predicción al cuadrado o residuos al cuadrado (método de mínimos cuadrados), que asegura la predicción máxima del criterio. De igual manera estos pesos facilitan la interpretación de la influencia de cada variable en el modelo, aunque la presencia de correlaciones entre las variables independientes dificulta su comparación.

El análisis de regresión es una técnica de dependencia que puede utilizarse únicamente con variables dependientes e independientes métricas o cuantitativas (de intervalo o de razón), aunque utilizando procedimientos de codificación para variables "dummy", es posible incluir datos no métricos o cualitativos (escala nominal u ordinal), es posible incluir datos no métricos como variables independientes. La regresión múltiple seguramente es la técnica mas utilizada dentro del análisis multivariado, que puede usarse para muchos propósitos. Aunque se podrían resumir sus principales aplicaciones en dos grupos: problemas de predicción y problemas de explicación. Sea cual sea la finalidad hay que tener

presente que el análisis de regresión revela relaciones entre variables. En general, el análisis de regresión permite a los investigadores preguntar "cual es el mejor predictor para...".

Regresión Logística

Sus orígenes se remontan al siglo XIX, cuando fue inventada la función de crecimiento de la población, su nombre fue dado por el matemático Belga Verhulst. Los eventos subsiguientes han sido determinados decisivamente por acciones individuales e historias personales de un grupo de estudiantes: el redescubrimiento de la función de crecimiento se debió a Peral y Reed, el término logístico a Yule y la introducción de la función bio-assay a Beckson (Cramer, 2003) [11].

Es una forma alternativa del análisis de regresión. Su uso está empleado en situaciones en que la variable dependiente toma únicamente dos valores, que indican la pertenencia a uno de dos grupos, normalmente etiquetados como 0 y 1.

Su utilización fundamental se encuentra en el análisis de variables dependientes o fenómenos que son dicotómicos por naturaleza. Aplicar regresión lineal en estas situaciones violaría muchos supuestos de los que el más crítico es que el término error sigue la distribución binomial en vez de la normal. El modelo de Regresión Logística permite predecir o estimar la probabilidad de que un individuo caiga en un estado en función de determinadas características individuales, las variables predictorias. El modelo permite identificar la importancia y contribución relativa de determinadas características individuales, denominados en el contexto epidemiológico factores de riesgo. El modelo permite estimar o predecir la magnitud del riesgo global cuando coinciden dos o más factores de riesgo.

Análisis de Series de Tiempo

El Análisis de Series de Tiempo contemporáneo tiene sus orígenes en las ciencias físicas y sociales. Los conceptos básicos han aparecido en ambas áreas través de transferencia de tecnología. Los investigadores importantes históricos en el desarrollo de ésta área son: Thiele, Hooker, Einstein, Wiener, Yule, Fisher, Tukey, Whittle y Bartlet. Las herramientas empleadas para direccionar problemas de series de tiempo incluyen: modelos matemáticos, métodos asintóticos, análisis funcional y transformaciones [5] (Brillinger,2000:4).

Este método es útil para analizar datos de series de tiempo, que son, secuencias de medidas que no siguen un orden aleatorio. A diferencia de los análisis de muestras aleatorias, el análisis de series de tiempo esta basado en la suposición de que los valores sucesivos en el archivo de datos representan medidas consecutivas tomadas a intervalos de tiempo iguales.

Existen dos principales objetivos en el análisis de series temporales, uno es el de identificar la naturaleza del fenómeno representado por la secuencia de observaciones y el otro es predecir (predecir valores futuros de la variable serie de tiempo). Ambos objetivos requieren que el patrón de datos de series de tiempo observadas sea identificado y más o menos descrito formalmente. Una vez que el patrón es establecido, se puede interpretar e integrar con otro dato. A pesar del profundo entendimiento y de la validación de la interpretación (teórica) del fenómeno, se puede extrapolar el patrón identificado para predecir los eventos futuros.

Objetivo	Técnica Multivariada
Reducción de Datos o Simplificación Estructural	Análisis de Componentes Principales Análisis Factorial Análisis de Correspondencias Escalamiento Multidimensional Análisis Conjunto Análisis de Correlación Canónico
Clasificación	Análisis de Conglomerados Análisis de Discriminantes
Relación entre dos conjuntos de variables	Análisis de Regresión Regresión Logística Series de Tiempo
Pruebas de Hipótesis	MANOVA Análisis de la varianza múltiple MANCOVA Análisis de la covarianza múltiple

Cuadro 3.1: Clasificación de Técnicas de acuerdo al Objetivo

3.4.4. Técnicas usadas para pruebas de hipótesis

Estas técnicas tienen un carácter esencialmente descriptivo, pero con otras pueden ponerse a prueba hipótesis sobre modelos complejos basados en poblaciones multivariantes.

Análisis de la Varianza Múltiple (MANOVA)

Fue desarrollado de forma teórica por S.S. Wilkins en el año de 1932 y publicado en Biométrica.

El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medidas de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes. Las variables dependientes son métricas y las variables independientes son no métricas.

Análisis de la Covarianza Múltiple (MANCOVA)

El objetivo del análisis multivariado de la covarianza, es el de eliminar el efecto de una o más variables perturbadoras de la variable dependiente. Así mismo es una técnica estadística utilizada para analizar la relación entre varias variable dependiente métrica o cuantitativa (de intervalo o de razón) y varias variables independientes que son mezcla de variables cuantitativas y cualitativas. En el análisis de la covarianza, tanto simple como múltiple, las variables cuantitativas independientes (covariables) tienen como objetivo eliminar determinados efectos que puedan sesgar los resultados incrementando la varianza dentro de los grupos. El análisis de la covarianza se suele comenzar eliminando, mediante una regresión lineal, la variación de efectos no deseados.

Esta exposición-resumen de técnicas se pueden resumir en el cuadro 3.1, de acuerdo al propósito u objetivo de la técnica.

3.5. Comentarios

A lo largo de este capítulo se definió qué es el Análisis Multivariado, se dieron ejemplos de aplicaciones actuales en diferentes áreas del conocimiento, se listaron

Objetivo	Técnica Multivariada	Tipo de Variable	Usos y Propósitos Principales
Clasificación	Análisis de Conglomerados	v.i. Nominal u Ordinal v.d. Intervalo o de razón	El Análisis de Conglomerados es la clasificación de objetos dentro de grupos diferentes, o mas precisamente, la partición de un conjunto de datos en subconjuntos (conglomerados), tal que los datos en cada subconjunto (idealmente) comparten algunas caracteristicas.
	Análisis de Discrimiantes	v.d. Nominal u Ordinal v.i. de Intervalo o de razón	El Análisis Discriminante intenta establecer si un conjunto de varibles puede ser usado para distinguir entre dos o más grupos.

Cuadro 3.2: Sintesis de Técnicas Multivariadas usadas para Clasificación

las técnicas que conforman los métodos multivariados -así como sus objetivos-, asimismo se señaló el tipo de variable apropiado para las técnicas- aunque este último no de forma puntual. De igual modo, se indicó que la correcta selección no sólo depende del propósito del análisis, que tiene que ser definido por el analista, sino también del correcto manejo y entendimiento de las variables que se han de manejar durante el análisis. Por todo lo anterior y a forma de resumen se puede decir que existen dos aspectos muy importantes para la selección de una técnica multivariada: el propósito del análisis y el tipo de variable. El tipo de variable será abordado en los cuadros-resumen.

Esto es los cuadros muestran una sintesis que contempla el objetivo, la técnicas multivariadas, el tipo de variable idónea y una breve descripción de sus usos y propósitos. Se dividen en cuatro tablas de acuerdo al objetivo: Reducción de datos o simplificación estructural 3.3, Clasificación 3.2, Relación entre dos conjuntos de variables 3.4 y pruebas de hipótesis 3.5.

Objetivo	Técnica Multivariada	Tipo de Variable	Usos y Propósitos Principales
Reducción de Datos o Simplificación Estructural	Análisis de Componentes Principales	Intervalo o de razón	El Análisis de Componentes Principales (PCA) trata de determinar un conjunto pequeño de varibles sintetizadas que puedan explicar el conjunto original
	Análisis Factorial	Intervalo o de razón	Factor analysis es una tecnica usada para explicar variabilidad entre variables aleatorias observadas en terminos de algunas variables no observadas llamadas factores.
	Análisis de Correspondencias	Nominal	El Análisis de Correspondencia es conceptualmente similar a PCA, pero escala los datos (los cuales tienen que ser positivos) tal que las filas y columnas son tratada equivalentemente.
	Escalamiento Multidimensional	Ordinal	El Escalamiento Multidimensional (MDS) envuelve varios algoritmos para determinar un conjunto de variables sintetizadas que mejor representan la distancias entre distancias por pares entre los puntajes.
	Análisis Conjunto	v.d. Ordinal v.i. Intervalo o de razón	El Análisis Conjunto tambien llamado modelo composicional multi-atributo o Análisis de preferencias.
	Análisis de Correlación Canónico	Intervalo o de razón	Un uso típico de correlación canonica es tomar a dos conjuntos de variables y ver que es comun entre los dos conjuntos.

Cuadro 3.3: Sintesis de Técnicas Multivariadas usadas para Reducir Datos

Objetivo	Técnica Multivariada	Tipo de Variable	Usos y Propósitos Principales
Relación entre dos conjuntos de variables	Análisis de Regresión Múltiple	Intervalo o de razón	El Análisis de Regresión trata de determinar una formula lineal que puede describir como algunas variables responden a cambios en otras. El Análisis de Regresión esta basado en modelo de lineal general.
	Regresión Logística	v.d. Dicotómica o Multinomial v.i. Nominal, Ordinal o de Intervalo o de razón	Regresión Logística permite realizar un análisis de regresión para estimar y probar la influencia de covariantes sobre una respuesta dicotomica múltiple.
	Series de Tiempo	v.i. intervalo (tiempo) v.d. Intervalo o de razón	El analisis de series de tiempo se usa para entender series de tiempo para entender la teora subyacente de los datos en el tiempo o bien para hacer predicciones. Especificamente para predecir puntos antes de que sean medidos.

Cuadro 3.4: Sintesis de Técnicas Multivariadas usadas para relacionar dos conjuntos de variables

Objetivo	Técnica Multivariada	Tipo de Variable	Usos y Propósitos Principales
Pruebas de Hipótesis	MANOVA Análisis de la Varianza Múltiple	v.i. Nominal u Ordinal v.d. Intervalo o de razón	La MANOVA es una extensión del metodo de análisis de la varianza (ANOVA) que subre casos donde hay mas de una variable dependiente y donde las variables dependientes no pueden ser simplemente combinadas. Tambien como indentificar si los cambios en la variable independiente tiene un efecto significativo sobre las variables dependientes, esta técnica tambien sirve para identificar la interacción entre las variables independientes y la asociació entre variables dependientes, si hay alguna.
	MANCOVA	v.i. Nominal u Ordinal v.d. Nominal, Ordinal o de Intervalo o de razón	Multiple Análisis de la Covariaza (MANCOVA) es similar al análisis de varianza multiple, pero permite controlar los efectos de variables independientes suplementarias (covariantes). Si existen covariantes se tiene que usar MANCOVA en lugar de MANOVA.

Cuadro 3.5: Sintesis de Técnicas Multivariadas usadas para Pruebas de Hipótesis

Capítulo 4

Medidas de Proximidad entre grupos y objetos

El ser humano, desde sus inicios, ha querido simplificar, de forma sistemática, lo que le rodea. El objetivo fundamental es separar o diferenciar objetos de acuerdo a si poseen o no diferentes características. Existen diferentes clasificaciones, pero lo que todas las clasificaciones tienen en común es el objetivo que persiguen: separar o dividir lo que se estudia en diferentes clases o grupos, de modo tal que todos los objetos que pertenecen a una misma clase sean parecidos entre sí y diferentes a los objetos que pertenecen a otras clases.

La motivación de la clasificación es hacer una investigación sistemática sobre objetos, con el fin de establecer si pueden ser resumidos en términos de un número pequeño de clases de objetos similares. Una gran variedad de medidas de similaridad y disimilaridad se han propuesto para lograr una clasificación válida. Al resumir un conjunto de datos se espera que éste describa una colección grande de objetos que sea relevante para el estudio. Estos resúmenes de datos pueden permitir hacer predicciones o descubrir hipótesis que describan la estructura de los datos. Estas predicciones pueden variar en niveles de sofisticación.

La terminología en esta sección se concentra en la descripción de medidas de disimilaridad y similaridad entre un par de objetos descritos por un conjunto de variables, aunque medidas en donde cada objeto es un grupo de objetos también son presentadas. Así mismo, se presentan métodos usados para medir la proximidad entre objetos.

En otras palabras, en esta sección se examinarán medidas que se utilizarán en técnicas de reducción de datos. Algunas técnicas se centran en la reducción del número de variables o columnas de la matriz de datos X. Las técnicas que se tratarán se centraran en la reducción del número de filas de X. Puesto que las filas de X representan unidades de observaciones, el procedimiento consiste en combinar las unidades en grupos de unidades de relativa homogeneidad. Estas técnicas o procedimientos de reducción de datos serán usados en el llamado Análisis de Conglomerados (Cluster Analysis).

En algunas situaciones, es difícil obtener medidas precisas en la forma de una matriz de datos X. De cualquier manera es posible obtener una matriz proximidad que provea información acerca de los grados de similaridad entre las unidades observadas.Y con estas será posible determinar dimensiones que son consistentes con la proximidad dada. Estas técnicas son comúnmente referidas como Escalas Multidimensionales.

4.1. Matrices de Proximidad Derivadas de las Matrices de Datos

La matriz de datos multivariada X de orden $(n \times p)$, consiste de observaciones obtenidas de las medidas de n objetos con respecto a p aspectos o características. Las p columnas de X son usualmente referidas como variables mientras que las n filas son comúnmente llamadas características de las unidades observadas. Una característica (profile) es simplemente un vector de medidas cuyos elementos serán comparados. En este trabajo las características son n los vectores de orden $(1\times)p$ que constituyen X.

$$X = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}$$

Los elementos son denotados por a_ij, $i = 1, 2, \ddots, n$, $j = 1, 2, \ddots, p$, en donde el primer subíndice i se refiere a la fila y el segundo subíndice j se refiere a la columna.

Una matriz de proximidad es una matriz de orden $(n\times n)$ que resume los grados de similaridad o disimilaridad entre todos los pares posibles de características en X. Esta matriz esta denotada por P con elementos p_{rs}, $r, s = 1, 2, 3, ..., n$. El elemento p_{rs} denota las medida de proximidad entre unidades observacionales r y s . La matriz XX' es un ejemplo de una matriz de proximidad.

Una variedad de medidas de proximidad se introducirán en esta sección. Estas Medidas de proximidad son también presentadas para relacionar dos grupos de unidades de observación. Las medidas de proximidad mostradas serán usadas en el estudio de Análisis de Conglomerados y Escalas Multidimensionales.

4.1.1. Proximidad

Proximidad literalmente significa cercanía en espacio, en tiempo o en otra dimensión. La "cercanía" de objetos necesita ser definida y medida antes del análisis estadístico. Las medidas de proximidad son de dos tipos: similares y disimilares, con la obvia interpretación de medida de qué tan similares o disimilares los objetos son entre sí.

Dados los objetos bajo consideración de un conjunto X. La medida de similaridad/disimilaridad entre dos objetos es entonces una función real definida sobre $X \times X$, dando origen a la similaridad s_{rs} o disimilaridad d_{rs} entre el r-ésimo y s-ésimo objeto. Usualmente $d_{rs} > 0$ y $s_{rs} > 0$ y la disimilaridad de un objeto consigo mismo es cero. La similaridad es usualmente estimada de tal manera que la máxima similaridad es la unidad.

Trevor y Michael Cox (2001) retoman las doce estructuras de proximidad posibles, S, que es necesario considerar antes de que una medida de proximidad, en particular, sea escogida, de acuerdo con Cormack (1971) y Hartigan (1967), las estructuras se pueden ver en el cuadro 4.1.

4.1.2. Similaridad

Un coeficiente de similaridad, también conocida como proximidad, indica la fuerza de la relación entre dos objetos dados los valores de un conjunto de p variables en común. La similaridad entre dos objetos r y s puede ser alguna función de sus valores observados, esto es:

Propiedad	Definición
$\mathbf{s_1}$	s definida en $(X \times X)$ es distancia euclidiana
$\mathbf{s_2}$	s definida en $(X \times X)$ es una métrica
$\mathbf{s_3}$	s definida en $(X \times X)$ es simétrica
$\mathbf{s_4}$	s definida en $(X \times X)$ es un valor real
$\mathbf{s_5}$	s completamente ordenado $\leq$ en $(X \times X)$
$\mathbf{s_6}$	s es parcialmente ordenado $\leq$ en $(X \times X)$
$\mathbf{s_7}$	s es una muetra τ definida en X un orden de similaridad parcial $(r, s) \leq (r^t, s^t)$ donde $\sup_1(r, s) \geq \sup_1(r^t, s^t)$
$\mathbf{s_8}$	s es una similaridad relativamente completa ordenada $\leq$ en X $\forall r$ en X, $s \leq_r t$ significa que s es no más similar a r que a t
$\mathbf{s_9}$	s es una similaridad relativamente completa ordenada $\leq_r$ en X
$\mathbf{s_{10}}$	s es una similaridad dicotómica en $(X \times X)$ en la cual $(X \times X)$ esta dividida en un conjunto de pares similares y un conjunto de pares disimilares
$\mathbf{s_{11}}$	s es una similaridad dicotómica en $(X \times X)$ que consiste en pares similares, pares disimilares y pares restates
$\mathbf{s_{12}}$	s es una partición de X en conjuntos de objetos similares

Cuadro 4.1: Doce Posibles Estructuras de Proximidad

$$p_{rs} = f(\bar{x}_j, \bar{x}_s)$$

donde y son los valores de las variables observadas para los objetos. Muchas funciones han sido propuestas dependiendo en parte por el tipo de variables concernientes (cuantitativas, categóricas, binarias, ordinales) y en parte por el tipo de objetos.

Similaridad es usualmente referida como una relación simétrica requerida $(X \times X)$. La mayoría de los coeficientes de similaridad son no negativos y escalados de tal manera que la máxima medida de la escala es igual a la unidad, aunque algunos son de una correlación natural (Everitt 1993).

Dados dos objetos r y s, la medida de proximidad p_{rs} es una medida de similaridad sip_{rs} satisface lo siguiente:

1. $0 \leq p_{rs} \leq 1$ $\forall r, s$;
2. $p_{rs} \Leftrightarrow r, s$ son idénticos;
3. $p_{rs} = p_{sr}$

La medida de similaridad más común es el coeficiente de correlación de Pearson. Como un coeficiente de correlación tiene un rango $(-1, 1)$, es usual el uso tanto como el valor absoluto del coeficiente o como el de sumar 1.0 al valor del coeficiente para después dividirlo entre 2, esto es:

$$q^*_{rs} = \frac{(1 + q_{rs})}{2}$$

En el cuadro 4.2 muestra la síntesis sobre los coeficientes de similaridad propuesta por Trevor y Michel Cox (2001), síntesis de diferentes autores.

4.1.3. Matrices de Similaridad

Las medidas de similaridad derivadas de aproximaciones tipo correlación pueden resumirse en una matriz $n \times n$ resultante de la suma de cuadrados y

Nombre del Coeficiente	**Fórmula**
Braun, Blanque	$s_{rs} = \frac{a}{max(a+b),(a+c)}$
Czekanowski, Sorensen, Dice	$s_{rs} = \frac{2a}{2a+b+c}$
Hamman	$s_{rs} = \frac{a-(b+c)+d}{a+b+c+d}$
Coeficiente Jaccard	$s_{rs} = \frac{a}{a+b+c}$
Kulczynski	$s_{rs} = \frac{a}{maxb+c}$
Kulczynski	$s_{rs} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$
Michel	$s_{rs} = \frac{4(ad-bc)}{(a+d)^2+(a+c)^2}$
Mountford	$s_{rs} = \frac{2a}{a(b+c)+2bc}$
Mozley, Margalef	$s_{rs} = \frac{a(a+b+c+d)}{(a+b)+(a+c)}$
Ochiai	$s_{rs} = \frac{a}{((a+b)(a+c))^2}$
Phi	$s_{rs} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Rogers, Tanimoto	$s_{rs} = \frac{a+d}{a+2b+2c+d}$
Rusell, Rao	$s_{rs} = \frac{a}{a+2b+2c+d}$
Coeficiente de aproximación simple	$s_{rs} = \frac{a+d}{a+b+c+d}$
Simpson	$s_{rs} = \frac{a}{min(a+d),(a+c)}$
Sokal, Sneath, Andenberg	$s_{rs} = \frac{a}{a+2(b+c)}$
Simpson	$s_{rs} = \frac{ad-bc}{ad-dc}$

Cuadro 4.2: Coeficientes de Similaridad Trevor y Michel Cox (2001)

productos punto para objetos en lugar de variables. Para la matriz de datos X la matriz $X\acute{X}$ de $n \times n$ es una suma de cuadrados y producto punto para los n objetos. Para el coeficiente coseno, la matriz de similaridad se deriva de la matriz X^+ en donde X^+ es la matriz estandarizada X que se muestra a continuación:

$$X^+ = \begin{pmatrix} \frac{x_{11}}{\sqrt{\sum_{j=1}^p x_{1j}^2}} & \frac{x_{12}}{\sqrt{\sum_{j=1}^p x_{1j}^2}} & \cdots & \frac{x_{1p}}{\sqrt{\sum_{j=1}^p x_{1j}^2}} \\ \frac{x_{21}}{\sqrt{\sum_{j=1}^p x_{2j}^2}} & \frac{x_{22}}{\sqrt{\sum_{j=1}^p x_{2j}^2}} & \cdots & \frac{x_{2p}}{\sqrt{\sum_{j=1}^p x_{2j}^2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}}{\sqrt{\sum_{j=1}^p x_{nj}^2}} & \frac{x_{n2}}{\sqrt{\sum_{j=1}^p x_{nj}^2}} & \cdots & \frac{x_{np}}{\sqrt{\sum_{j=1}^p x_{nj}^2}} \end{pmatrix}$$

Para el coeficiente de correlación q_{rs}, la matriz de datos esta dada por $X{+}{+}$, la cual contiene medidas de media central y estandarizada como se muestra abajo.

$$X^+ = \begin{pmatrix} \frac{(x_{11}-\bar{x}_1)}{s_1} & \frac{(x_{12}-\bar{x}_1)}{s_1} & \cdots & \frac{(x_{1p}-\bar{x}_1)}{s_1} \\ \frac{(x_{21}-\bar{x}_2)}{s_2} & \frac{(x_{22}-\bar{x}_2)}{s_2} & \cdots & \frac{(x_{2p}-\bar{x}_2)}{s_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(x_{n1}-\bar{x}_n)}{s_n} & \frac{(x_{n2}-\bar{x}_n)}{s_n} & \cdots & \frac{(x_{np}-\bar{x}_n)}{s_n} \end{pmatrix}$$

con

$$\bar{x}_r = \frac{\sum_{j=1}^p x_{rj}}{p}$$

$$s_r = \frac{\sum_{j=1}^p (x_{rj} - \bar{x}_r)^2}{p}$$

Distancia	**Fórmula**
Distancia Euclidiana	$d_{rs} = \sqrt{\sum_{j=1}^{p}(x_{rj} - x_{sj})}$
Distancia Euclidiana Estandarizada	$d_{rs} = \sqrt{\sum_{j=1}^{p} w_j(x_{rj} - x_{sj})}$
Distancia Mahalanobis	$d_{rs} = \sqrt{(\bar{x}_r - \bar{x}_s)^t \sum^{-1}(\bar{x}_r - \bar{x}_s)}$
Métrica City Block	$d_{rs} = \sum_{j=1}^{p} \lvert x_{rj} - x_{sj}\rvert$
Métrica Minkowski	$d_{rs} = \sqrt[\lambda]{\sum_{j=1}^{p} w_j(\lvert x_{rj} - x_{sj}\rvert)^\lambda}$
Métrica Camberra	$d_{rs} = \frac{\sum_{j=1}^{p} \lvert x_{rj} - x_{sj}\rvert}{(x_{rj} - x_{sj})}$
Divergencia	$d_{rs} = \frac{1}{p}\frac{\sum_{j=1}^{p}(x_{rj} - x_{sj})^2}{(x_{rj} - x_{sj})^2}$
Bray-Curtis	$d_{rs} = \frac{1}{p}\frac{\sum_{j=1}^{p} \lvert x_{rj} - x_{sj}\rvert}{\sum_{j=1}^{p}(x_{rj} - x_{sj})}$
Soergel	$d_{rs} = \frac{\sum_{j=1}^{p} \lvert x_{rj} - x_{sj}\rvert}{\sum_{j=1}^{p} max(x_{rj}, x_{sj})}$
Distancia Bhattacharyya	$d_{rs} = \sqrt{\sum_{j=1}^{p}(x_{rj}^{\frac{1}{2}} - x_{rj}^{\frac{1}{2}})^2}$
Wave-Hedges	$d_{rs} = \frac{1}{p}\sum_{j=1}^{p}[1 - \frac{min(x_{rj}, x_{sj})}{max(x_{rj}, x_{sj})}]$
Separación Angular	$d_{rs} = 1 - \frac{\sum_{j=1}^{p} x_{rj}x_{sj}}{\sqrt{\sum_{j=1}^{p} x_{rj}^2 \sum_{j=1}^{p} x_{sj}^2}}$

Cuadro 4.3: Coeficientes de Disimilaridad Trevor y Cox (2001)

y

$$r = 1, 2, 3, ..., n$$

Una tercera forma de matriz X involucra la característica en forma media central que no ha sido estandarizada. En este caso la matriz XX es proporcional a una matriz de covarianzas entre las características.

4.1.4. Disimilaridad

La medida de disimilaridad, también conocida como distancia, será introducida como el complemento de la medida de similaridad. Una medida de proximidad p_{rs} es una medida de disimilaridad si p_{rs} satisface lo siguiente:

1. $p_{rs} \geq 0 \ \forall r, s$;
2. $p_{rs} = 0 \Leftrightarrow r, s$ son idénticos;
3. $p_{rs} = p_{sr}$

La medida de disimilaridad más usada es la distancia Euclidiana. Una medida de disimilaridad alternativa es la distancia de Mahalanobis entre dos observaciones. En el cuadro 4.3 se da una lista de posibles medidas de disimilaridad para datos cuantitativos que son en particular, continuos, discretos posiblemente pero no binarios.Ésta es una síntesis sobre medidas de disimilaridad (distancias) propuesta por Trevor y Michel Cox (2001) que a su vez es una síntesis de diferentes autores Anderberg (1973), Sneath and Sokal (1973), Gordon (1999), entre otros.

4.2. Distancia

Existen diversas formas de considerar la distancia que existe entre dos objetos, en matemáticas, y no sólo la distancia que se usa habitualmente (la distancia

entre dos puntos es la línea recta). Una distancia,d , es una aplicación entre el producto cartesiano, $X \times X$, de un conjunto –el conjunto formado por parejas de elementos de ese conjunto-, y los números reales no negativos, $X^+ \cup 0 = [0, \infty]$, de modo que a cada par de elementos del conjunto X, (r, s), se le asigna un número real no negativo, j:

$$d : X \times X \rightarrow R^+ \cup 0 = [0, \infty)$$
$$d(r, s) = j$$

Todas las distancias en Matemáticas se caracterizan por cumplir ciertas propiedades y solo se consideran distancias aquellas que cumplen con las siguientes propiedades:

1. Esta definida positiva, es decir, la distancia entre dos elementos cualesquiera es mayor o igual a cero:

   $$d_{rs} \geq 0$$

   y es cero si $r = s$, $x_{rj} = x_{sj}$

   $$d_{rs} = 0$$

2. Es simétrica es decir, la distancia de s a r es la misma que la de r a s

   $$d_{rs} = d_{sr}$$

   $$\forall 1 \leq r, s \leq n$$

3. Cumple con la desigualdad del triangulo, es decir, la distancia entre dos puntos cualesquiera,r y s, es menor o igual que la suma de la distancia de r a un tercer punto t, más la distancia de t a s.

   $$d_{rs} \leq d_{rs} + d_{rs}$$

4. Si $r \not= s$

   $$d_{rs} > 0$$

   Si las coordenadas en un diagrama de dispersión de dimensión p representan los valores de p variables, medidas sobre n individuos, en general, los puntos que están muy juntos representan individuos que son similares en características, mientras que los punto que están alejados representan individuos que son disimilares. Es decir, la distancia en un diagrama de este tipo es una medida de disimilaridad. La distancia ordinaria entre dos puntos $(x_1, \ldots, x_p)$ y $(x_1^*, \ldots, x_p^*)$ es llamada distancia Euclidiana y está dada por:

   $$d = \sqrt{(x_1 - x_1^*)^2 + \ldots + (x_p - x_p^*)^2}$$

4.3. Medidas de Proximidad entre Objetos

Las medidas de proximidad usualmente reflejan grados de similaridad o grados de disimilaridad. Cuando dos objetos se vuelven más similares, el valor de una medida de similaridad se incrementa mientras que la correspondiente medida de disimilaridad declina en valor. Un ejemplo de una medida de similaridad entre dos objetos es un coeficiente de correlación entre los objetos basados en p medidas. Otro ejemplo de una medida de similaridad basada sobre la muestra

de p observaciones es la distancia Euclidiana entre dos objetos. Los dos tipos de medida de proximidad serán definidos de manera más general. Se consideran cuatro grandes tipos de medidas de proximidad, de acuerdo con Sneath y Sokal (1973):

1. Distancias: son diferentes medidas entre los puntos del espacio definido por los individuos. En realidad, las distancias son las medidas inversas de las similaridades, es decir, son disimilitudes. La distancia más conocida es la distancia euclidiana, aunque existen otras distancias. Entre estas se encuentran: la Distancia euclidiana al cuadrado, Distancia Mahalanobis y la Distancia City Block o Manhattan.

2. Coeficientes de asociación: se basan en algoritmos en los que se utilizan datos cualitativos. También se pueden aplicar a datos cuantitativos si se está dispuesto a sacrificar alguna información proporcionada por los individuos a las variables. Estas medidas son, básicamente, una forma de medir la concordancia o conformidad entre los estados de dos columnas de datos. Entre estos se encuentran; el Coeficiente de Jaccard, Coeficiente de Emperejamiento Simple, Coeficiente de Yule, etc.

3. Coeficientes de Correlación y otros coeficientes angulares: miden la proporcionalidad e interdependencia entre los vectores que definen los individuos. El más conocido es el coeficiente de correlación. Entre estos se encuentra: El Coeficiente de Correlación de Pearson y la Distancia Coseno.

4. Coeficiente de similaridad probabilística: incluyen información estadística y miden la homogeneidad del sistema por particiones o subparticiones del conjunto de los individuos. La idea de utilizar estos coeficientes se basa en relacionarlos con diferentes clasificaciones utilizando para ellas criterios de bondad de ajuste. Las principales propiedades de estos coeficientes es que son aditivos, se distribuyen como chi cuadrado X^2 y son probabilísticas. Esta última propiedad permite, en aquellos casos en que es posible, establecer una hipótesis nula y contrastarla por los métodos estadísticos tradicionales. Entre estos se encuentra la Medida del desorden o entropía

4.3.1. Distancias

La distancia es una descripción numerica de que tan lejos estan los objetos en cualquier momento del tiempo y las principales distancias empleadas como coeficientes de disimilaridad son las siguientes:

Distancia Euclidiana al cuadrado

Esta distancia se define entre dos individuos como la suma de los cuadrados de las diferencias de todas las coordenadas de los dos puntos, es decir, la distancia entre la r-ésima y la s-ésima fila de datos de la matriz X; será denotada por $(x_{r1}, x_{r2}, \ldots, x_{rp})$ y $(x_{s1}, x_{s2}, \ldots, x_{sp})$ respectivamente. Estas dos filas corresponden a las observaciones sobre dos objetos para todas las variables p. Geométricamente, la característica de estos objetos puede ser vista como las coordenadas de dos puntos en un espacio de dimensión p. Una medida conveniente de disimilaridad entre dos objetos r y s puede ser obtenida de la distancia Euclidiana entre dos puntos. Esta distancia es denotada por d_{rs}^2 donde:

$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2$$

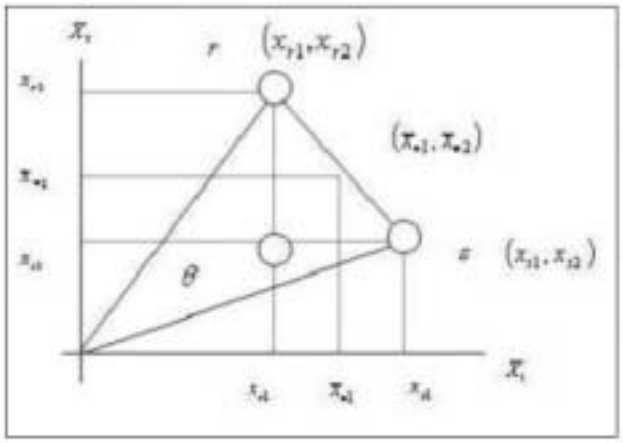


Figura 4.1: Representación en dos Dimensiones de la Proximidad entre dos Puntos

La medida d_{rs}^2 será referida como la distancia Euclidiana cuadrada. Para dos dimensiones ($p = 2$) la distancia entre dos objetos r y s puede ser representada como lo muestra la figura 4.1.

En este caso el cuadrado de la distancia Euclidiana esta dado por:

$$d^2 = (x_{r1} - x_{s1})^2 + (x_{r2} - x_{s2})^2$$

Sus medidas pueden estar relacionadas a los lados de un triángulo formado por los puntos r, s y t.Distancia de la media de las variables: la expresión de d_{rs}^2 no será afectada si se obtiene la media de las variables x_1 y x_2 cada una por separado. La media $\frac{(x_{r1}+x_{s1})}{2}$ será obtenida de ambos x_{r1} y x_{r2} similarmente la media $\frac{(x_{r2}+x_{s2})}{2}$ será obtenida de x_{r2} y x_{s2}. El resultado del valor de d_{rs}^2 seguirá sin cambios y este resultado se extiende al caso de p variables, de una madera similar.

Distancia Mahalanobis

una extensión del sistema de pesos que toma en cuenta las covarianzas entre las variables esta dada por la distancia Mahalanobis. Esta distancia es llamada una medida multivariada de distancia a partir de que toma en cuenta las estructuras de la covarianza entre las variables. Si las variables originales son transformadas primero en componentes principales antes de calcular la distancia Euclidiana, entonces la distancia Euclidiana basada sobre todos los componentes principales son equivalentes a las distancias Mahalanobis.

$$d_{rs} = \sqrt{(\bar{x}_r - \bar{x}_s)' \sum^{-1} (\bar{x}_r - \bar{x}_s)}$$

Distancia City Block o Manhattan

Distancia City Block o Manhattan: una alternativa de la distancia tipo métrica es la Manhattan o City Block, la cual esta basada en al valor absoluto de las diferencias entre las coordenadas. Esta métrica1[1] esta dada por.

[1]Métrica: también llamada dato cuantitativo, esta medida identifica o describe objetos no únicamente con la posesión del atributo pero también por la cantidad o grado al cual el tema puede ser caracterizado por el atributo.

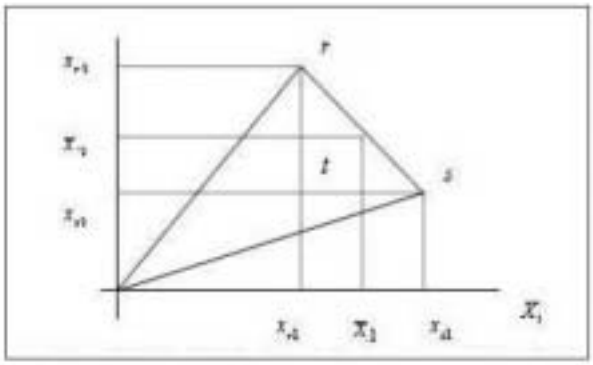


Figura 4.2: Representación de la Metrica City Block o Manhattan

$$d_{rs} = \sum_{j=1}^{p} \mid x_{rj} - x_{rs} \mid$$

Para el caso de dos dimensiones, la distancia d_{rs} representa la suma de las distancias derivadas de los puntos r a t y de los puntos derivados de s a t. Esto se puede apreciar en la figura 4.2. Con la métrica City Block, una diferencia constante entre cada una de las p coordenadas en la cantidad a tiene el mismo efecto sobre la distancia total como el cambio de diferencias entre un conjunto de coordenadas mediante la cantidad pa. Esto no es cierto para la distancia métrica Euclidiana donde las distancias en el segundo caso podrían ser más grandes que el primer caso. De aquí que la métrica City Block sea mucho menos sensitiva a los datos extremos (outliers).

Distancia (o Métrica) Minkowski

La distancia Euclidiana y la métrica City Block son casos especiales de la métrica Minkowski la cual esta dada por:

$$m_{rs} = \sqrt[\lambda]{\sum_{j=1}^{p} \mid x_{rj} - x_{rs} \mid^{\lambda}}$$

La distancia Euclidiana y la métrica City Block corresponden a $\lambda = 2$ y $\lambda = 1$ respectivamente. En general, entre más grande sea el valor de λ más grande será el énfasis dado a las diferencias en las coordenadas sobre una variable dada. En adición a las tres propiedades de una medida de disimilaridad, la métrica Minkowski también satisface las siguientes propiedades:

1. $p_{rs} = 0$ únicamente si $\bar{x}_r = \bar{x}_s$
2. $p_{rs} \leq p_{rm} + p_{ms}$ $\forall r, s, m$

El caso límite, para λ tendiendo al infinito, de la medida de Minkowski, es la *Distancia Chebychev* que es el máximo de las diferencias absolutas de los valores de todas las coordenadas.

Distancia Camberra

Es una modificación de la distancia Manhattan, definida por:

$$d_{sr} = \frac{\sum_{j=1}^{p} \mid x_{rj} - x_{sj} \mid}{(x_{rj} - x_{sj})}$$

La ventaja de esta distancia es que es una propiedad que se aplica a valores positivos, que depende únicamente de los individuos o grupos que están siendo comparados y no ésta afectada por el rango de sus valores. Además es sensible a proporciones y no sólo a valores absolutos.

Distancia Euclidiana en Forma Matricial

Dada una matriz de datos de orden $(n \times p)$ de una matriz X con $(1 \times p)$ vectores fila $(x_1', x_2', x_3', \ldots, x_n')$, el cuadrado de la distancia Euclidiana, d_{rs}^2, entre los objetos r y s puede ser escrita como:

$$d_{rs}^2 = (\bar{x}_r - \bar{x}_s)'(\bar{x}_r - \bar{x}_s)$$

$$r, s = 1, 2, \ldots, n$$

La matriz $(n \times n)$ de valores d_{rs}^2 es usualmente llamada matriz de la distancia Euclidiana cuadrada.

Si a las p variables de X se les calcula la media, la matriz de datos será denotada mediante $X*$. De aquí que como la distancia entre dos puntos no se afecta si se ignorará la media, la distancia Euclidiana cuadrada entre dos objetos r y s está dada por:

$$d_{rs}^2 = (\bar{x}_r^* - \bar{x}_s^*)'(\bar{x}_r^* - \bar{x}_s^*)$$

$$r, s = 1, 2, \ldots, n$$

donde $\bar{x}_r^*$ y $\bar{x}_s^*$ son las correspondiente filas de X^*.

Distancia Euclidiana Estandarizada

Una desventaja de la distancia Euclidiana como una medida de proximidad es su sensibilidad a las escalas de medida. Es posible, que una o más variables dominen la medida de distancia por la diferencia grande en escala. En general, si las escalas de medida no son comunes para todas las p variables, es preferible el uso de una distancia de pesos dada por $\sum_{j=1}^{p} w_j (x_{rj} - x_{sj})^2$ donde los pesos w_j reflejan la importancia de las variables $j = 1, 2, 3, \ldots, p$ Un caso especial de distancia Euclidiana es la distancia euclidiana estandarizada, que esta dada por:

$$d_{rs}^2 = \sum_{j=1}^{p} \frac{1}{s_j^2} (x_{rj} - x_{sj})^2 = (\bar{x}_r^* - \bar{x}_s^*)' D^{-1} (\bar{x}_r^* - \bar{x}_s^*)$$

donde s_j^2, $j = 1, 2, 3. \ldots, p$ denota la varianza de la variable X_j sobre los n objetos y D es la matriz diagonal son elementos sobre la diagonal dados por s_j^2 con $j = 1, 2, 3. \ldots, p$. Los vectores x_r y x_s de $(1 \times p)$ denotan las observaciones sobre las dos características.

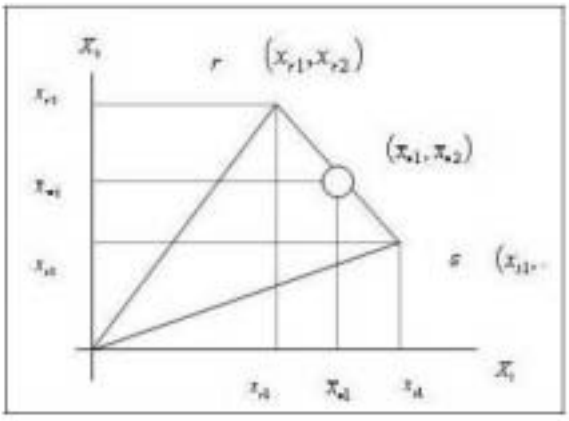


Figura 4.3: Representación del Centroide

La Distancia Euclidiana y el Centroide

La distancia Euclidiana entre dos características, d_{rs}^2, pueden ser relacionada al centroide entre dos objetos. Denotando la media sobre la variable j por medio de $\bar{x}_{\bullet j}$ con $\bar{x}_{\bullet j} = \frac{(x_{rj}+x_{sj})}{2}$, el centroide esta dado por $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_p)$. La distancia Euclidiana al cuadrado puede escribirse como:

$$d_{rs}^2 = 2[\sum_{j=1}^{p}(x_{rj} - \bar{x}_{\bullet j})^2 + \sum_{j=1}^{p}(x_{sj} - \bar{x}_{\bullet j})^2]$$

Cada una de las partes del lado derecho de la ecuación representa el cuadrado de una distancia entre una característica y el centroide de las dos características. La suma representa la suma del cuadrado de las desviaciones de las dos características derivadas de sus centroide. Tal que la variación de las dos características alrededor de sus centroide es proporcional al cuadrado de la distancia entre ellos. Esta variación entre las dos características puede también ser caracterizadas como la variación dentro del grupo formado por la unión de las dos características. Para el caso de dos variables el centroide es el punto medio de la línea que une a los puntos r y s, como se muestra en la figura 4.3.

Medidas de Distancia Medias sobre Variables

A veces es preferible dividir la distancia entre el número de variables. Por ejemplo la distancia Euclidiana cuadrada y la distancia Euclidiana son dadas en ocasiones como:

$$\frac{1}{p}\sum_{j=1}^{2}(x_{rj} - x_{sj})^2$$

y

$$\frac{1}{p^{1/2}}[\sum_{j=1}^{2}(x_{rj} - x_{sj})^2]^{1/2}$$

respectivamente. Esta modificación no afecta el resultado de la medida de proximidad, puesto que todas las medidas son divididas por la misma cantidad.

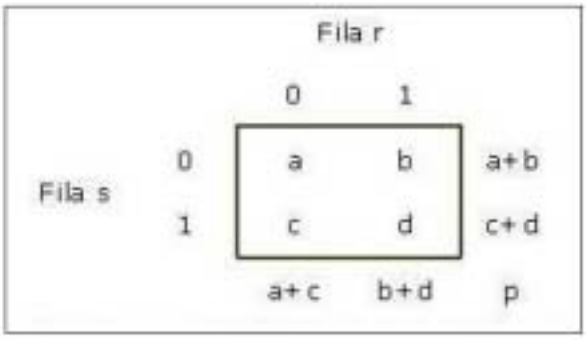


Figura 4.4: Relación entre dos Variables Binarias

4.3.2. Coeficientes de Asociación

En el caso especial de que las p categorías contengan únicamente dos categorías (binarias), un acercamiento alternativo puede ser usado como medida de proximidad entre condiciones. Cada variable, cada condición es codificada con 0 o 1 para indicar cual de los dos atributos esta presente o ausente una característica. Cuando los estados de las variables se comparan en parejas de columnas, en una matriz de datos, el resultado se suele resumir en una tabla de frecuencias tamaño (2×2). Una tabla (2×2) puede ser usada para describir la proximidad en términos de las cuatro posibles categorías para cada una de las variables p .

En la figura 4.4 a representa el número de variables en el cual ambas condiciones son codificadas 0 y d representa el número de variables en el cual ambas condiciones son codificadas 1, mientras que b y c representan el numero de variables con valor 1 y 0 para cada individuo. La suma representa el número de variables en la cual las condiciones tienen diferente codificación. El total número de variables observadas es $p = (a + b + c + d)$.

Dos usos comunes de medida de similaridad en este caso son el coeficiente de aproximación simple y el coeficiente Jaccard. Para el coeficiente de aproximación simple la medida de similaridad esta dada por $\frac{(a+b)}{(a+b+c+d)}$ la cual mide la proporción de variables en las cuales las características tienen la misma codificación.Los principales coeficientes de asociación son los siguientes:

Coeficiente de Jaccard

Para el coeficiente Jaccard la medida de similaridad esta dada por $\frac{d}{(b+c+d)}$. En este caso él número de variables en la cual ambas condiciones son codificadas cero ha sido omitido. Si el propósito de la medida de similaridad es indicar que tan similares son las condiciones con respecto a los atributos presentes (codificados 1) y el ignorar el impacto de los atributos ausentes (codificados 0), entonces el coeficiente Jaccard es el más apropiado. Mediante la exclusión de variables en el cual ninguna característica tiene el atributo, la similaridad es únicamente medida con respecto de un atributo en común. Si dos características son faltantes en un gran número de atributos puede que no se desee decir que son similares. Por ejemplo en taxonomía numérica el coeficiente de Jaccard es preferido.

Coeficiente de emparejamiento simple

Se define como el cociente entre el número de emparejamientos y el número total de casos considerados:

$$s_{rs} = \frac{a+b}{a+b+c+d}$$

Coeficiente de Yule

Se define por

$$s_{rs} = \frac{ad-bc}{ad+dc}$$

y varía entre -1 y +1.

Si la variable dummy de la matriz X es usada para variables binarias (2 columnas para cada variable X) como en el caso de datos categóricos, las cantidades K, p y f_{rs} son relacionadas a a, b, c y d mediante, $p = (a+b+c+d)$, $K = 2p$ y $f_{rs} = (a+b)$. La medida de proximidad introducida para datos categóricos esta ahora dada por $c_{rs} = \frac{(a+b)}{(a+b+c+d)}$, $q_{rs} = \frac{(a+d)-(b+c)}{(a+b+c+d)}$ y $d^2_{rs} = (b+c)$. La medida c_{rs} es por lo tanto equivalente al coeficiente de correlación simple y d^2_{rs} es obtenido mediante la substracción de 1 del coeficiente de correlación. El coeficiente de correlación q_{rs} en este caso es llamado *coeficiente Hamann.*

Si en el caso de variables binarias una matriz X esta formado por el uso del código 0-1 (1 columna por variable), un 0 aparece en una columna donde las características no tienen el atributo y aparece un 1 donde el atributo se presenta. En este caso la matriz XX' admite el coseno y la medida de correlación esta dada por:

$$c_{rs} = \frac{d}{[(d+b)(c+d)]^{1/2}}$$

y

$$q_{rs} = \frac{(ad-dc)}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$$

El coeficiente es usualmente llamado *coeficiente Ochiai* y q_{rs} es referido como el coeficiente φ. En cualquiera de los casos el coeficiente cae en el rango requerido que satisface la medida de similaridad [22] (Jobson 1992).

4.3.3. Coeficientes de Correlación y otros Coeficientes Angulares

Una sugerencia alternativa, a la medida de proximidad entre dos puntos r y s en un espacio de dimensión, es el uso del ángulo entre los dos vectores $(1 \times p)$ de observaciones x_r y x_s. Las características de los objetos r y s son mostrados como puntos en una espacio de dos dimensiones. Los dos puntos pueden ser vistos como el vértice formado entre los vectores que tienen su origen en el mismo punto con un ángulo θ entre dos vectores figura 4.5. Una medida de similaridad útil es el coseno del ángulo θ. En general el coseno del ángulo entre dos vectores x_r y x_s esta dado por:

$$c_{rs} = \frac{\sum_{j=1}^{p} x_{rj}x_{sj}}{\sqrt{\sum_{j=1}^{p} x^2_{rj} \sum_{j=1}^{p} x^2_{sj}}}$$

Como podemos ver en la figura 4.5, que c_{rs} no depende de la longitud de los dos vectores, y consecuentemente los cambios proporcionados en las coordenadas

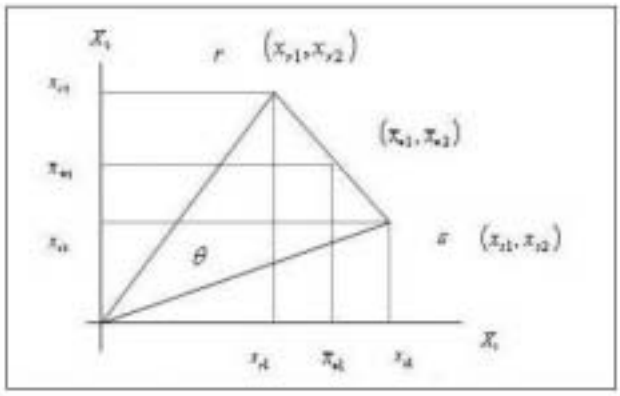


Figura 4.5: Relación del Angulo entre dos Vectores

de x_r y/o en x_s no cambiará. Las cantidades $\sum_{j=1}^{p} x_{rj}^2$ y $\sum_{j=1}^{p} x_{sj}^2$ son los cuadrados de las distancias de los vectores. El coeficiente coseno es también referido en algunas ocasiones como coeficiente congruencia. Las características x_r y x_s pueden ser usadas en forma media central para forzar $(\bar{x}_r - \bar{x}_r e)$ y $(\bar{x}_s - \bar{x}_s e)$ en donde e es un vector unitario de $(1 \times p)$:

$$\bar{x}_r = \frac{\sum_{j+1}}{p}$$

así como

$$\bar{x}_r = \frac{\sum_{j=1}^{p} x_{sj}}{p}$$

son las medias para las características x_r y x_s respectivamente.

Coeficiente de Correlación de Pearson

El coseno del ángulo entre los vectores media central es equivalente a la correlación de Pearson entre dos vectores x_r y x_s. La medida de similaridad, coeficiente de correlación resultante esta dado por:

$$q_{rs} = \frac{(\bar{x}_r - \bar{x}_r e)'(\bar{x}_s - \bar{x}_s e)}{(\sqrt{(\bar{x}_r - \bar{x}_r e)'})(\bar{x}_r - \bar{x}_r e)(\bar{x}_s - \bar{x}_s e)'(\bar{x}_s - \bar{x}_s e)'}$$
$$= \frac{\sum_{j=1}^{p}(x_{rj} - \bar{x}_r)(x_{sj} - \bar{x}_s)}{\sqrt{\sum_{j=1}^{p}(x_{rj} - \bar{x}_r)^2 \sum_{j=1}^{p}(x_{sj} - \bar{x}_s)^2}}$$

Las medidas $c_r s$ y q_{rs} son llamadas medidas de similaridad. Una desventaja de la medida de similaridad es que es sensible a la dirección de la escala de medida para cada una de las variables. Si algunos de los objetos son medidos en dirección positiva y otros son medidos en dirección negativa, el tipo de medida positivo o negativo afectará el significado en general de la característica del valor de q_{rs}. En cuyos casos algunas escalas deberían ser convertidas y así las direcciones serán las mismas.

Distancia del Coseno

Viene definida por el coseno del ángulo subtendido por los individuos i y j con el origen de coordenadas, y se expresa de la siguiente manera:

$$\cos \alpha_{ij} = \frac{\sum_k x_{ik} x_{jk}}{\sqrt{\sum_k (x_{ik})^2 \sum_k (x_{ij})^2}}$$

El rango de esta medida está comprendido, claramente, entre -1 y $+1$. El valor $+1$ se obtiene para una perfecta similitud si los valores son del mismo signo y -1 son de signo contrario, mientras que los valores próximos a 0 se obtienen para individuos disimilares.

4.3.4. Coeficientes de Similaridad Probabilística

Los coeficientes de similitud probabilística calculan la probabilidad acumulada de que un par de individuos i y j, sean tan similares, o más, que lo que empíricamente se puede afirmar sobre la base de la distribución observada.Se define, en primer lugar, la *medida del desorden* para la variable k como

$$H(k) = - \sum_{g=1,\ldots,m(k)} p_{kg} \ln p_{kg}$$

siendo $m(k)$ el número de posibles estados diferentes para la variable k, donde p_{kg} es la proporción observada del individuo t que muestra el estado g para la variable k. Por tanto, se tiene que

$$\sum_{g=1,\ldots,m(k)} p_{kg} = 1$$

Si las n variables no están correlacionadas, se pueden sumar los valores separados de los $H(k)$ para obtener la información total del grupo:

$$I = t \sum_{g=1,\ldots,m(k)} H(k)$$

En ocasiones, la frecuencia de los datos se presenta con sólo dos estados (presencia o ausencia de una característica), en estas situaciones, se definen los valores anteriores como sigue:

$$H(k) = -(p_k \ln p_k + (1 - p_i) \ln(1 - p_i))$$

$$I = -t \sum_{g=1,\ldots,m(k)} (p_k \ln p_k + (1 - p_i) \ln(1 - p_i))$$

4.3.5. Combinación de Variables Categóricas y Variables de Escala Intervalar

Si la matriz de datos X contiene una combinación de variables dummy y variables de escala intervalar es difícil combinar las variables para determinar la medida de similaridad. Un acercamiento podría ser el de estandarizar las variables o las columnas de X antes de calcular la matriz XX'.

Una segunda alternativa podría ser la de calcular separadamente las medidas de proximidad para los datos categóricos y variables de tipo intervalar y después combinar las dos medidas de proximidad usando pesos apropiados. Una tercera alternativa podría ser convertir las variables tipo intervalar en variables tipo categórico mediante la construcción de clases y después tratar a todas las variables como categóricas.

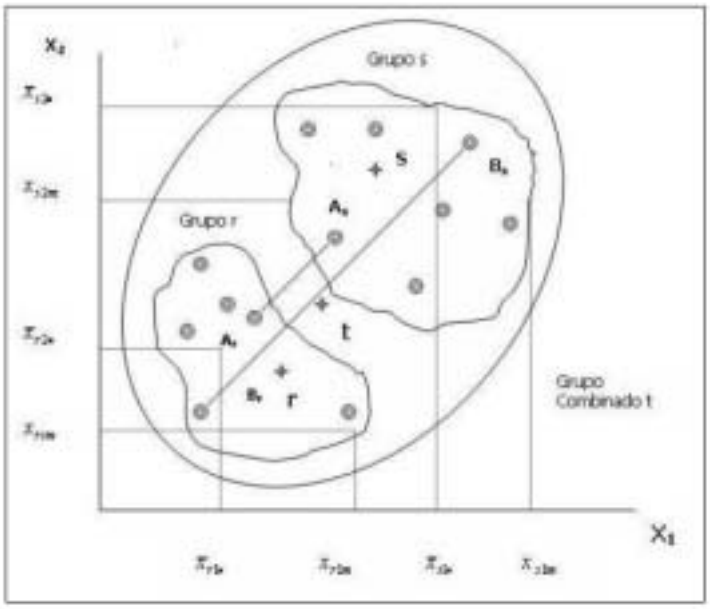


Figura 4.6: Medidas de Proximidad entre Grupos

4.4. Medidas de Proximidad entre Grupos

Anteriormente fueron presentadas técnicas para la medida de aproximación entre objetos o filas de una matriz de datos. Ahora trataremos una variedad de métodos para la medida de proximidad entre dos grupos de objetos. El propósito de estudiar las medidas de proximidad entre grupos será demostrado cuando se indiquen las principales características o las diferentes partes del Análisis de Conglomerados.

Asumiendo que dos grupos de objetos denotados por r y s contienen n_s y n_r objetos respectivamente, las observaciones sobre las p variables para los n_r objetos en un grupo r están denotadas por x_{rjm} ,$j = 1, 2, \ldots, p$, $m = 1, 2, \ldots, n_r$ y similarmente para el grupo de s observaciones esta denotada por x_{smj}, $j = 1, 2, \ldots, p$, $m = 1, 2, \ldots, n_s$. La figura 4.6 ilustra la notación para $p = 2$ variables.

En la figura 4.6 el grupo r contiene $n_r = 6$ objetos y el grupo s contiene $n_s = 7$ objetos. El m-ésimo objeto del grupo r tiene coordenadas (x_{r1m}, x_{r2m}) mientras que el m-ésimo objeto del grupo s tiene coordenadas (x_{s1m}, x_{s2m}). La función $p_{rs}(j, k)$ es usada para denotar la medida de proximidad entre el j-ésimo objeto de grupo r y el k-ésimo objeto del grupo s.

4.4.1. Vecino más cercano

El método vinculación simple (single linkage) o el vecino más cercano (Nearest Neighbor) es una medida de proximidad entre dos grupos, esta basada en la medida de proximidad más fuerte entre objetos de dos grupos. Por esta razón, y aunque pueden haber muchos objetos involucrados, la medida de proximidad usada está basada en un par de objetos.

Si a una medida de disimilaridad como lo es la distancia Euclidiana está siendo usada como medida de proximidad, entonces el método single linkage usara la mínima distancia Euclidiana posible entre objetos de dos grupos. Para medidas de similaridad como lo es el coeficiente de correlación, el método de medida single linkage estará basado en la máxima correlación posible entre objetos de

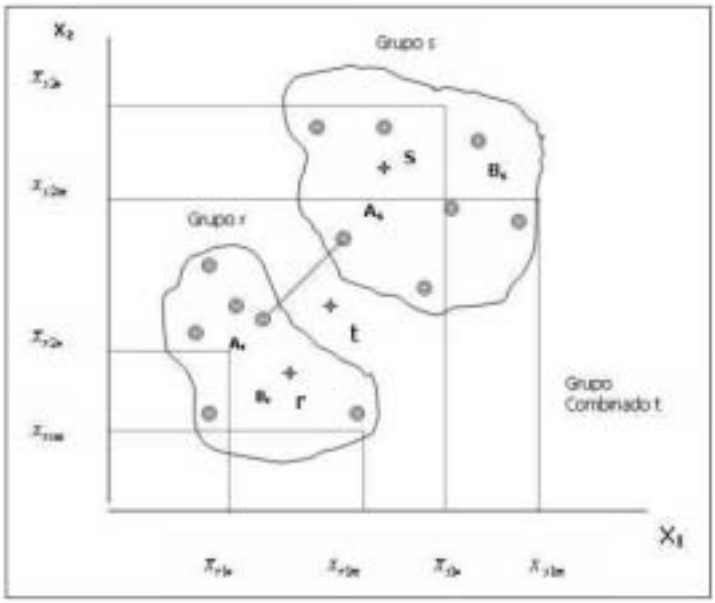


Figura 4.7: La distancia entre Ar y As representa la menor distancia entre los puntos de los dos grupos

dos grupos. En la figura 4.7 las distancia entre A_r y A_s representa la distancia Euclidiana mínima entre puntos de dos grupos.

4.4.2. Vecino más lejano

La vinculación completa (Complete linkage) o el vecino más lejano (Furthest neighbor) es una medida de proximidad entre dos grupos se deriva de la unión más débil entre objetos de dos grupos. El método de medida complete linkage es por lo tanto lo contrario al método de medida single linkage. Para una medida de disimilaridad como la distancia Euclidiana, la distancia más grande posible entre objetos de dos grupos es usada para representar la proximidad entre grupos. Para una medida de similaridad como lo es el coeficiente correlación el valor más pequeño posible sobre todos los pares posibles es usado como una medida de proximidad. La distancia entre B_r y B_s en la figura 4.8 es la distancia más grande entre objetos de dos grupos.

4.4.3. Vinculación de medias

Como una alternativa del uso de la medida de proximidad basada en un solo par de objetos posibles, la media puede ser determinada sobre todos los pares de objetos. Si n_r y n_s son los objetos en los grupos r y s respectivamente, entonces hay un total de $(n_r)(n_s)$ pares de medidas de proximidad. La medida vinculación de medidas (average linkage) esta dada por la media de las medidas $(n_r)(n_s)$ dada por: $\frac{\sum_{j=1}^{n_r}\sum_{k=1}^{n_s} p_{rs}(j,k)}{n_r n_s}$. La medida de proximidad vinculación de medias es a veces referida como UPGMA por sus siglas en ingles *unweighted pair group method using averages*.

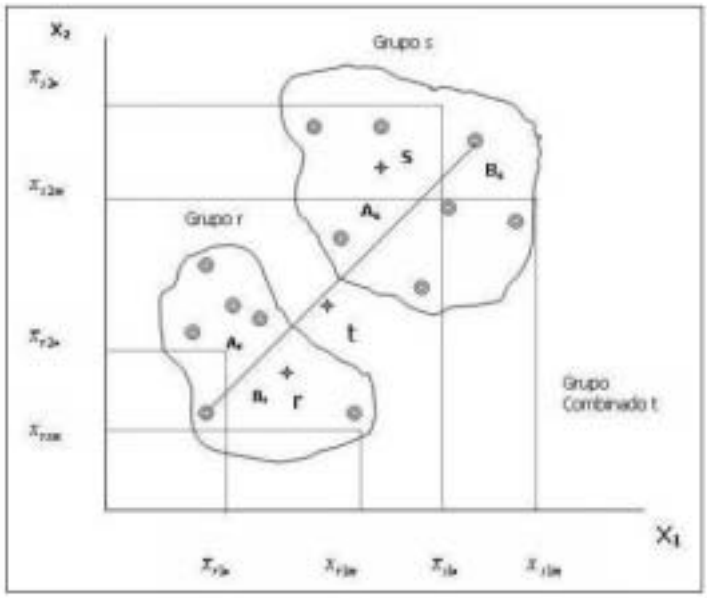


Figura 4.8: La distancia más grande entre objetos de dos grupos Br y Bs

4.4.4. Distancia entre Centroides

La distancia Euclidiana entre grupos de centroides puede ser usada para medir la proximidad entre dos grupos de objetos. Si las coordenadas para los centroides en los grupos r y s están dadas por $(\bar{x}_{r1}, \bar{x}_{r2}, \ldots, \bar{x}_{rp})$ y $(\bar{x}_{s1}, \bar{x}_{s2}, \ldots, \bar{x}_{sp})$, el cuadrado de la distancia Euclidiana entre dos centroides esta dada por:

$$d_{rs}^2 = \sum_{j=1}^{p} (\bar{x}_{rs\bullet} - \bar{x}_{sj\bullet})^2$$

Esta medida esta referida como la distancia entre centroides. Para en caso de dos variables en la figura 4.9 representa los centroides son denotados por $*r$ y $*s$. La distancia Euclidiana entre los dos puntos es la medida de proximidad requerida entre los grupos r y s .

4.4.5. Método Ward o Suma de Cuadrados Aumentada

Una alternativa de medida de proximidad basada en la distancia Euclidiana entre centroides usa el hecho de que son un total de $(n_r)(n_s)$ distancias entre dos grupos. Una medida de la distancia total entre los dos grupos esta dada por $n_r n_s d_{rs}^2$. Entonces hay un total de $(n_r + n_s)$ objetos, la media esta dada por $\frac{n_r n_s d_{rs}^2}{(n_r+n_s)}$. Esta medida de distancia media es equivalente al cambio de la suma de cuadrados dentro del grupo o *la suma de cuadrados aumentada*, resultado de la combinación de grupos r y s. Este concepto es ilustrado a continuación. Para el grupo - ésimo grupo, la suma de cuadrados dentro del grupo esta dada por

$$SSW_r = \sum_{m=1}^{n_r} \sum_{j=1}^{p} (x_{rmj} - \bar{x}_{rj\bullet})^2$$

De manera similar para el s-ésimo grupo, la suma de cuadrado dentro del grupo esta dado por

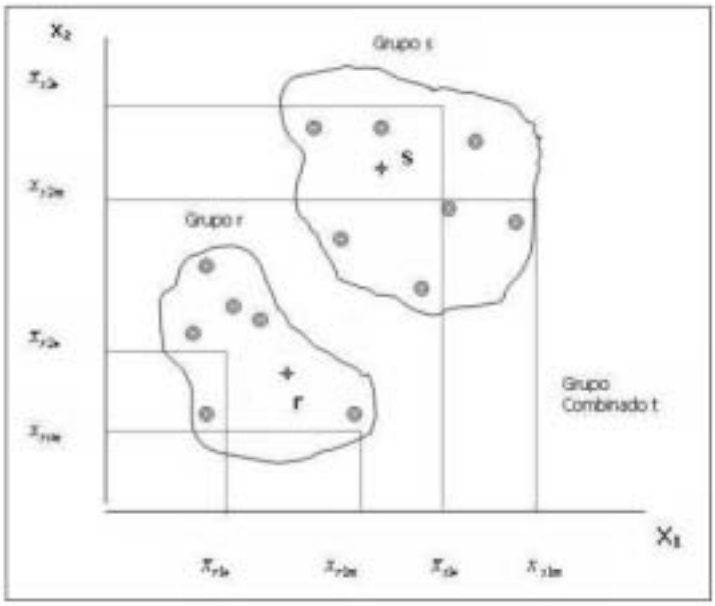


Figura 4.9: Representación de centroides r y s

$$SSW_s = \sum_{m=1}^{n_r} \sum_{j=1}^{p} (x_{smj} - \bar{x}_{sj\bullet})^2$$

Si los grupos r y s son combinados para formar un nuevo grupo t, un nuevo centroide $(\bar{x}_{t1\bullet}, \bar{x}_{t2\bullet}, \ldots, \bar{x}_{tp\bullet})$ es obtenido y el nueva suma de cuadrados dentro del grupo t esta dada por:

$$SSW_t = \sum_{m=1}^{n_r+n_s} \sum_{j=1}^{p} (x_{tmj} - \bar{x}_{tj\bullet})^2$$

El aumento en el total de la suma de cuadrados aumentada dentro del grupo considerando como un resultado de grupos conjuntos r y s esta dado por $SSW_t - (SSW_r + SSW_s)$. Este aumento en el total de la suma de cuadrados es equivalente a la distancia total $\frac{n_r n_s d_{rs}^2}{(n_r+n_s)}$. Este aumento a la suma de cuadrados es comúnmente usado como una medida de proximidad entre dos grupos. Hay que recordar que para $n_r = n_s = 1$ la distancia Euclidiana al cuadrado ente los puntos y el centroide esta relacionado a la distancia euclidiana ente los puntos y que el centroide para el grupo t esta dado por $*t$ el cual cae a través de la línea que une $*r$ y $*s$.

4.4.6. Un algoritmo para Modificar la Medida de Proximidad

Cuando los objetos o grupos de objetos son combinados para formar un nuevo grupo, es necesario poder revisar o actualizar las medidas de proximidad entre los nuevos grupos y para los conjuntos de grupos de datos restantes. Un algoritmo útil puede ser usado para revisa la matriz de medidas de proximidad, que esta dado por:

$$p_{tu} = \alpha_r p_{ru} + \alpha_s p_{su} + \beta p_{rs} + \gamma \mid p_{rs} + p_{su} \mid$$

donde

1. t es la nueva referencia para el grupo resultante de la combinación de grupos r y s.

2. u es la referencia para algún otro grupo, otros de r y s.

3. $\alpha_r, \alpha_s, \beta$ y γ son coeficientes que depende de las medidas de proximidad que son usadas.

Asumiendo que la medida de proximidad que ha sido usada es una medida de disimilaridad, el método *complete linkage* emplea,$\beta = 0$ y $\gamma = 1/2$ mientras que para el método *single linkage* los valores son, $\beta = 0$ y $\gamma = -1/2$. Si la medida de proximidad es una medida de similaridad los valores para γ son intercambiados para los métodos *single linkage* y *complete linkage*. Para el método *average linkage*, $\alpha_s = \frac{n_s}{(n_r+n_s)}$, $\alpha_r = \frac{n_r}{(n_r+n_s)}$, $\beta = 0$ y $\gamma = 0$.

Usar la relación $p_{tu} = \alpha_r p_{ru} + \alpha_s p_{su} + \beta p_{rs} + \gamma \mid p_{rs} + p_{su} \mid$ permite a la matriz de proximidad el actualizar grupos posteriores r y s sean combinados para formar el nuevo grupo t. Es importante notar que la matriz de proximidad actual no requiere la matriz de datos originales. La matriz de proximidad original entre los objetos es el único enlace requerido de los datos originales.

4.4.7. Relación del Análisis de la Varianza

El total de la suma de cuadrados dentro del grupo para el grupo combinado t puede ser visto como el total de la suma de cuadrados como en la ANOVA. Los subgrupos r y s admiten una suma de cuadrados dentro del grupo y un total de la suma de cuadrados entre grupos. Así como en el análisis de la varianza, la suma total de cuadrados contenidos en los dos componentes, la suma de cuadrados dentro del grupo y las sumas de cuadrados entre ellos. La suma de cuadrados total es equivalente a SSW_t y la suma de cuadrados dentro del grupo es $(SSW_t + SSW_s)$. La diferencia entre los dos es el cuadrado de la suma entre ellos y esta dado por:

$$SSG_t = \sum_{j=1}^{p} [n_r(\bar{x}_{rj\bullet} - \bar{x}_{tj\bullet})^2 + n_s(\bar{x}_{rj\bullet} - \bar{x}_{tj\bullet})^2]$$

$$= \frac{n_r n_s}{(n_r + n_s)} \sum_{j=1}^{p} (\bar{x}_{rj\bullet} - \bar{x}_{sj\bullet})^2 = \frac{n_r n_s}{(n_r + n_s)} d^2_{rs\bullet}$$

Esta suma de cuadrado entre ellos es la suma de cuadrados incrementada usado para medir la proximidad entre grupos r y s.

4.4.8. Algoritmo para Determinar la Distancia entre Centroides y el Método Ward

Para el método del centroide (también referido como UPGMC el cual es la abreviación de la expresión "unweighted pair group method using centroids") y el método Ward (o método de la suma de cuadrados aumentada) la medida de proximidad puede ser obtenida en una forma secuencial usando el algoritmo mostrado abajo. La secuencia involucra una serie de paso en los cuales una medida de proximidad es determinada entre dos objetos o grupos de objetos en cada paso. El proceso empieza con una matriz de proximidad basada en la distancia euclidiana cuadrada entre los objetos originales. A un punto dado en la secuencia, la medida de proximidad entre los grupos r, s y u son denotados por

p_{rs}, p_{ru} y p_{ss}. Los grupos r y s son combinados para formar un nuevo grupo denotado por t, y una nueva medida de proximidad p_{tu} son necesarios para relacionar t y u. Para el método de suma de cuadrados aumentada, el proceso empieza con la medida de proximidad dada por $p_{rs} = \frac{1}{2}d_{rs}^2$ donde $2d_{rs}^2$ denota el cuadrado de la distancia euclidiana entre los objetos r y s. Después de combinar objetos r y s para formar el nuevo grupo t, la medida de proximidad suma de cuadrados aumentada entre t y u esta dada por:

$$p_{tu} = \frac{1}{(n_t + n_s)}[(n_u + n_r)p_{ru} + (n_u + n_s)p_{su} - n_u p_{rs}]$$

en donde $n_t = (n_r + n_s)$.Para el método del centroide el proceso empieza son la medida de proximidad dada por $p_{rs} = d_{rs}^2$. Después de combinar objetos r y s para formar el grupo t la medida de proximidad entre t y u esta dado por:

$$p_{tu} = \frac{n_r}{(n_t + n_s)}p_{ru} + \frac{n_s}{(n_r + n_s)}p_{su} - \frac{n_r n_s}{(n_r + n_s)^2}p_{rs}$$

Los dos algoritmos anteriores con casos especiales de:

$$p_{tu} = \alpha_r p_{ru} + \alpha_s p_{su} + \beta p_{rs} + \gamma \mid p_{rs} + p_{su} \mid$$

Para el método Ward,$\alpha_r = \frac{(n_r+n_u)}{(n_t+n_u)}$, $\alpha_s = \frac{(n_s+n_u)}{(n_t+n_u)}$, $\beta = \frac{-n_u}{(n_t+n_u)}$ y $\gamma = 0$. Para el método de centroide $\alpha_r = \frac{n_r}{(n_r+n_s)}$, $\alpha_s = \frac{n_s}{(n_r+n_s)}$, $\beta = \frac{-n_r n_s}{(n_r+n_s)^2}$ y $\gamma = 0$. Aunque los dos algoritmos pueden ser vistos como casos especiales del algoritmo anterior es importante tener en mente que la matriz de proximidad original se asuma como la distancia euclidiana cuadrada.

4.4.9. Desigualdad Ultramétrica

El algoritmo general $p_{tu} = \frac{n_r}{(n_t+n_s)}p_{ru} + \frac{n_s}{(n_r+n_s)}p_{su} - \frac{n_r n_s}{(n_r+n_s)^2}p_{rs}$ provee un método secuencial para medidas de proximidad actualizadas entre grupos como grupos expandidos en tamaño y de aquí que se vuelvan menos numerosos. Usualmente la medida de proximidad es tal que la distancia entre dos grupos que han sido combinados (grupos y) es pequeña en comparación con la distancia entre el nuevo grupo combinado t y cualquier otro grupo u. Si esto es verdad entonces la medida de proximidad distancia-tipo satisface la desigualdad ultrametrica. Esta desigualdad implica que la no medida de distancia en la nueva matriz puede ser más pequeña que la medida de distancia más pequeña en la matriz anterior. El propósito de identificar esta propiedad es que esta garantiza que considerando grupos combinados la medida de proximidad distancia-tipo será monótona no decreciente.

Si los parámetros de $p_{tu} = \frac{n_r}{(n_t+n_s)}p_{ru} + \frac{n_s}{(n_r+n_s)}p_{su} - \frac{n_r n_s}{(n_r+n_s)^2}p_{rs}$ satisfacen las condiciones $(\alpha_r + \alpha_s + \beta) \geq 1$ y $\gamma \geq max(-\alpha_r, -\alpha_s)$, entonces la medida de proximidad tendrá la propiedad ultrametrica. Para las medidas discutidas, únicamente el método de centroide no satisface esta propiedad. Para el método de centroide $(\alpha_r + \alpha_s + \beta) = \frac{(n_r n_s)}{(n_r+n_s)^2}$ que no es $\ngeq 1$. Por esta razón el método de centroide es usado raro vez.

4.4.10. Método Ward Derivado de las Matrices MANOVA

Dados g grupos de objetos con tamaños de grupos $n_1, n_2, \ldots, n_g$; en donde cada objeto es medido sobre una dimensión p la variable x $(p \times 1)$. La MANOVA puede ser usada para caracterizar diferencias entre los g grupos. En la MANOVA tres sumas de cuadrados y matrices producto punto se usan para medir la variación y las denotaremos como T, W y G. Las tres matrices fueron

nombradas la suma de cuadrados total, la suma de cuadrados dentro y entre la suma de cuadrados respectivamente y satisface la relación $T = W + G$.Las matrices T, W y G son matrices con j-ésimos elemento diagonales dados por

$$t_{jj} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{\bullet j \bullet})^2$$

$$w_{jj} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{\bullet jk})^2$$

y

$$g_{jj} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\bar{x}_{\bullet jk} - \bar{x}_{\bullet j \bullet})^2 = \sum_{k=1}^{g} n_k (\bar{x}_{\bullet jk} - \bar{x}_{\bullet j \bullet})$$

respectivamente.

Para cada uno de estos tres elementos, el primer o la sumatoria interna representa la suma de cuadrados sobre las observaciones en un grupo para la variable j, mientras que la suma externa determina la suma sobre todos los grupos. Cada elemento de la diagonal consecuentemente representa respectivamente la suma de cuadrados total, la suma de cuadrados de grupos incluidos y la suma de cuadrados entre los grupos g sobre todos los grupos para una variable.

Para determinar las tres sumas de cuadrados sobre todas las p variables las sumas de los elementos de la diagonal de estas matrices son requeridos. Las tres sumas de cuadrados están dados por:

$$trT = \sum_{j=1}^{p} t_{jj} = \sum_{j=1}^{p} \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{\bullet j \bullet})^2$$

$$trW = \sum_{j=1}^{p} w_{jj} = \sum_{j=1}^{p} \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{\bullet jk})^2$$

$$trG = \sum_{j=1}^{p} g_{jj} = \sum_{j=1}^{p} \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{\bullet jk} - \bar{x}_{\bullet j \bullet})^2$$

$$= \sum_{j=1}^{p} \sum_{k=1}^{g} n_k (\bar{x}_{\bullet jk} - \bar{x}_{\bullet j \bullet})^2$$

Además mediante el reordenamiento de la secuencia de los signos de las sumatorias estas cantidades puede ser vistas como sumas sobre los grupos de suma de cuadrados determinadas sobre p variables. Denotando la suma de cuadrados para grupos k mediante SSW_k, SST_k, y SSG_k tenemos $trT = \sum_{k=1}^{g} SST_k$, $trW = \sum_{k=1}^{g} SSW_k$ y $trG = \sum_{k=1}^{g} SSG_k$ donde:

$$SSW_K = \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{\bullet jk})^2$$

$$SSG_K = \sum_{i=1}^{n_k} n_k (\bar{x}_{\bullet jk} - \bar{x}_{\bullet j \bullet})^2$$

y

$$SST_K = \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{\bullet j \bullet})^2$$

Sabemos que la medida de proximidad entre grupos usando el método Ward muestra la relación entre las tres sumas de cuadrados. Así como que cuando los grupos r y s se conjuntan para formar un nuevo grupo t que:

$$SSW_t = \frac{n_r n_s}{(n_r + n_s)} \sum_{j=1}^{p} (\bar{x}_{\bullet jr} - \bar{x}_{\bullet js})^2 + SSW_r + SSW_t$$

De aquí que $SST_t = SST_r + SST_s$ continua definida, el efecto sobre la suma de cuadrados del grupo restante es:

$$SSG_t = SSG_r + SSG_s - \frac{n_r n_s}{(n_r + n_s)} \sum_{j=1}^{p} (\bar{x}_{\bullet jr} - \bar{x}_{\bullet js})^2$$

Para la j-ésima los elementos de la diagonal de las matrices W y G elementos w_{jj} incrementados por $\frac{n_r n_s (\bar{x}_{\bullet jr})^2}{(n_r+n_s)}$ mientras que los elementos g_{jj} se decrementan en esta misma cantidad. El efecto acumulado sobre trW y trG de los grupos combinados r y s es por lo tanto un incremento en trW de $\frac{n_r n_s \sum_{j=1}^{p} (\bar{x}_{\bullet jr})^2}{(n_r+n_s)}$ y un correspondiente decremento en trG es esta misma cantidad.

4.4.11. Doble Media Central

Si a las observaciones primero que nada se les obtiene la media por variables y después la media por filas, las observaciones resultantes estarán dadas por $(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})$ donde

$$\bar{x}_{i\bullet} = \frac{\sum_{j=1}^{p} x_{ij}}{p}$$

$$\bar{x}_{\bullet j} = \frac{\sum_{j=1}^{p} x_{ij}}{n}$$

y

$$\bar{x}_{\bullet\bullet} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}}{np}$$

En este caso los elementos de la matriz de datos modificada no tendrá efecto en las filas y en las columnas a esto se le llama doble mean-centered. Esta transformación será usada en escalas multidimensionales.

4.5. Comentarios

En este capítulo se presentaron medidas de proximidad, que permiten llevar acabo clasificaciones. Las medidas de proximidad, ya sean similaridades o disimilaridades, separan objetos en diferentes clases o grupos. Aunque existe un número amplio de medidas para lograr una clasificación de objetos, se definieron las de mayor importancia de acuerdo a varios autores, éstas seran usadas en el Análisis de Escalas Multidimensionales. También se analizaron métodos para separar dos grupos de objetos. El propósito de estudiar las medidas de proximidad entre grupos será demostrado cuando se indiquen las principales características o las diferentes partes del Análisis de Conglomerados. Una síntesis de las medidas de proximidad, que permiten clasificar objetos 4.4 y grupos 4.5, a manera de síntesis. Éstas son las que se encuentran presentes en el paquete estadístico SPSS.

Objetivo	Proximidad	Medida
Medidas de Proximidad entre Objetos	Disimilaridades	Distancia Euclidiana Distancia Euclidiana al cuadrado Chebychev City Block o Manhatan Minkowski Chi-cuadrada Phi-cuadrada
	Similaridades	Correlación de Pearson Coseno Russel y Rao Jaccard Dice Rogers y Tanimoto Sokal y Sneath Kulczynski Hamann Lambda Andenberg Yules Ochiai Phi

Cuadro 4.4: Sintesis de Medidas de Proximidad entre Objetos

Objetivo	Medida
Medidas de Proximidad entre Grupos	Single linkage o Nearest Neighbor (Vinculación Simple) Complete linkage o Furthest Neighbor (Vinculación Completa) Average linkage (Vinculación Media) Distancia entre centroides Método Ward o Suma de cuadrados aumentada Between Groups linkage Within Groups linkage

Cuadro 4.5: Sintesis de Medidas de Proximidad entre Grupos

Capítulo 5

Análisis de conglomerados

El objetivo del *Análisis de Conglomerados* es la partición de un conjunto de objetos en grupos o conglomerados de tal manera que las características de los objetos en el mismo conglomerado son similares, mientras que las características de objetos en diferentes conglomerados son considerablemente distintas. De aquí que el conglomerado esta ligado al concepto de proximidad entre objetos y grupos de objetos, las técnicas de proximidad discutidas en el Capítulo 4 juegan un importante papel en la identificación de conglomerados.

En algunas aplicaciones de análisis de conglomerados los objetos se consideran pertenecientes a un grupo natural pequeño, mientras que en otros casos el reto es simplemente encontrar un agrupamiento conveniente. El primer caso es a veces llamado clasificación, la otra aplicación es referida como disección. Otros términos comúnmente usados para el procedimiento de análisis de conglomerado tipo incluyen patrones de reconocimiento y taxonomía numérica. El desarrollo de técnicas de conglomerados y la aplicación de tales técnicas son recurrentes en diferentes campos de estudio, ingeniería, zoología, medicina, lingüística, antropología, psicología y mercadotecnia son algunos de los campos de aplicación.

Así como el análisis de componentes principales, el análisis de conglomerados puede ser visto como una técnica de reducción de datos. En lugar de reducir el número de variables o columnas requerido para caracterizar X como en el análisis de componentes principales, análisis de conglomerados reduce el número de objetos distintos o filas de X mediante la creación de grupos de objetos llamados conglomerados. Las técnicas del Análisis de Conglomerados puede ser clasificada en diferentes tipos; algunos de ellos son: jerárquico, no-jerárquico (k-medias), densidad, Q-sort y aglomerativos o clumping entre otros.

Mientras que la estructura del conglomerado puede variar dependiendo de las diferentes aplicaciones de investigación, para la mayoría de las técnicas resumidas aquí, se asume que el reto es encontrar un conglomerado natural. El conglomerado natural asume que satisface las propiedades de aislamiento externo y cohesión interna. El aislamiento externa sugiere que puntos que se encuentran en un conglomerado deben estar separados de otros puntos que se encuentran en otros conglomerados por medio de espacio vacíos. Cohesión interna requiere que puntos dentro de un conglomerado deben estar cercanamente juntos. Esta caracterización de conglomerado natural por lo tanto no permite la superposición de conglomerados.

Se enfatizará en el método jerárquico y el no-jerárquico en este capítulo. Una diversidad de algoritmos para realizar un análisis de conglomerados jerárquico será resumida. Así como técnicas para evaluación de la solución de conglomerados.

5.1. Análisis de Conglomerados

El termino "Análisis de Conglomerados" alberga a diferentes algoritmos y métodos para agrupar objetos de tipo similar en categorías respectivamente. Una pregunta que confronta a investigadores en muchas áreas es la de cómo organizar datos observados en estructuras significativas, esto es, el desarrollo de taxonomías. En otras palabras el Análisis de Conglomerados es una herramienta de análisis exploratorio de datos cuyo objetivo es el de clasificar diferentes objetos en grupos de tal forma que el grado de asociación entre dos objetos es máxima si ellos pertenecen al mismo grupo y mínima de lo contrario. El Análisis de Conglomerados puede ser usado para descubrir estructuras en los datos sin proveer una explicación-interpretación, en otras palabras, simplemente descubre estructuras en los datos sin explicar por qué existen.

La creación de grupos basados en similaridades exige una definición de similaridad o de su complemento (distancia). Existen muchas formas de medir estas distancias y diferentes reglas matemáticas para asignar los individuos a distintos grupos, dependiendo del fenómeno estudiado y del conocimiento previo del posible agrupamiento que se tenga.

El Análisis de Conglomerados suele comenzar estimando las similitudes entre los individuos (u objetos) a través de correlación (distancia o asociación) de las distintas variables (métricas o no métricas) de que se dispone. A continuación se establece un procedimiento que permite comparar los grupos en virtud de las similaridades. Por último se decide cuántos grupos se construyen, teniendo en cuenta que cuanto menos sea el número de grupos, menos homogéneos serán los elementos que integran cada grupo. Se buscará formar el mínimo número de grupos lo más homogéneos posibles dentro de sí y lo más heterogéneos posibles entre sí.

En resumen el propósito del análisis de conglomerados es reducir un conjunto grande de datos en subgrupos significativos de individuos u objetos. La división se logra por medio de similaridades que existen entre objetos. Los outliers son un problema en esta técnica, causada frecuentemente por muchas variables irrelevantes. La muestra tiene que ser representativa de la población. Existen tres métodos principales de conglomerados: jerárquico, el cual es un proceso de árbol apropiado para un conjunto pequeño de datos; no-jerárquico, el cual requiere especificaciones del número de conglomerados por adelantado y el tercero es una combinación de los dos. Existen cuatro reglas principales para crear conglomerados: los conglomerados tienen que ser diferentes, tienen que ser mensurables, tienen que ser localizables y los conglomerados tienen que ser útiles y significativos.

5.2. Tipos de Análisis de Conglomerados

El método jerárquico es un procedimiento secuencial, tal que en cada paso solo un objeto o grupo de objetos cambian de grupos de pertenencia y los grupos en cada paso son anidados con respecto a grupos previos. De esta manera un objeto que ha sido asignado a un grupo nunca será removido del grupo después sobre el proceso de conglomeración. El método jerárquico produce una secuencia completa de soluciones de conglomerados empezando con n conglomerados (uno por cada objeto) y termina con un solo conglomerado que contiene a todos los n objetos. En algunas aplicaciones un conjunto anidado de conglomerados es la solución requerida mientras que en otras aplicaciones solo una de las soluciones de conglomerados en la jerarquía es seleccionada como la solución.

El método de no-jerárquico (k-medias) empieza con un número de conglomerados, digamos g, como el objetivo y después particionar los objetos para

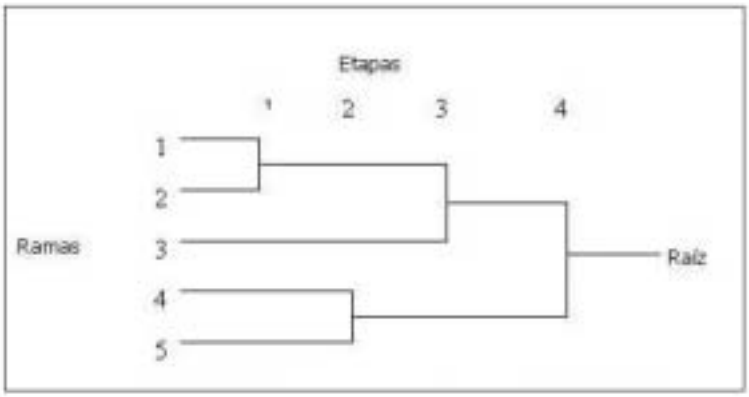


Figura 5.1: Diagrama de árbol para Conglomerado

obtener los requeridos g conglomerados. En contraste al método jerárquico, la técnica de particionar permite a los objetos cambio de agrupo de pertenencia a través del proceso de formación de conglomerados.

El método Q-sort incluye una variedad de técnicas que son similares a las de análisis de factores. Estos métodos usualmente empiezan con la matriz XX' y están implicados con el agrupamiento de objetos junto con aquellos elementos que fuera de la diagonal son relativamente grandes.

El método de densidad o moda, intenta procedimientos asumiendo que los objetos serán colocados en un espacio de tal forma que haya muchas áreas densas con regiones entre ellas que son dispersas. Este método asume la existencia de conglomerados naturales.

El método de dos fases usa un enfoque de conglomerado secuencial. Escanea los registros de los datos uno a uno y decide si los registros presentes debieran ser fusionados con conglomerados formados previamente o comenzar con un nuevo conglomerado basado en el criterio de distancia.

Finalmente, el método clumping diferente a las tres técnicas anteriores permite conglomerados superpuestos. Otro término usado para este tipo de análisis de conglomerados es el conglomerado indefinido (fuzzy clustering) porque este conglomerado permite la superposición.

5.3. Métodos Jerárquicos

El método más común al análisis de conglomerados es el método jerárquico. El método procede secuencialmente cediendo un arreglo anidado de objetos en grupos. El proceso jerárquico puede ser representado convenientemente usando un diagrama de árbol como se ilustra en la figura 5.1. Ésta ilustra el proceso de aglomeración jerárquica para una muestra de cinco objetos.

Empezando con el lado izquierdo de la figura, hay cinco objetos que pueden ser vistos como cinco conglomerados cada uno conteniendo un único objeto. El paso 1 del proceso (moviéndose un paso a la derecha) el objeto 1 y 2 son asociados para formar un grupo. Similarmente en el paso 2 objetos 4 y 5 son asociados para formar un grupo. Después de que el paso 2 esta completado, hay tres conglomerados ahora, dos que contienen dos objetos y uno que contiene solamente el objeto 3. En el paso 3, el objeto 3 se asocia al conglomerado que contiene el objeto 1 y 2, finalmente en el paso 4 todos los objetos son asociados para formar un solo conglomerado de cinco objetos.

5.3.1. Aglomerativo vs Proceso Divisible

En cada paso del proceso conglomerado jerárquico, procediendo de izquierda a derecha, dos grupos son asociados. Una vez que los grupos son asociados no se separaran durante el proceso. Este proceso jerárquico es llamado aglomerativo porque, como el proceso se mueve secuencialmente de n conglomerados a un conglomerado, el tamaño de los conglomerados crece y el número de conglomerados decrece. El proceso aglomerativo se mueve de las ramas del árbol del lado izquierdo a la raíz del árbol del lado derecho. Un proceso jerárquico que se mueve en orden contrario es llamado divisible. El proceso divisible empieza con todos los objetos en un conglomerado del lado derecho de la figura y se mueve hacia la izquierda. Este proceso por lo tanto se mueve de la raíz del árbol a las ramas del árbol. La aproximación divisible no es comúnmente usada y no se discutirá en esta sección.

En cada paso del proceso jerárquico, el valor de una función objetivo o criterio de conglomeración tiene que ser calculada para determinar cuales dos grupos tienen que asociarse. La función objetivo es usualmente basada en una medida de proximidad entre grupos que nosotros llamamos criterios de agrupamiento que son usados en el proceso jerárquico para determinar cuales grupos serán asociados en cada paso.

Empezando con la matriz de proximidad de los n objetos los dos objetos más cercanos son asociados en el paso 1 para formar un grupo. Antes de seleccionar el siguiente par de objetos o conglomerados que serán asociados, la matriz de proximidad tiene que ser revisada para reflejar las proximidades entre los nuevos conglomerados y los objetos restantes. El algoritmo para actualizar las medidas de proximidad entre grupos puedes ser usada para revisar las proximidades entre el nuevo conglomerado y los objetos restantes, este algoritmo es:

$$p_{tu} = \alpha_r p_{ru} + \alpha_s psu + \beta p_{rs}\gamma + \mid p_{rs} + psu \mid$$

Donde

1. t es la nueva referencia para el grupo resultante de la combinación de grupos r y s .
2. u es la referencia para algún otro grupo, otros de r y s.
3. α_r, α_s β y γ son coeficientes que depende de las medidas de proximidad que son usadas.

Después de revisar la matriz de proximidad una selección de los dos conglomerados más cercanos serán asociados en el paso 2 puede hacerse. Este proceso continua hasta que todos los objetos son contenidos en un solo conglomerado. Después de cada paso del proceso la matriz de proximidad es revisada para reflejar la relación entre los grupos que existen al momento. La matriz de proximidad es entonces usada para determinar los grupos que serán asociados en el siguiente paso.

El proceso de conglomerado jerárquico aglomerativo no provee una solución conglomerada única. De hecho cada paso del proceso es una solución conglomerada. La determinación del número apropiado de conglomerados involucra la selección de uno de los pasos del proceso jerárquico usando un segundo criterio óptimo. Una variedad de criterios óptimos serán resumidos.

5.3.2. Comparación de Criterios de Agrupamiento

Asumiendo que la distancia cuadrada euclidiana es una medida de proximidad subyacente, será útil comparar los criterios de agrupamiento en el contexto

de un proceso de conglomerado jerárquico. El single linkage (vecino mas cercano) y el complete linkage (vecino mas lejano) son comparados en la figura 5.2 Usando el criterio de agrupamiento single linkage, los grupos r y u están mas cercanos que los grupos r y s. La distancia entre r y u es la distancia de D_r a D_u y la distancia de r y s es la distancia de C_r a C_s. En el caso del criterio de agrupamiento complete linkage r y s están cercanos porque la distancia de B_r a B_u es mayor que la distancia de A_u a A_s. Como un resultado de estas diferencias es fácil imaginar como el criterio de agrupamiento single linkage sugiere un conglomerado tipo cadena, mientras que complete linkage sugiere un conglomerado compacto. Como se puede ver en la figura 5.2 es posible un conglomerado single linkage para un objeto en un conglomerado estar cercano a un objeto en otro conglomerado que para algunos objetos en su mismo conglomerado.

Una segunda comparación interesante entre single y complete linkage es el impacto en el tamaño del conglomerado sobre la medida de proximidad. Imagine dos distintos conglomerados que crecen en tamaño en un espacio limitado. La medida del criterio de agrupamiento single linkage entre dos conglomerados seguirá constante, mientras que la medida del criterio de agrupamiento complete linkage se incrementará. De igual manera suponga que un punto aislado o un outlier y su medida de proximidad a un conglomerado crecen en tamaño. La medida de proximidad del criterio single linkage a un outlier seguirá inmóvil, pero la medida de proximidad del criterio complete linkage tendera a incrementarse. En un proceso jerárquico la medida de proximidad se incrementa conforme se incrementa el conglomerado en tamaño. Puesto que la medida del criterio complete linkage está basado en la asociación más débil, un punto aislado se vuelve relativamente más cercano a un conglomerado existente que con el criterio single linkage. Con el método single linkage los outliers tienden a permanecer como un punto aislado hasta el final en el proceso jerárquico. El método single linkage se dice ser espacio conservador, mientras que el método linkage es llamado espacio diluido o espacio suficiente.

Ambos criterios single y complete linkage emplean únicamente una sola medida de proximidad para representar proximidades grupales y por esta razón son muy susceptibles a observaciones extremas. En un proceso jerárquico single linkage, un solo outlier falsea la información entre dos conglomerados, que puede resultar en una asociación eventual de los dos grupos. En el caso de un proceso complete linkage, pequeños cambios en el localización de puntos particulares o errores pueden tener un impacto sustancial en la solución jerárquica.

Los criterios average linkage (vinculación media), centroide y Ward son usualmente preferidos a los criterios single linkage y complete linkage por su relativa insensibilidad a los extremos o outliers. Dependiendo de los tipos de conglomerados esperados esta propiedad puedes ser también una desventaja. El criterio average linkage se determina mediante un promedio de proximidades entre todos los pares de objetos (un objeto de cada grupo). Este proceso average linkage tiene interesantes propiedades. En la Figura 5.3 dos grupos son mostrados, A con un punto y B con dos puntos (B_L, B_U). La distancia promedio en una dimensión es:

$$\bar{d} = \frac{[(X_1 - X_2 - h)^2] + [(X_1 - X_2 - h)^2]}{2} = (X_1 - X_2)^2 + h^2$$

El método average linkage entre A y B se incrementa cuando B_L y B_U se alejan del centro de B. Así la distancia promedio entre el punto A y el grupo B se incrementa cuando la distancia entre los puntos en B se incrementan. El criterio average linkage entre dos grupos basado en la distancia cuadrada euclidiana por lo tanto crece cuando los dos grupos se vuelven menos compactos.

Una segunda propiedad del criterio average linkage puede ser ilustrada tam-

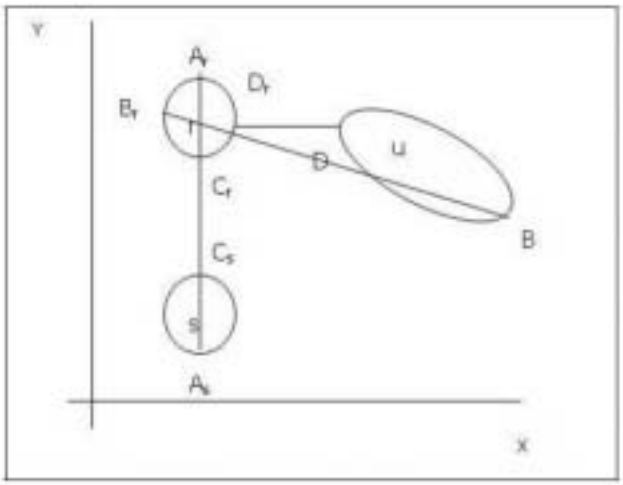


Figura 5.2: Single linkage vs Complete Linkage

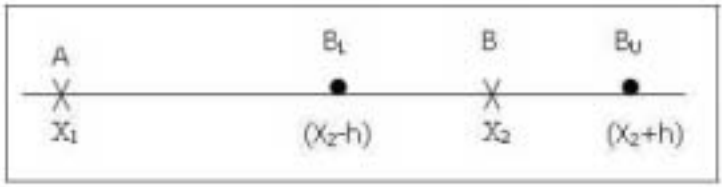


Figura 5.3: Distancia entre Conglomerados usando Vinculación Media (Average Linkage)

bién en la figura 5.3. Si B representa un punto grupal y A representa un punto grupal la distancia euclidiana cuadrada promedio esta dada por $(X_1 + X_2)^2$. En comparación con el ejemplo anterior, A esta cercano al grupo B, cuando B fue considerado un grupo de dos puntos. De esta manera cuando el tamaño de los grupos se incrementa, la medida average linkage se incrementa a menos que todos los puntos en el grupo sean localizados en el centroide.

El comportamiento del método average linkage del conglomerado jerárquico en la presencia de outliers puede ser explicado usando las propiedades antes mencionadas. Inicialmente aunque al principio los conglomerados son pequeños y compactos, los outliers tienden a permanecer aislados. Cuando los conglomerados crecen en tamaño la medida de proximidad media entre conglomerados crece hasta que un punto es alcanzado donde los outliers tienen la medida de proximidad de magnitud similar. Puesto que el tamaño y compactes de los conglomerados influye en la medida average linkage, los outliers son los mas probables de asociar con otros outliers y menos probables de asociar con los menos compactos y/o grandes conglomerados. El método average linkage es caracterizado por su tendencia para formar grupos de outliers cerca del proceso final de jerarquización.

Los métodos de centroide y Ward son fáciles de comparar porque las medidas de proximidad están basadas en d_{rs}^2, la distancia euclidiana cuadrada entre grupos de centroides. Puesto que el coeficiente d_{rs}^2 es 1 para el método del centroide y $\frac{n_r n_s}{n_r+n_s}$ para el método Ward es únicamente necesario determinar el impacto que el tamaño de los conglomerados n_r y n_s tienen sobre la medida Ward.

El coeficiente $\frac{n_r n_s}{n_r+n_s}$ puede ser escrito en otras formas como son $\frac{n_s}{(1+\frac{n_s}{n_r})}$ y $\frac{1}{(\frac{1}{n_r}+\frac{1}{n_s})}$. A partir de estas expresiones podemos concluir que n_r y n_s tienen un gran tamaño así como el coeficiente $\frac{n_r n_s}{n_r+n_s}$ y también cuando n_r crece relativamente a n_s el coeficiente se incrementa. Dados dos conglomerados cuyos centroides son una distancia d_{rs}^2, podemos concluir a parte, que el método Ward podría aceptar una medida de proximidad grande entre los centroides comparativamente el tamaño de los conglomerados se incrementa y comparativamente el tamaño de los conglomerados se vuelves menos iguales. Podemos concluir que en comparación con el método centroide, el método Ward tiene gran tendencia a formar conglomerados de igual tamaño y/o pequeños. El método centroide tiende a asociar conglomerados cuyos centroides están cercanos, mientras que en el método Ward tiende a asociar conglomerados pequeños fuera de alcance de otro conglomerado pequeño más distante. Así como single linkage el método centroide es un espacio contraído, mientras que el método Ward así como el método complete linkage es espacio diluido.

Es también interesante examinar el impacto que los outliers podrían tener sobre el proceso de conglomerado jerárquico así como el comparar estos procesos con el método average linkage. Considerando la medida de proximidad entre un outlier y un grupo con n_s objetos. El coeficiente de d_{rs}^2 en este caso esta dado por $\frac{n_s}{(1+n_s)}$.Cuando n_s es pequeño el coeficiente es también relativamente pequeño y cuando n_s se incrementa el coeficiente se aproxima a 1. Por la razón de que el coeficiente podría ser relativamente pequeño en el proceso jerárquico en forma anticipada habría sido una tendencia para puntos insolados para ser más compatible con pequeños conglomerados en los estados de procesos de manera anticipada. Dos puntos aislados pudieran tener un coeficiente de $\frac{1}{2}$ y de aquí que serán probablemente asociados en el proceso de forma anticipada. Dos conglomerados grandes podrán tener un coeficiente $\frac{n_r n_s}{(n_r+n_s)}$. Al mismo tiempo comparado a un coeficiente de $\frac{1}{2}$ para dos conglomerados de puntos distintos pueden ser visto que una distancia grande entre outliers pueden eventualmente

ser afectado cuando $\frac{n_r n_s}{(n_r+n_s)}$ se incrementa relativamente a $\frac{1}{2}$. Cuando el método average linkage es comparado con el método Ward no es excepcional encontrar que en el método Ward los outliers son agrupados de manera anticipada en el proceso jerárquico.

5.3.3. Algunas Aproximaciones Multivariadas para Conglomerado Jerárquico

En la sección de criterios de agrupamiento se hablo de la suma de cuadrados aumentada y se demostró estar relacionado a los cambios en trW y trG donde W y G están en el interior y entre grupos de matrices de sumas de cuadrados respectivamente. Después de que los grupos r y s sean asociadas en el paso l, las nuevas matrices W_1 y G_1 pueden ser expresadas en términos de estas matrices en el paso previo $(l-1)$ como:

$$W_l = W_{l-1} + \frac{n_r n_s}{(n_r+n_s)}(\bar{x}_r - \bar{x}_x)(\bar{x}_r - \bar{x}_s)'$$

y

$$G_l = G_{l-1} - \frac{n_r n_s}{(n_r+n_s)}(\bar{x}_r - \bar{x}_x)(\bar{x}_r - \bar{x}_s)'$$

en cada paso del algoritmo de Ward trata de minimizar el incremento en trW_l y minimizar el decremento en trG_l. Porque este criterio no toma en cuenta los elementos fuera de la diagonal de W y G este contribuye a producir conglomerados de forma esférica. Este método es óptimo consecuentemente si la matriz de covarianzas subyacente $\sum$, es esférica, es decir $\sum = \sigma^2 I$.

Un criterio alternativo, el cual toma en cuenta las covarianzas entre las variables, minimizando $\mid W \mid$ en lugar de trW. Cuando los grupos r y s son asociados el valor corregido de $\mid W \mid$ esta dado por:

$$\mid W \mid = \mid W_{l-1} \mid (\frac{n_r n_s}{(n_r+n_s)})(\bar{x}_r - \bar{x}_s)' W_{k-1}^{-1}(\bar{x}_r - \bar{x}_s)$$

Por esta razón cada paso de los dos grupos seleccionado para asociarse tienen que minimizar una función de la distancia Mahalanobis entre los centroides grupales. Es importante recordar que la distancia Mahalanobis es equivalente a la distancia euclidiana con componentes principales usados en lugar de las X variables. En comparación al criterio trW el criterio $\mid W \mid$ toma en cuenta efectos de correlación. El uso de los criterios anteriores podría por lo tanto tener una tendencia a generar conglomerados de forma elíptica. Ambos criterios podrían por lo tanto tener una tendencia a producir conglomerados de la misma forma por la suposición de homogeneidad de las matrices de covarianzas entre grupos. Un criterio alternativo, el cual permite variaciones grandes en la forma del conglomerado, es el minimizar $\Pi_{k=1}^{g} = \mid W_k \mid^{nk}$ donde W_k y n_k corresponden a el grupo de conglomerado k.

Un criterio multivariado alternativo para escoger conglomerado esta relacionado al análisis discriminante y la Manova. El criterio busca escoger conglomerados que maximicen $trGW_{-1}$. Puesto que este criterio tiene una tendencia a maximizar los eigenvalores grandes de GW^{-1} conglomerados obtenidos tienden a expandirse.

5.3.4. Estimando la Solución Jerárquica y la Selección del Conglomerado

Como se describió anteriormente el método jerárquico del análisis de conglomerados produce una secuencia anidada de soluciones de conglomerados que va

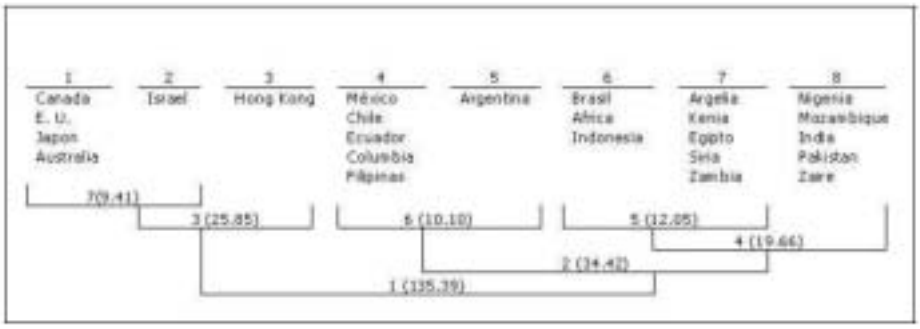


Figura 5.4: Dendograma Parcial

desde -el número total de objetos- hasta 1. Las soluciones de conglomerados que son seleccionados de esta jerarquización para usos posteriores dependen de la aplicación en particular. En algunos casos una jerarquización anidada dentro de un rango de soluciones puede ser usada para resumir las relaciones entre varios subgrupos. Un ejemplo podría ser un grupo de plantas o animales en taxonomía numérica. Alternativamente, una solución particular de conglomerado (únicamente un paso en la jerarquía) puede ser seleccionada para ser usada como un agrupamiento conveniente para análisis posteriores. Dependiendo de la aplicación la solución jerárquica podría requerir un estudio adicional antes de hacer la selección de la solución particular (o soluciones). En esta sección algunas técnicas serán presentadas para estimar la calidad de la solución jerárquica y para escoger una solución conglomerada apropiada.

5.3.5. Dendogramas y Proximidades Derivadas

En el conglomerado jerárquico un diagrama de árbol como en la figura 5.1 puede ser usado para guardar el historial del proceso secuencial conglomerado, el diagrama indica el valor de la proximidad por cada paso. Esta medida de proximidad derivada indica el grado de similaridad entre dos conglomerados que han sido asociados en la etapa correspondiente. Cuando estos valores de proximidades son incluidos con el árbol, el diagrama de árbol es usualmente llamado dendograma. Un dendograma por lo tanto contiene una escala de proximidad derivada que muestra el valor de la medida de proximidad en cada paso del proceso jerárquico. La Figura 5.4 es un ejemplo de un dendograma particionado puesto que únicamente los estados finales del proceso son mostrados. Los valores de la proximidad monótonamente se incrementan si el proceso de conglomerado jerárquico satisface la desigualdad ultrametrica que se menciono en la sección de criterios de agrupamiento. Una tabla la cual resume la información del dendograma es llamada diagrama aglomerativo.

Para los $(n-1)$ en el proceso secuencial el orden de los pasos forma una relación uno a uno con la medida de proximidad en el dendograma. El número de los pasos al cual dos objetos dados primero aparecen juntos en el mismo conglomerado es llamado posición de la partición. El conjunto de las proximidades derivadas correspondientes en el dendograma puede ser usado para obtener una nueva matriz de proximidades llamada matriz de proximidades derivadas. Cuando dos grupos de objetos son asociados, todos los posibles pares de objetos derivados, de los objetos en grupos opuestos, son asignados al mismo valor de proximidad derivado. Puesto que son únicamente $(n-1)$ valores de proximidades únicas en el dendograma, la matriz de proximidades derivada que contiene un total de $\frac{n(n-1)}{2}$ valores tiene que tener muchos valores en común.

5.3.6. Correlación Cofonética y la Validez del Conglomerado

Como una medida de validación del conglomerado es a veces de interés el comparar la matriz de proximidades derivada con la matriz de proximidades original. El método más común de comparación es el de calcular la correlación de Person entre los valores originales y los valores derivados. La correlación resultante es llamada correlación cofonética. La magnitud de esta correlación debería ser muy cercana a 1 para tener una solución de alta calidad. Esta medida puede también ser usada para comparar soluciones de conglomerados alternativos obtenidos de diferentes algoritmos.

5.3.7. Stress

Un método alternativo de comparación de dos conjunto de proximidades en el caso de distancia euclidiana es el de calcular la medida stress (énfasis) que es:

$$\frac{\sum_{i}^{n}\sum_{<j}^{n}(p_{ij}-\hat{p}_{ij})^2}{\sum_{i}\sum_{<j}\hat{p}_{ij}^2}$$

donde p_{ij} denota la proximidad original y $\hat{p}_{ij}$ la proximidad derivada. Esta medida es comúnmente usada en escalas multidimensionales para evaluar la solución escalar.

5.3.8. Proximidad Derivada Alternativa Basada en Centroides

Una aproximación alternativa para derivar proximidades de los resultados del proceso jerárquico podría ser el de regresar sobre cada uno de los pasos y calcular la distancia entre los centroides de los conglomerados asociados en cada paso. Estas distancias pudieran entonces ser usadas como la medida de proximidad derivas. En este caso, el método del centroide no está siendo usado como un criterio para formación de conglomerados pero es usado como medida para proximidades derivadas. En este caso, de cualquier manera las proximidades derivadas podrían no satisfacer la desigualdad ultrametrica.

5.3.9. Eligiendo el Número de Conglomerados

En le proceso de conglomerado jerárquico, una secuencia de soluciones de conglomerados es obtenida con una solución ideal para cada número posible de conglomerados desde n hasta 1. Un segundo paso del análisis de conglomerados es el de seleccionar un número óptimo de conglomerados. Para facilitar la determinación de la solución apropiada un criterio de optimización será usado. Cuando el número de conglomerados g declina de n a 1 la solución de conglomerados es evaluada mediante el cálculo de uno o varios criterios de optimización. Al final del proceso jerárquico el criterio de optimización es estudiado para determinar el valor óptimo de g.

La aproximación más simple para escoger los conglomerados usa el valor de la medida de proximidad grupal para los dos grupos asociados en cada paso. Como el proceso se mueve del paso 1 a 1 paso $(n-1)$ el valor de la medida de proximidad grupal, dígase s, se incrementara (por medidas disimilaridades). Si n es grande en s incremento en será inicialmente pequeño pero tendera a crecer exponencialmente. De esta manera cuando g decrece de n a 1 el valor de la medida de proximidad grupal debería comportarse como en la figura 5.5 a).

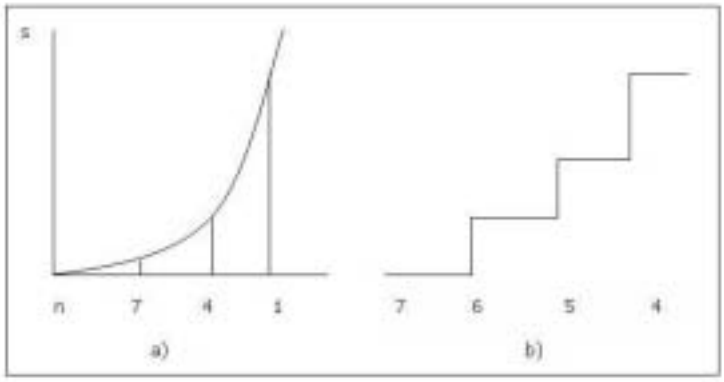


Figura 5.5: Esquema del Criterio Conglomerado

Una aproximación de la selección de un valor apropiado de g será para examinar el comportamiento de s en una área deseada. Si ocurre un gran cambio de s en algún valor de g entonces la solución $(g+1)$ inmediatamente anterior a este paso debería ser la escogida. La figura 5.5 b) ilustra este concepto; la función da la impresión de tener un mucho mayor salto en el paso 4 que en los dos pasos anteriores y de aquí que la solución con 5 conglomerados deberá ser seleccionada. Una alternativa de aproximación gráfica involucra esquematizar los cambio en S,ΔS, como una función del número de conglomerados. Inicialmente la curva ΔS debería crecer despacio pero eventualmente crecería rápidamente cuando conglomerados distintos son combinados. Una inflexión o un cambio dramático en la inclinación de ΔS podría ser el indicativo del final conveniente del proceso de conglomerado.

5.3.10. Prueba Estadística para el Número de Conglomerados

A cada paso del proceso jerárquico, una medida de proximidad derivada es posible indicando la medida de proximidad grupal correspondiente a los dos conglomerados asociados en ese paso. Si el algoritmo empleado satisface la desigualdad ultrametrica las medidas derivadas se incrementaran monótonamente durante el proceso.

El primer estadístico es llamado la regla de cola superior y esta basado en la aceptación del hecho de que no hay conglomerados. La medida de proximidad derivada es simplemente un conjunto de estadísticos de orden correspondientes a la muestra de alguna distribución de probabilidad subyacente. Si la medida de proximidad derivada subyacente puede asumir una distribución normal, los valores de la medida obtenidos del proceso jerárquico pueden ser tratados como una muestra de una distribución normal. Denotando las medidas derivadas mediante $s_1, s_2, \ldots, s_{n-1}$ correspondiente a $1, 2, \ldots, (n-1)$ conglomerados, la media de la muestra

$$\bar{s} = \frac{\sum_{j=1}^{n-1} s_j}{(n-1)}$$

y la desviación estándar de la muestra

$$v = \sqrt{\frac{\sum_{j=1}^{n-1} (s_j - \bar{s})^2}{(n-2)}}$$

pueden ser usadas para derivar una prueba estadística. Los valores estandarizados de las medidas de proximidad observadas están dadas por $\binom{(s_j-\bar{s})}{v}$. Si esta prueba estadística es relativamente grande a una estadística normal estandarizada entonces puede concluirse que el conglomerado formado en el paso j no es óptimo. El valor de la medida de proximidad correspondiente s_j al paso j en este caso es considerado muy grande. Esta primera regla es conservada porque si s_j es relativamente grande el valor de $\bar{s}$ y v serán también grandes y por esto el valor estandarizado será muy pequeño.

Una segunda prueba es llamada regla media móvil (moving average rule) emplea una media móvil ajustando a una línea obtenida mediante el ordenamiento de los valores de proximidad $s_1, s_2, \ldots, s_{n-1}$. Asumiendo que no hay conglomerados el ordenamiento de s_j valores se desea que sean aproximadamente lineales con alguna inclinación b_j. Usando un punto r basado en la media móvil sobre los puntos anteriores a s_{j+1}, s_{j+1} el valor ajustado de $\hat{s}_{j+1}$, está dado por $(\bar{s}_j + L_j + b_j)$, donde b_j es la inclinación cuadrada media móvil de la línea y $L_j = \binom{(r-1)b_j}{2}$. Denotando una estimación de la desviación estándar de $\hat{s}_{j+1}$ mediante v_j, una prueba estadística esta dada por $\binom{(s_{j+1}-\hat{s}_{j+1})}{v_j}$. Una vez más valores grandes de este estadístico, relativo a una estadística normal estándar podrían indicar que el $(j+1)$ agrupamiento no es óptimo. En comparación con la regla de cola superior, este estadístico tiene la ventaja de que las cantidades usadas para derivar $\hat{s}_{j+1}$ no dependen de y de los valores más grandes.

5.3.11. Estadísticos tipo ANOVA

Sin tomar en cuenta cual criterio es usado para llevar a cabo la proceso jerárquico (average linkage, complete linkage, single linkage, Ward, etc.) si la matriz de proximidad original es la distancia euclidiana cuadrada, entonces la suma de matrices cuadradas T, W y G pueden ser usadas para construir una variedad de medidas para ayudar en la selección de conglomerados. Como se menciono las cantidades trW, trT y trW miden el total de suma de cuadrados, dentro de la suma de cuadrados y entre las sumas de cuadrados correspondientes. En cada paso del proceso jerárquico trW se incrementa apoyado por $(SSW_t - SSW_r + SSW_s)$ mientras que trG decrece en la misma cantidad. La suma total de cuadrados continúa definida sobre todo el proceso.

5.3.12. Seudo F, Seudo t^2 y Razón F Beale

Tres estadísticos tipo F que son a veces usados para seleccionar conglomerados son derivados de los cambios en las sumas de cuadrados descritas anteriormente. Dos estadísticos producidos son llamados seudo F y seudo t^2. El seudo estadísticoestá dado por:

$$F^* = \frac{\left[\binom{trG}{(g+1)}\right]}{\left[\binom{trW}{(n-g)}\right]}$$

también se ha llamado el criterio proporción varianza. Bajo la hipótesis de normalidad multivariada con matriz de covarianzas esféricas, este estadístico es el convencional estadístico ANOVA de una cola para probar igualdad de medias de conglomerados. Con esta hipótesis alternativa fuerte,F^* tiene una distribución F con $p(g-1)$ y $p(n-g)$ grados de libertad si los vectores medios conglomerados con iguales. Esta estadística puede ser comparada a una tabulación F usando un apropiado p-value Bonferroni para evaluar la significancia de los conglomerados.

Un uso alternativo del estadístico F^* es el de monitorear el comportamiento de F^* sobre los pasos del proceso. Inicialmente cuando g decrece,F^* declinara en valor cuando trW se incrementa gradualmente y trG decrece gradualmente. El decline gradual casi monótono en F^* que ocurre cuando objetos similares son asociados. En algún punto en el proceso un decline inesperado relativamente grande en F^* debería ocurrir si la asociación de dos conglomerados como resultado en un cambio grande en trW y trG. El valor de g inmediato anterior a este punto debería ser considerado como un valor optimo posible de g. Este estadístico podría funcionar bien si hay una pequeña cantidad de conglomerados de forma esférica que sean distintos.

Un segundo estadístico que es similar en esencia a F^* es el seudo estadístico t^2 dado por:

$$t^* = \frac{[SSW_t - SSW_r + SSW_s](n_r + n_s - 2)}{[SSW_r + SSW_s]}$$

El numerador de la suma de cuadrados en t^{*2} mide la suma de cuadrados incrementada resultante de la asociación de conglomerados$r$ y s para formar un nuevo conglomerado t. El denominador de la suma de cuadrados es la suma de la suma interior para los dos conglomerados que han sido asociados. Los tamaños de los conglomerados son n_r y n_s respectivamente. Bajo la suposición de normalidad multivariada con matriz de covarianza esférica, el seudo estadístico t^2 tiene una distribución F con p y $p(n_r + n_s - 2)$ grados de libertad si los dos conglomerados asociados no son distintos. Así como arriba este estadístico puede ser comparado con un p-value Bonferroni para representar una prueba aproximada de significancia del conglomerado. Como en el caso de F^*, t^{*2} puede ser usada para monitorear el proceso jerárquico. Un valor relativamente grande de t^{*2} de conglomerados podría sugerirse $(g + 1)$ como una selección de conglomerado posible.

Monitorear el seudo estadístico t^2 es equivalente a monitorear el estadístico $\frac{[SSW_t + SSW_s]}{SSW_t}$ sobre el proceso jerárquico. Un decline inesperado en esta medida podría ser indicativo de la asociación de dos conglomerados muy distintos.

Un tercer estadístico tipo F es comúnmente referido como razón F Beale y esta dado por

$$F \doteq \frac{[trW_1 - trW_2]}{trW_2} \Big/ [\binom{(n-g_1)}{(n-g_2)}\binom{g_2}{g_1}^{\binom{2}{p}}] - 1$$

donde W_1 y W_2 denotan la matriz W correspondiendo a los conglomerados g_1 y g_2 respectivamente y donde $g_2 > g_1$. Si la solución g_2 es significativamente mejor que g_1, F' puede ser comparada con una estadística F para $p(g_2 - g_1)$ y $p(n - g_1)$ grados de libertad. Esta es solamente una aproximación a F como en los dos estadísticos previos.

Si los dos conglomerados solución g_1 y g_2 son consecutivos $g_2 = (g_1 + 1)$ entonces F' esta dada por:

$$F' = \frac{[SSW_t - SSW_r + SSW_s]}{trW_2} \Big/ [\binom{(n-g_1)}{(n-g_1-1)}\binom{g_1+1}{g_1}^{\binom{2}{p}}] - 1$$

En este caso el incremento en la interior del grupo las sumas de cuadrados son comparadas al total dentro del grupo de las suma de cuadrados anteriores a este punto en el proceso. El radio debería ser por lo tanto una medida confiable de cambio en rtW que en el estadístico t^{*2} , cuando el proceso jerárquico esta controlado por el valor de $[SSW_t - SSW_r + SSW_s]$ como en el método Ward.

5.3.13. Medidas tipo R^2

Una medida de la partición del total de la suma de cuadrados trT, entre trW y trG, esta dada por $R^2 = \frac{trG}{trT}$. Esta razón indica la proporción del total de la variación entre los objetos que es justificada por la variación entre los grupos de conglomerados. Cuando el número de conglomerados declina R_g^2 también declina. Un decremento inesperado en R_g^2 podría indicar que las asociaciones de dos conglomerados son considerablemente distintas. Otro estadístico relacionado a R_g^2 es llamado semiparcial R_g^2 y esta dado por $\Delta R^2 = R_g^2 + R_{(g-1)^2}$. El estadístico semiparcial R_g^2 calcula la razón de $[SSW_t - SSW_r + SSW_s]$ a trW. Puesto que el numerador de ΔR^2 representa la suma de cuadrados incrementada, esta cantidad puede ser monitoreada a través del proceso aunque el método Ward podría no ser el criterio de conglomerado jerárquico el usado. Si el método average linkage esta siendo usado para seleccionar el conglomerado, la semiparcial R^2 provee información usando un criterio alternativo. El estadístico ΔR^2 es también útil para comparar dos o más alternativas de soluciones jerárquicas basada en diferentes criterios.

5.3.14. Medidas Tipo Correlación y la Calidad del Conglomerado

La medida tipo correlación y la calidad del conglomerado están basadas en una comparación de la matriz de proximidad original y la localización del grupo conglomerado para cada objeto. Las medidas están basadas en el principio de que los objetos que están en el mismo conglomerado en cualquier paso deberían tener medidas de proximidad originales cercanas que aquellos objetos que están en diferente conglomerado. En cada paso del proceso conglomerativo todos los pares de objetos están asignados a una nueva o proximidad derivada valor basado en que si o no los pares están en el mismo grupo conglomerado. Pares en los cuales objetos pertenecientes al mismo conglomerado les son asignados el valor 0, mientras que aquellos cuyos objetos están en diferentes conglomerados son asignados el valor 1. Todos los pares codificados 0 son llamados pares internos (within pairs) y todos los pares codificados 1 son llamados entre pares (between pairs). Los coeficientes de correlación entre las proximidades originales y los valores asignados pueden ser usados para determinar la calidad del conglomerado.

5.3.15. Correlación Puntual-Biserial

El coeficiente de correlación de Person entre las proximidades $\binom{n(n-1)}{2}$ originales y el valor asignado correspondiente (0 o 1) es llamado la correlación puntual biserial. La correlación puede también ser determinada usando la expresión:

$$r_b = \frac{(\bar{d}_b - \bar{d}_w)\left(\frac{n_b n_w}{n_d^2}\right)^{\frac{1}{2}}}{s_d}$$

donde los subíndices b y w corresponden a los grupos de pares codificados 1 (pares medios) y codificados 0 (pares internos) respectivamente. Las medias de la proximidades originales para los dos grupos son denotadas por $\bar{d}_b$ y $\bar{d}_w$. El número de pares en cada uno de los dos grupos es denotado mediante n_b y n_w. El número total de pares $\binom{n(n-1)}{2} = (n_b + n_w)$ esta denotado por n_b y n_w la desviación estándar de las proximidades originales esta denotada por s_b. Un valor relativamente alto de este coeficiente de correlación (cercano a 1) podría indicar que los pares codificados 1 tienden a tener valores de proximidad bajos

(relativamente disimilares) mientras que los pares codificados 0 tienden a tener valores de proximidad altos (relativamente similares).

5.3.16. Gamma y G(+)

Una alternativa de medida de correlación entre las proximidades originales y los valores numéricos asignados puede ser obtenida usando un coeficiente de concordancia. Los dos grupos de pares definidos arriba, interno y medio, son comparados, uno de cada grupo, para generar un total de $(n_w)(n_b)$ comparaciones. Todas las comparaciones en cuyas proximidades originales de los pares internos, exceden (más similares) la proximidad original de los pares medios, son clasificados como concordantes pero en contraste el caso opuesto es clasificado como discordante. El número total de comparaciones es por lo tanto dividido en dos grupos conteniendo $S(+)$ concordantes y $S(-)$ comparaciones discordantes, de aquí que $S(+) + S(-) = n_w n_b$. El coeficiente gamma de concordancia esta dado por:

$$\gamma = \frac{(S(+) - S(-))}{(S(+) + S(-))}$$

Un valor de γ cercano a 1 por lo tanto indica una concordancia cercana entre proximidades originales y pertenencia de grupos conglomerados. El coeficiente de Kendall de concordancia entre proximidades originales y los valores de codificación 0-1 son equivalentes al coeficiente gamma si todas las coincidencia son eliminadas de los cálculos. Una medida de alternativa de concordancia esta dada por $G(+) = \frac{S(-)}{[S(+)+S(-)]}$. Para este coeficiente, un valor cercano a 0 es indicativo de concordancia cercana entre grupos conglomerados y las proximidades originales.

5.3.17. Combinando Análisis de Conglomerado Jerárquico con otros Métodos Multivariantes

Antes de que una solución de análisis conglomerado sea obtenida, es de interés la caracterización del conglomerado con respecto a las variables usada para derivarlas. Como un resultado del análisis conglomerado, una nueva variable categórica denotando la pertenencia grupal conglomerada puede ser adicionada a la matriz de datos original X. El reto de la caracterización del conglomerado por lo tanto consiste en relacionar la nueva variable categórica a las variables p originales en X. Cada una de las p variables puede ser relacionada a la variable individual conglomerada usando la ANOVA o simultáneamente usando la MANOVA. Por la razón de que son usualmente un número grande de variables un análisis de componentes principales puede ser usado inicialmente para reducir el número de variables. Alternativamente, un análisis discriminante basado en las p variables originales pueden ser usadas para caracterizar las diferencias entre los grupos conglomerados.

5.4. El Algoritmo k-medias

Un uso común del método de particiones es el algoritmo de k-medias cuyas medidas de proximidad entre grupos usando la distancia euclidiana entre grupos de centroides. En este método la búsqueda es para k conglomerados. Usando la notación empleada aquí será renombrada algoritmo medias. Iniciando con una selección de g grupos los objetos son reasignados hasta que son colocados en el grupo con el centroide más cercano. Cuando los objetos son reasignados

los centroides del grupo tienen que ser revisados. Para un nivel dado de g, el equilibrio es alcanzado cuando todos los objetos son colocados en el grupo cuyo centroide es el más cercano.

5.4.1. Seleccionando la Partición Inicial

La fase del grupo inicial puede ser con un conjunto preseleccionado de g conglomerados o, con un conjunto preseleccionado de g puntos semilla u objetos que son después usados para colocar los objetos restantes. El conjunto inicial de g conglomerados puede ser obtenido de otra solución conglomerada, de un estudio previo, o bien puede ser dictado por algunas hipótesis subyacentes.

El número de soluciones iniciales posibles es grande si n es grande relativamente a g. El número de particiones posibles de n objetos en g grupos esta dado por:

$$\frac{1}{g!}\sum_{i=1}^{g}\binom{g}{1}(-1)^{g-i}i^n$$

el cual es de orden $\frac{g^n}{g!}$ cuando n es grande. Para evaluar todas las posibles particiones entonces se tendría que prohibir una n grande. Por lo tanto, en la práctica, soluciones iniciales bien seleccionadas serán requeridas así como diferentes configuraciones iniciales deberían ser usadas para asegurar la valides de la solución final.

En la ausencia de un conjunto inicial de g conglomerados el paso inicial empieza con un conjunto de g objetos o puntos semilla alrededor de los cuales se formaran g conglomerados. Las medidas de proximidad son calculadas entre cada uno de los $(n-g)$ objetos restantes. Otro criterio que podría ser usado incluye trW,$\mid W \mid$ y $G^{-1}W$. En adición un procedimiento sugerido es RELOCATE (recolocar) esta basado en la división de distancias euclidianas cuadradas en dos componentes.

5.4.2. Usando Recolocamiento

El procedimiento recolocar es una generalización del algoritmo k medias. Dando una partición particular de los n objetos en g grupos, el procedimiento recolocar usa un criterio dado para medidas de proximidad entre cada objeto y g grupos. Cada objeto es entonces colocado dentro del grupo más cercano. El proceso es ejecutado secuencialmente de tal manera que después de que cada objeto es colocado, el criterio es calculado para colocar el siguiente objeto. El proceso continúa hasta que todos los objetos son colocados en su grupo más cercano. Por esto es posible para un objeto moverse más de una vez antes de que el equilibrio sea alcanzado. Un procedimiento de conglomerado jerárquico puede ser ejecutado mediante la asociación de dos conglomerados más cercanos en cada equilibrio. La partición inicial puede ser basada en una colocación aleatoria o puedes ser basada en una solución del análisis conglomerado anterior.

Para el propósito de probar la validez del conglomerado se sugiere que el resultado de tres alternativas de procedimiento dos pasos sea comparada. Inicialmente una colocación aleatoria es hecha para g' conglomerados donde g' es más grande que la solución esperada. Una solución primer paso consistente en recolocar la solución inicial usando el componente tamaño donde una segunda solución primer paso consistente en la recolocación de la solución inicial usando el componente forma. La solución tercera del primer paso consiste de la colocación aleatoria inicial con no recolocación. El segundo paso del proceso recomendado produce una solución jerárquica usando la medida distancia

euclidiana para reducir el número de conglomerados y en cada paso un recolocamiento de objetos. El proceso es ejecutado para cada una de las soluciones de los tres primeros pasos y produce tres conjuntos de solución jerárquica cada uno que va desde g' conglomerados hasta 1.

5.5. Clasificación Tipológica y Métodos Q-sort

La aproximación clasificación tipología para conglomerar objetos ha sido usada ampliamente en psicología y psiquiatría para la clasificación de personalidades humanas y varios desordenes psicológicos. La aproximación esta basada en las propiedades de una descomposición espectral de la matriz XX' de orden $(n \times n)$. Por la razón de que el numero de variables p es usualmente mucho menor que el numero de individuos n, el rango de XX' es usualmente p. La descomposición espectral por lo tanto cae en un conjunto de p componentes que pueden ser vistos como un conjunto de clases "puras". El proceso de clasificación entonces involucra la asignación de n individuos a las más cercanas clases puras.

Este método puede ser visto como un análisis factorial ejecutado sobre de las n filas de X en lugar de las p columnas de X y por esta razón es también referido como método Q-sort en lugar del convencional R-sort del método de análisis factorial. Para una matriz de datos X de orden $(n \times p)$ con una apropiada estandarización, la matriz $X'X$ es llamada una matriz tipo R, mientras que la matriz XX' es llamada matriz tipo Q. La matriz tipo R resume la correlación entre las columnas o variables de X, mientras que la matriz tipo Q resume las correlaciones entre las filas o características de X.

Los eigenvectores de XX' proveen vectores de coeficientes que pueden ser usados para obtener las clases puras como una combinación lineal de n objetos. Una rotación de la solución inicial es usualmente ejecutada en un intento de obtener clases puras que dependen de únicamente un número pequeño de individuos. Idealmente cualquier individuo podría ser en un principio un determinante de una y solo una clase pura. Así después de rotar cada individuo debería cargar altamente sobre un solo factor o clase ideal.

5.6. Método Densidad

En aplicaciones donde conglomerados naturales son deseados, son usados métodos que buscan regiones de densidad alta comúnmente llamados modas. Conglomerados naturales usualmente sugieren que debería haber muchos puntos en el espacio que estén muy cercanos a otros puntos y que estos conglomerados están separados por una área con muy pocos puntos. La aproximación single linkage es usada en esta categoría. Una técnica popular es llamada análisis moda. Este método determina puntos densos los cuales son usados para definir conglomerados iniciales. Un radio r y un número de puntos k son seleccionados inicialmente. Alrededor de cada punto u objeto una esfera de radio r es determinada y el número de puntos k contenidos en la esfera es entonces calculado. Todos los puntos con al menos otros puntos contenidos en la esfera son llamados puntos densos. Los conglomerados iniciales son definidos por los puntos densos de tal manera que si un punto denso pertenece a más de un conglomerado el conglomerado relevante es combinado. Los conglomerados son también combinados si la distancia entre ellos es menor que un valor inicial c, el cual es el promedio de las $2k$ distancias más pequeñas entre los puntos originales n. Cualquier punto separado de todos los puntos densos por al menos r, forma su propio

conglomerado. Después de que la solución inicial ha sido determinada r puede ser incrementada y el proceso repetido.

El método taxmap también usa una aproximación single linkage. Empezando con una matriz de proximidad, los dos individuos más cercanos son seleccionados para formar el primer conglomerado. Una nueva matriz proximidad es calculada relacionando el conglomerado a los otros puntos. El punto más cercano al conglomerado es determinado nuevamente. La medida de proximidad promedio entre los tres es calculada y comparada a la medida de proximidad entre los primeros dos. La diferencia entre las dos medidas es referida a la medida de discontinuidad. Si esta medida es mas grande que algunos valores preasignados, entonces el individuo no es asociado al conglomerado y a un nuevo conglomerado es iniciado con el punto rechazado. El proceso es ahora repetido para los nuevos conglomerados.

5.7. Técnicas Clumping o Conglomerado Borroso (fuzzy)

Las técnicas clumping empiezan con una matriz de proximidad. Empezando con un nivel de preselección de proximidades p. Un conglomerado es formado encontrando los más grandes subconjuntos posibles en el cual todos los puntos son asociados a todos los otros puntos en el conjunto. Si uno de los conglomerados tiene puntos que son asociados a al menos k puntos en otro conglomerado, los dos conglomerados son combinados. El entero k es preseleccionado. En este método los conglomerados se superponerse a una cierta extensión.

5.8. Validación Conglomerada y Metodología del Análisis Conglomerado

La discusión de análisis conglomerado presentada en este capítulo sugiere que la metodología es exploratoria. El efecto del análisis depende de una extensa selección de técnicas, sobre las variables seleccionadas, y sobre la estructura conglomerada subyacente – si de hecho hay una. La siguiente cita de Milligan (1981) pone el escenario para la discusión presentada en esta sección.

> An inherent problem in the use of a clustering algorithm in practice is the difficulty of validating the resulting data partition. This is a particularly serious issue since virtually any clustering algorithm will produce partitions for any data set, even random noise data which contains no cluster structure. Thus, an applied researcher is often left in a quandary as to whether the obtained clustering of a real life data set actually represents significant cluster structure or an arbitrary partition of random data[1].

[1]Un problema inherente en el uso de un algoritmo jerárquico en la práctica es la dificultad de validación de la resultante partición de datos. Esta es un tema particularmente serio por la razón de que virtualmente cualquier algoritmo conglomerado producirá particiones para cualquier conjunto de datos, incluso datos discordantes aleatorios los cuales no contienen estructura conglomerada. Así, un investigador aplicado es frecuentemente dejado en una situación difícil en la decisión de cuando la obtención de un conglomerado de un conjunto de datos reales realmente representa una estructura conglomerada significante o una partición arbitraria de datos aleatorios. (Trad. del Autor)

5.8.1. Validación del Conglomerado

Para ser valido entonces, una solución conglomerada debería no ser una estructura que pudiera tener ocurrencia por casualidad de una muestra aleatoria de una población homogénea. La estructura debe ser inusual para ser valida. La suposición de aleatoriedad subyacente podría ser expresada de diferentes formas. El arreglo de los puntos en el espacio euclidiano es aleatorio, o la asignación de puntos en conglomerados es aleatorio, o finalmente, el rango de orden de las proximidades observadas es aleatorio. Las pruebas de aleatoriedad usualmente demandan uno de estos conceptos de aleatoriedad.

La validez de una estructura conglomerada puede ser examinada de diferentes formas. La medida criterio externo (external criteria) de la solución conglomerada contra información anterior concerniente a una estructura. La evaluación de algoritmos conglomerados usando muestras de conglomerados conocidos es un ejemplo de una evaluación externa de metodología conglomerada. Criterio interno (internal criteria) es usado para evaluar una solución conglomerada relativa a la matriz de datos y la matriz de proximidad correspondiente. Con el criterio interno el punto en cuestión es la bondad de ajuste del conglomerado solución relativa a la matriz de proximidad original. Un tercer criterio para la evaluación de una solución conglomerada es la replicalidad la cual involucra el uso de procedimientos validación cruzada. La comparación de resultados de muestras divididas (split half) podrían ser un ejemplo una evaluación replicada. Finalmente, una comparación de soluciones conglomeradas obtenidas de un algoritmo conglomerado alternativo aplicado a la misma matriz de datos, constituye lo que es usualmente referido a como la aplicación del criterio relativo. Este caso los índices de ajuste pueden ser calculados entre soluciones de conglomerados alternativos.

5.8.2. Medidas de recuperación de conglomerados y criterio de medida externa

Basado en un número de estudios, la medida recomendada de validación externa es el índice ajuste de Rand desarrollado por Hubert y Arabie (1985). Este índice, el cual es usado para comparar una solución conglomerada derivada a una solución conglomerada verdadera, esta dada por:

$$Ra = \frac{(a+b+n_c)}{(a+b+c+d+n_c)}$$

donde n_c es un ajuste para corregir posibles concordancias casuales. Los parámetros a,b,c y d están definido en la Figura 5.6 Las concordancias casuales,n_c, están dadas por:

$$n_c = \frac{[n(n^2+1) - (n+1)\sum n_{i\bullet} - (n+1)\sum n_{\bullet j} + {}^{2}\sum\sum_{n} n_{i\bullet}^2 n_{\bullet j}^2]}{[2(n-1)]}$$

Donde n_{ij} denota el número de puntos en el conglomerado i para la solución derivada la cual esta también en el cluster j de la solución verdadera.

5.9. Comentarios

El término de Cluster Análisis (usado primeramente por Tryon en 1939) incluye un número de algoritmos diferentes y métodos para agrupar objetos de tipo similar en categorías respectivamente. Una interrogante que en general

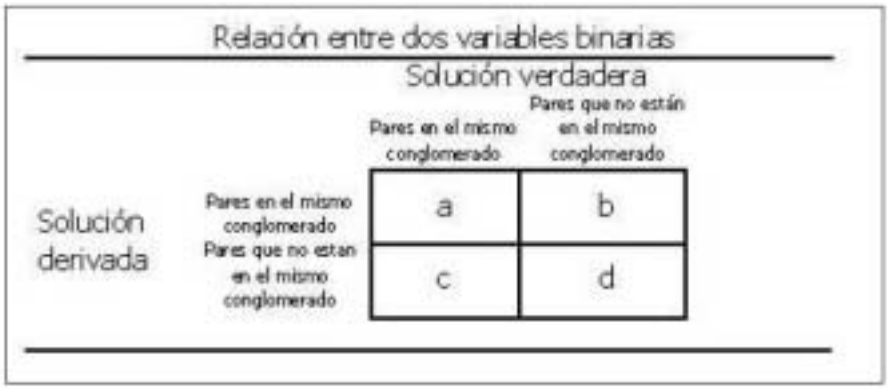


Figura 5.6: Comparación en la Localización de Pares de Puntos entre Soluciones Verdaderas y Derivadas

enfrentan investigadores en muchas áreas es como organizar datos observados en una estructura significativa, esto es, como desarrollar taxonomías. En otras palabras el Análisis de Conglomerados es una herramienta para el análisis de datos exploratorio el cual intenta colocar diferentes objetos en grupos de tal manera que el grado de asociación entre dos objetos sea máximo si pertenecen al mismo grupo y mínimo de otra forma. El análisis de conglomerados puede ser usado para descubrir estructuras en los datos sin proveer una interpretación/explicación. En otras palabras, el análisis de conglomerados simplemente descubre estructuras en los datos sin explicar por qué existen.El procedimiento básico es el siguiente

- Formulación del problema: seleccionar las variables a las que se desea aplicar la técnica de Conglomerados.
- Seleccionar una medida de distancia: distancia euclidiana cuadrada, distancia Manhatan, Chevychev, etc.
- Seleccionar un procedimiento de conglomerados.
- Decidir el número de conglomerados.
- Interpretación del conglomerado: sacar conclusiones, ilustrar la técnica usando mapas preceptúales y dendogramas son útiles.
- Evaluar fiabilidad y validez: repetir el análisis pero con una medida de distancia diferente, repetir el análisis pero usar diferente técnica de conglomerados, dividir los datos aleatoriamente y analizar cada parte separadamente, repetir el análisis varias veces eliminando una variable por vez, repetir el análisis varias veces usando un orden diferente cada vez.

Capítulo 6

Escalamiento Multidimensionales MDS

Escalas multidimensionales (MDS) se ha convertido en una técnica popular del análisis multivariado y exploratorio. MSD es un conjunto de métodos de análisis de datos, el cual permite a uno inferir sobre las dimensiones del espacio de forma intuitiva de los sub-objetos. Los datos son típicamente una medida de similaridad o disimilaridad de los objetos. El resultado primario es una configuración espacial, en el cual los objetos son representados como puntos. Los puntos en esta representación espacial son arreglados de tal manera que sus distancias corresponden a las similaridades de los objetos: objetos similares son representados mediante puntos se encuentran más cercanos, objetos disimilares mediante puntos que están alejados.

Roger Shepard (1979) describe el escalamiento multidimensional como: la idea de representar objetos (como lo son colores, sonidos, formas, caras, significados de palabras, etc.) como puntos en el espacio, de tal forma que las distancias entre los puntos representan las similaridades percibidas entre los objetos; a encontrado una aplicación amplia en las ciencias cognitivas, del comportamiento y biomédicas de igual manera. Joseph Kruskal (1979) menciona que aparte de la psicología, MDS tiene diversas aplicaciones, que incluyen el arreglo de las macromoleculas de las cuales están constituidos lo ribosoma, y las relaciones entre especies basados en reacciones de la cera. Aunque el ejemplo por excelencia del uso de Escalas Multidimensionales sea la reconstrucción de mapas de ciudades, mediante el uso estimado de tiempos de viaje, respetando el orden de recorrido. Greenacre y Underhill (1982) usaron el tiempo de vuelo entre aeropuertos situados al Sur de Africa; Mardia et. al. (1979) uso distancias entre carreteras de algunas ciudades de Inglaterra. Estos son algunos ejemplos que contemplan el uso de esta técnica del análisis multivariado.

En este capítulo se describen diferentes técnicas de MDS para analizar datos y así como su la interpretación de resultados. Se hará la distinción entre MDS métrica y no métrica. Se desarrollara la técnica de reducción de datos, bajo ciertos supuestos sobre las medidas de proximidad de tal manera que sea posible determinar dimensiones que sean consistentes con las proximidades dadas. Por último se dará una muy breve explicación de los algoritmos ASCAL y PROXSCAL.

6.1. Tipos de Escalas Multidimensionales MDS

Lo que se llama genéricamente MDS puede ser considerado como una gran familia de procedimientos del análisis multivariado, aplicables a múltiples tipos de datos y diferentes niveles de medida, y que tienen distintos presupuestos de partida y diversas finalidades. La distinción entre las principales variantes de MDS puede hacerse en función de unos cuantos factores. De éstos, los más relevantes son los siguientes:

1. *Tipos de datos de entrada.* Representa uno de los factores más importantes a la hora de decidirse por el uso de un tipo de MDS u otro. Existen distintas posibilidades de análisis dependiendo de la forma concreta en que se encuentren organizados los datos, de la escala en que vengan medidos éstos, así como de determinados supuestos que se hagan sobre ellos.

2. *Números de matrices de proximidad empleados.* Dependiendo fundamentalmente del método de recogida de datos empleados, podemos encontrarnos con situaciones variadas en cuanto al número de matrices de proximidad que se pueden emplear: desde una sola matriz de entrada para todos los sujetos hasta una matriz de proximidades por sujeto, o una matriz de proximidades por grupo. En los dos últimos casos es posible tratar cada matriz como una fuente de datos diferente, o también como replicaciones de una misma fuente de datos.

3. *Modelo de escalamiento empleado.* Este factor está fuertemente relacionado con los dos anteriores. Por un lado, en cuanto al tipo de datos de entrada, existe una distinción fundamental entre los modelos que hacen uso de datos medidos en escala de intervalo o de razón (llamados modelos métricos) frente a los que utilizan datos medidos en escala ordinal (llamados modelos no-métricos). Por otro lado, en cuanto al número de matrices de entrada, se tienen cuatro tipos de modelos diferentes:

   - aquellos que utilizan una sola matriz de proximidades como entrada (lo que se conoce como MDS Clásico o CDMS);
   - aquellos que utilizan varias matrices como entrada, pero las tratan como replicaciones de una misma fuente de datos (conocido como MDS replicado o RMDS)
   - los que utilizan varias matrices de entrada y las tratan como representaciones de un mismo espacio pero ponderadas de forma diferente (conocido como MDS de diferencias individuales, INDSCAL, o MDS ponderado)
   - los que utilizan varias matrices de entrada y las tratan como representaciones de un mismo espacio pero ponderadas y rotadas de forma diferente (conocido como modelo euclídeo generalizado o GEMSCAL)

Finalmente, también existe una distinción entre aquellos modelos que toman como entrada una matriz de proximidades simétrica (donde las proximidad entre el estímulo a y el estímulo b es la misma que existe entre el estímulo b y el estímulo a), y aquellos que permiten tratar con una matriz de proximidades asimétricas (modelo ASCAL) o varias matrices de proximidades asimétricas (modelo AINDS).

6.2. Modelos de Escalamiento Multidimensional

Los distintos modelos de MDS dependen fundamentalmente de la relación asumida entre los datos de entrada (las proximidades) y la distancias entre

estímulos obtenidas como solución. La asunción básica de cualquier modelo de escalamiento es que las distancias son función de las proximidades. Formalmente:

$$\delta_{ij} = f(d_{ij})$$

La elección de la transformación efectuada (f) determina el modelo de escalamiento a utilizar. Aunque f puede representar una función potencial, exponencial, lineal o de cualquier otro tipo, existen dos casos importantes:

1. Cuando se asume que f es una función lineal con pendiente positiva. En ese caso la relación entre proximidades y distancias es del tipo: $\delta_{ij} \rightarrow a+b\delta_{ij} = d_{ij}$. El modelo que plantea este tipo de relación entre proximidades y distancias es el llamado modelo *métrico*.

2. Cuando se asume que f es una función monótona creciente. En este caso la relación entre proximidades y distancia es mucho más laxa: si $\delta_{ij} < \delta_{kl}$, entonces $d_{ij} \leq d_{kl}$. Es el llamado modelo *no-métrico*.

La utilización de uno u otro modelo no es en absoluto arbitraria, sino que viene determinada, en la mayoría de los casos, por la escala de medida de los datos. Si la escala es de tipo intervalo o razón, el modelo a aplicar es el métrico. Si los datos están medidos en una escala ordinal, son datos binarios o frecuencias, el modelo a utilizar será el no-métrico.

En cualquier caso el uso de un modelo no-métrico, a diferencia de lo que ocurre con otros análisis, no supone ningún tipo de desventaja si se utiliza un número de estímulos razonablemente elevado. Se ha comprobado en múltiples ocasiones que el escalamiento no métrico (rasgos) de datos métricos (proximidades medidas en una escala de intervalo o razón) ofrece soluciones casi idénticas a las proporcionadas por el modelo de escalamiento métrico. La razón de esta aparente paradoja es que, a pesar de la pérdida de información que supone transformar datos medidos originalmente en escala de intervalo o razón en datos en escala ordinal, la representación espacial impone grandes restricciones a las soluciones posibles. El número de distancias que es necesario estimar para n estimulos es de

$$\begin{pmatrix} 2 \\ n \end{pmatrix}$$

Sin embargo, cada una de estas distancias puede ser comparada con cualquier otra. Por tanto, el número de relaciones posibles entre distancias es de

$$\begin{pmatrix} \begin{pmatrix} n \\ 2 \end{pmatrix} \\ 2 \end{pmatrix}$$

Dado un número tan elevado de relaciones posibles, aún siendo éstas de tipo ordinal, será muy difícil encontrar muchas disposiciones alternativas de los estímulos que satisfagan todos sus requerimientos. Por ello, la solución final está muy determinada a pesar de que los datos de partida no sean métricos.

6.3. Escalamiento multidimensional

Formalmente Escalamiento Multidimensional (MDS) es un grupo de técnicas que usa proximidades entre objetos para producir una representación espacial de los objetos. La matriz de proximidad es usualmente una matriz de disimilaridad. La representación espacial derivada consiste en una configuración geométrica de puntos en un mapa, cada punto corresponde a uno de los objetos. Entre mas

grande sea la similaridad entre los objetos mas cerca los objetos estarán en el mapa. Una matriz de proximidad consistente en distancias entre ciudades, en particular puede ser usado par construir una representación espacial preservando las distancias. De forma análoga al mapa de ciudades, en muchas aplicaciones las medidas de proximidad usadas, para relacionar los objetos, no están basadas en una medida directa. En su lugar las medidas de proximidad están basadas en percepciones de similaridad derivadas del criterio humano.

Escalamiento multidimensional el cual esta basado sobre medidas de proximidad es conocido como métrica MDS, mientras que MDS no-métrico es usado para caracterizar la técnica cuando las proximidades esta basada en el sentido común. La representación espacial de la métrica MDS intenta preservar las distancias entre los objetos, mientras que la representación espacial MDS no-métrica únicamente preserva el rango de orden entre las disimilaridades. De esta manera en la MDS no-métrica, si los objetos (AyB) se perciben mas cercanos que (AyC) y (ByC), entonces la configuración espacial preservara esta similaridad jerárquica. En la configuración geométrica derivada la distancia entre los puntos (AyB) puede ser menor que las distancia entre (AyC) y (ByC). En el caso de las similaridades percibidas entre automóviles, AyB puede representar carros compactos producidos por dos compañías Norteamericanas, mientras que el automóvil C puede representar un carro compacto producido en Europa.

Una vez que las dimensiones o escalas han sido determinadas el segundo paso del análisis involucra la interpretación de los resultados. Gráficas de dispersión mostrando la localización de los objetos con respecto de las dimensiones derivadas son útiles porque proveen una representación gráfica de las relaciones de disimilaridad. Si las medidas de las características de los objetos están disponibles y se crea que pueden contribuir a la percepción de la disimilaridad, otros análisis pueden utilizarse para la interpretación. Las dimensiones derivadas pueden asociarse a las características medidas usando otras técnicas multivariadas. En un estudio de automóviles, características como son: economía de combustible, tamaño, estilo y lujo, pueden asociarse a escalas.

Si la matriz de disimilaridad esta basada en una media numérica de matrices de disimilaridad de una muestra de individuos, entonces las diferencias individuales pueden también ser estudiadas. Una vez que las dimensiones derivadas han sido determinadas para el total de la muestra, un conjunto de coordenadas puede ser determinado para localizar los puntos para cada individuo. Las diferencias entre individuos son manejadas mediante la asignación de pesos individuales a las diferentes dimensiones. La explicación para la aplicación del escalamiento de diferencias individuales es que las diferencias individuales son atribuidas a las diferencias en importancia que los individuos anexan a las varias escalas comunes. Las escalas subyacentes son constantes para los individuos.

Similar al análisis de conglomerados, escalas multidimensionales es una técnica de análisis de datos exploratorio. Análisis de conglomerados busca la clasificación de objetos en grupos usando medidas de disimilaridad derivadas de las medidas observadas, por otra parte, escalas multidimensionales buscan la determinación de dimensiones subyacentes que contribuyan a percibir diferencias entre objetos. Como en el caso de componentes principales y análisis de factores, en escalas multidimensionales su interés esta basado en el entendimiento de las dimensiones subyacentes que contribuyen a diferenciar los objetos. Análisis factorial usa medidas obtenidas de los objetos sobre dimensiones conocidas, mientras que escalas multidimensionales usa medidas de disimilaridad de los objetos sobre todo para derivar dimensiones subyacentes.

6.4. Escalamiento Multidimensional Métrico

La métrica de escalas multidimensionales empieza con una matriz D de disimilaridades de orden,$(n \times n)$, δ_{rs}, $r, s = 1, 2, \ldots, n$ la cual determina una medida de disimilaridad para todos los posibles pares de n objetos. Los elementos de la diagonal de D son ceros. El objetivo de la métrica MDS es el de definir a un conjunto de p dimensiones subyacentes definidas por las medidas $x_1, x_2, x_3, \ldots, x_p$ tal que:

a) Las coordenadas de n objetos a lo largo de las dimensiones derivadas p caen en una matriz de distancia Euclidiana y

b) Los elementos de la matriz de distancia euclidiana son equivalentes a, o próximamente cercana, los elementos δ_{rs} de D.

En contraste a la mayoría de las técnicas multivariadas en escalas multidimensionales, la matriz X de observaciones de orden $(n \times p)$ se deriva de la matriz D de disimilaridades.

6.4.1. Construcción de una Matriz Semidefinida Positiva Basada en D

Una matriz de disimilaridad D con ceros en la diagonal principal es no semidefinida positiva. Una matriz definida positiva A de orden $(n \times n)$, después de todo, puede ser construida basada en los elementos de δ_{rs}. Los elementos a_{rs} de la matriz nueva A pueden ser determinados usando la relación:

$$a_{rs} = -\frac{1}{2}[\delta_{rs}^2 - \delta_{r\bullet}^{-}\delta_{\bullet s}^2 + \delta_{\bullet\bullet}^2]$$

$$r, s = 1, 2, \ldots, n$$

donde

- $\delta_{r\bullet}^2 = \frac{1}{2}\sum_{s=1}^{n}\delta_{rs}^2$ Distancia cuadrática media por fila
- $\delta_{r\bullet}^2 = \frac{1}{2}\sum_{s=1}^{n}\delta_{rs}^2$ Distancia cuadrática media por columna
- $\delta_{r\bullet}^2 = \frac{1}{2}\sum_{s=1}^{n}\delta_{rs}^2$ Distancia cuadrática media de la matriz

En notación matricial la relación esta dada por:

$$A = -\frac{1}{2}[I_n - \frac{1}{2}i_n i_n^u]D^2[I_n - \frac{1}{n}i_n i_n]$$

Donde I_n es una matriz identidad de orden $(n \times n)$, i_n es un vector de unidades de orden $(n \times 1)$ y D^2 es la matriz cuyos elementos son los cuadrados de los elementos de D.La matriz $[I_n - \frac{1}{n}i_n\acute{i}_n]$ es llamada matriz centrada. La matriz A ha sido derivada de la matriz D^2 mediante la obtención de la media por filas y columnas de D^2 (doble media central). Las filas y columnas de por esta razón suman cero y de aquí que el rango de A es a lo mas $(n-1)$.

6.4.2. El teorema fundamental de MDS

La matriz D de disimilaridades δ_{rs}^2, δ_{rs}, $r, s = 1, 2, \ldots, n$ de orden $(n \times n)$ se dice euclidiana si existe una dimensión p y un conjunto de n puntos dados por $(1 \times p)$ de vectores $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ tales que:

$$\delta_{rs}^2 = (\bar{x}_r - \bar{x}_s)^t(\bar{x}_r - \bar{x}_s)$$

con

$$r, s = 1, 2, \ldots, n$$

En otras palabras si las observaciones de X son conocidas, D será la matriz de distancia euclidiana derivadas de $X(n \times p) = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ El teorema fundamental de MDS indica que la matriz de disimilaridad dada D es euclidiana $\Leftrightarrow$ la matriz definida por $a_{rs} = -\frac{1}{2}[\delta^2_{rs} - \delta^-_{r\bullet}\delta^2_{\bullet s} + \delta^2_{\bullet\bullet}]$, es semidefinida positiva. Este teorema fundamental provee la llave para obtener la solución de MDS para una matriz D de disimilaridades. Si la matriz D es euclidiana, entonces la matriz A puede ser escrita como $A = X^*\acute{X}^*$, donde X es la matriz de orden $(n \times p)$ que consiste en las coordenadas de los n puntos en un espacio de p dimensiones y X^* es la matriz de columnas media central de X. De aquí que no haya pérdida de generalidad al asumir que las pX variables tienen media cero. Se asumirá que las pX variables tienen media cero.

6.4.3. La solución MDS

Dada una matriz de disimilaridad D, la matriz A es construida usando:

$$a_{rs} = -\frac{1}{2}[\delta^2_{rs} - \delta^-_{r\bullet}\delta^2_{\bullet s} + \delta^2_{\bullet\bullet}]$$

$$r, s = 1, 2, \ldots, n$$

Los eigenvectores $v_1, v_2, v_3, \ldots, v_n$ y los correspondientes eigenvalores λ_1, λ_2, $\ldots$, λ_n de la matriz A son usados para obtener las medidas subyacentes $x_1, x_2, \ldots, x_n$. Si A es una matriz semidefinida positiva de rango p, entonces p de los eigenvalores de A serán positivos y los restantes $(n-p)$ eigenvalores serán cero. La matriz A se puede expresar como $A = V \wedge \acute{V}$, donde V es la matriz de eigenvectores de A y $\wedge$ es la matriz diagonal de eigenvalores.

El número de eigenvalores positivos permite la determinación de p mientras que la ausencia de cualquier eigenvalor negativo sustenta a la matriz semidefinida positiva necesaria. De esta manera dada la matriz D y el rango de A no se especifica de antemano, pero esta determinado por el número de eigenvalores positivos. Para los p eigenvalores no cero las coordenadas de X pueden ser definidas por:

$$x_j = v_j \sqrt{\lambda_j}$$

con

$$j = 1, 2, \ldots, p$$

donde se asume que $\bar{v}^t_j v_j = 1$. Equivalentemente

$$X(n \times p) = V \wedge^{\frac{1}{2}}$$

Las filas de X, dadas por $x^t_1, x^t_2, \ldots, x^t_n$ tienen la propiedad de que

$$\delta^2_{rs} = (\bar{x}_r - \bar{x}_s)^t(\bar{x}_r - \bar{x}_s)$$

con

$$r, s = 1, 2, \ldots, n$$

y por esta razón las relaciones de disimilaridad en D son preservadas por la solución escalar dada por X. Las pX variables o dimensiones (escalas) tienen media cero y son únicas solo por una constante.

6.4.4. Una solución aproximada

En la practica el objetivo es obtener un número pequeño de dimensiones digamos $k \leq p$ tal que la relaciones de disimilaridad derivadas son aproximadamente iguales a la matriz original D. Una aproximación común es retener los primeros r eigenvectores y los eigenvalores correspondientes de tal manera que A es aproximada por $\hat{A} = V_0 \wedge_0 V_0^t$, donde $\wedge_0$ y V_0 denotan submatrices de $\wedge$ y V correspondiendo a los r eigenvalores mas grandes y a los respectivos eigenvectores correspondientes. Los valores escalares correspondientes están dados por $X^{(0)} = V_0 \wedge_0^{\frac{1}{2}}$ y las distancias resultantes están dadas por

$$d_{rs}^{(0)^2} = (\bar{x}_r^{(0)} - \bar{x}_s^{(0)})^t (\bar{x}_r^{(0)} - \bar{x}_s^{(0)})$$

Si los primeros k eigenvalores cuentan para la mayoría de las variaciones en A entonces las aproximaciones de δ_{rs}^2 mediante $d_{rs}^{(0)^2}$ deberá ser buena.

Una medida útil de bondad de ajuste esta basada en el cuadrado de la correlación de Person RSQ entre δ_{rs}^2 y $d_{rs}^{(0)^2}$,$r, s = 1, 2, \ldots, n$. Este valor debería estar cercano a 1 para asegurar una razonable ajuste. El índice RSQ es una correlación cuadrática entre las disparidades derivadas por el modelo de escalamiento, de modo que puede ser interpretado como la proporción de varianza en las disparidades que es explicada por las distancias.

$$RSQ = \frac{[\sum_r \sum_s (\delta_{rs} - \delta_{\bullet\bullet})(\hat{\delta}_{rs} - \hat{\delta}_{\bullet\bullet})]}{[\sum_r \sum_s (\delta_{rs} - \delta_{\bullet\bullet})^2][\sum_r \sum_s (\hat{\delta}_{rs} - \hat{\delta}_{\bullet\bullet})^2]}$$

Otras medidas de bondad de ajuste son llamadas STRESS y S-STRESS. La función de bondad de ajuste para evaluar cuánto se aproximan las distancias obtenidas a partir de X a las disparidades obtenidas de la transformación de esas distancias es conocida como STRESS:

$$S = \sqrt{\frac{\sum_r \sum_s (\delta_{rs}^2 - \hat{\delta}_{rs})}{\sum_r \sum_s \hat{\delta}_{rs}^2}}$$

Cuanto mayor sea el valor de STRESS, mejor será el ajuste encontrado entre distancias y disparidades. Es decir, el STRESS no es propiamente un índice de bondad de ajuste, más bien de "maldad" de ajuste. Su valor mínimo se encontrará en 0, cuando no exista diferencia entre distancias y disparidades.

Puesto que se parte de una matriz de coordenadas aleatorias el ajuste nunca es muy bueno al principio. Por ello, es necesario llevar a cabo el proceso iterativamente para lograr que se minimice el valor de STRESS. Esto se consigue alterando los valores de las coordenadas de la matriz X de modo que la diferencia entre las distancias y disparidades derivadas a partir de ellos sea más pequeña ahora que en el paso anterior.

Para el algoritmo de convergencia se utiliza otra función de STRESS, conocida como S-STRESS, cuya fórmula es:

$$S - STRESS = \sqrt{\frac{\sum_r \sum_s (\delta_{rs}^2 - \hat{\delta}_{rs})}{\sum_r \sum_s \hat{\delta}_{rs}^2}}$$

La evaluación del ajuste proporcionado por el valor de STRESS para una solución determinada debe hacerse teniendo en cuenta varios factores:

1. El valor de STRESS suele ser más alto cuanto mayor sea el número de estímulos, debido a que cuando tenemos pocos estímulos, el número de proximidades a ajustar en la solución será también pequeño, pero a medida que aumenta el número de estímulos, el número de proximidades a ajustarse se incrementa rápidamente.

2. El valor de STRESS será siempre más alto para soluciones de menor dimensionalidad, e irá bajando a medida que la solución contenga un mayor número de dimensiones. Cuando el número de dimensiones es igual al número de estímulos menos $2(n-2)$, el ajuste será siempre perfecto. El objetivo en este caso será buscar un valor suficientemente bajo de STRESS (buen ajuste) unido a una dimensionalidad también baja (representación parsimoniosa de los datos).

3. Cuando se permite desempatar los empates presentes en los datos, el valor de STRESS obtenido suele ser menor que el obtenido cuando no se permite desempatar.

6.4.5. Métrica Escala Multidimensional Empezando con D

En general la matriz de disimilaridad original D es dada y la matriz subyacente X no esta disponible. Las dimensiones derivadas en MDS son entonces interpretadas usando otra información. Por ejemplo, tenemos que proceder como si la matriz subyacente X no estuviera disponible cuando se generan las dimensiones. La matriz original X fue entonces usada para relacionar las dimensiones derivadas de las variables originales. Este ejemplo ha sido usado por conveniencia y también para permitir comparaciones a otras técnicas como son análisis de conglomerados y análisis de componentes principales.

6.4.6. Mejorando la Solución

Es posible en este paso el mejorar la representación en dos dimensiones mediante la corrección de coordenadas. Una aproximación podría ser un procedimiento aproximatorio numérico para corregir las coordenadas, para crear las distancias cercanas derivadas de las distancias originales. Es posible llevar acabo esto mediante la determinación de coordenadas corregidas en las primeras dos dimensiones tal que $\sum_{r<s}\sum[d_{rs}(2)-\delta_{rs}]^2$ sea mínima. Los procedimientos numéricos como lo son Newton-Raphson o el procedimiento steepest descent son usados frecuentemente para obtener coordenadas corregidas y valores corregidos de $d_{rs}(2)$,$\hat{d}_{rs}(2)$ tal que la suma de cuadrados anterior se minimiza. En algunas instancias un procedimiento iterativo que es puesto en práctica revisa ambas $d_{rs}(2)$ y las coordenadas en una serie de pasos alternativos.

6.4.7. Usando Similaridades

Si la matriz proximidad es una matriz de similaridades (por ejemplo un coeficiente de correlación[1]) C con elementos c_{rs} que satisfacen $c_{rs}=1$, $c_{rs}=c_{sr}$,$0\leq c_{rs}\leq 1$,$r,s=1,2,\ldots,n$, la matriz puede ser transformada a una matriz de disimilaridad usando la expresión $\delta_{rs}^2=(2-2c_{rs})$. Esta fue la relación obtenida entre distancias euclidianas cuadradas y la correlación para variables estandarizadas.

6.5. Escalamiento Multidimensional No-métrico

En MDS no métrico una matriz D de disimilaridades δ_{rs} es frecuentemente derivado de respuestas a cuestionarios o procedimientos experimentales. Las respuestas o tema (el punto en cuestión) son usualmente requeridos para hacer

[1]Se definió una propiedad para que el coeficiente de correlación que esta entre $-1<r<1$ cumpla con estas propiedades, en la sección que correspondiente a medidas de proximidad.

comparaciones entre conjuntos de objetos o estímulos. El propósito del análisis escalamiento no métrico es el obtener una compresión dentro de la naturaleza de las disimilaridades percibidas. Tales análisis han sido usados para medir actitudes y preferencias en: legislación, política y sociología; para hacer comparaciones culturales en antropología; para estudiar percepciones humanas en psicología y lingüística; y para evaluar diseño de productos; la mercadotecnia ha proveído mucha de la literatura de investigación sobre técnicas de escalamiento y sobre el diseño experimental requerido.

En el escalamiento multidimensional no-métrico muchos de los esfuerzos son dirigidos al diseño de procedimientos de experimentos para medir disimilaridad. Una lista parcial de los tipos de procedimientos son:

a) Comparaciones apareadas: a los temas (o puntos en cuestión) se les requiere la comparación de todos los pares posibles de un conjunto de objetos y la estimación del grado de similaridad.

b) Particiones: a los temas se les requiere la división de conjuntos de objetos en un número pequeño o categorías mutuamente exclusivas.

c) Posiciones o grados: a los temas se les requiere el rango de los objetos con respecto a un criterio específico.

d) Comparaciones Triadas: a los temas se les requiere una posición de grado de similaridad entre tres posibles apareamientos de conjunto de tres objetos.

e) Tetrads: a los temas se les requiere el comparar todos los posibles pares de objetos y el indicar el más similar y/o los pares mas disimilares.

6.5.1. Escalamiento ordinal

En MDS no métrica, las disimilaridades δ_{rs} dadas son usadas para generar un conjunto de distancias derivadas d_{rs}, las cuales son relacionadas aproximadamente a las disimilaridades δ_{rs} dadas por una función f monótona creciente. En este caso lo escribimos como: $d_{rs} \approx f(\delta_{rs})$ en donde f es una función con la propiedad de:

$$\delta_{rs} < \delta_{uv} \Leftrightarrow f(\delta_{rs}) < f(\delta_{uv})$$

.

El grado de correlación entre δ_{rs} y $f(\delta_{rs})$ en unitaria, mientras que el grado de correlación entre δ_{rs} y d_{rs} es cercana a 1. Una gráfica de d_{rs} contra δ_{rs} tendría que estar muy cercana a una función monótona creciente. Porque únicamente el rango de orden es importante en la escala ordinal y el llamada escalamiento no métrico. Una aproximación para determinar los elementos d_{rs} y las configuraciones subyacentes es un proceso iterativo conocido como *algoritmo Shepard Kruskal*.

6.5.2. Algoritmo Shepard Kruskal

El algoritmo Shepard Kruskal para MDS no métrica es mostrado en la figura 6.1 Después de determinar la matriz de disimilaridad D y la correspondiente matriz escalar A usando:

$$a_{rs} = -\frac{1}{2}[\delta_{rs}^{-}\delta_{r\bullet}^{2} - \delta_{\bullet s}^{2} + \delta_{\bullet\bullet}^{2}]$$

$$r, s = 1, 2, \ldots, n$$

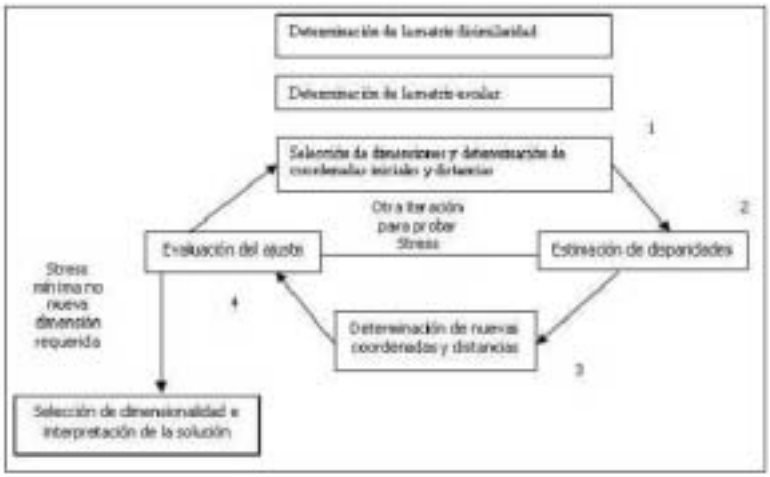


Figura 6.1: Algoritmo Shepard – Kruskal para Escalamiento No-Métrico

un proceso iterativo que revisa las disimilaridades y las coordenadas de los objetos hasta que es logrado un ajuste adecuado. El objetivo del proceso iterativo es el de obtener una representación espacial en una dada, tal que las distancias euclidianas son relacionadas monótonamente a las disimilaridades originales.

La parte iterativa del proceso contiene cuatro fases. La primera fase o *fase inicial* selecciona las p dimensiones y determina la configuración inicial $X^{(0)}$ y las distancias resultantes $d_{rs}^{(0)}$. La segunda fase o fase no métrica usa una regresión monótona para relacionar $d_{rs}^{(0)}$ y δ_{rs}. La regresión estimada produce una nuevo conjunto de seudo disimilaridades $\hat{\delta}_{rs}^{(0)}$ llamadas disparidades que son relacionadas monótonamente a δ_{rs}. La tercera fase del proceso llamada fase métrica revisa la configuración espacial para obtener $X^{(1)}$ para poder obtener las nuevas distancias $d_{rs}^{(1)}$ las cuales están más relacionadas a las disparidades $\hat{d}_{rs}^{(0)}$. Si el ajuste no es adecuado las fases 2 y 3 son repetidas. Para la repetición de la fase 2, las distancias $d_{rs}^{(1)}$ están relacionadas a las disimilaridades δ_{rs} originales usando una regresión monótona para generar una nueva representación espacial $X^{(2)}$ y las nuevas distancias $d_{rs}^{(2)}$. La fase evaluación entonces compara $\hat{d}_{rs}^{(1)}$ y $d_{rs}^{(2)}$. Finalmente después de obtener soluciones sobre un rango de dimensiones una solución dimensional es seleccionada. Esta solución entonces será interpretada. Esta etapa será referida como la selección y fase de interpretación. Una presentación de las técnicas involucradas en los pasos 2, 3, 4 y 5 es presentada a continuación.

6.5.3. Fase No-métrica y la Regresión Monótona

En la fase no métrica, disparidades $\hat{d}_{rs}^{(0)}$ son determinadas por medio de las distancias $d_{rs}^{(0)}$ de tal forma que las $\hat{d}_{rs}^{(0)}$ son relacionadas monótonamente a la disimilaridad original δ_{rs}. Las $\hat{d}_{rs}^{(0)}$ son el resultado de una regresión de las $d_{rs}^{(0)}$ sobre los δ_{rs} sujetas a la condición de que la relación de ajuste es monótona. Esta regresión es por lo tanto llamada *regresión monótona*. Un método de aproximación sucesiva útil para obtener una regresión estimada, es llamado el algoritmo pool-adjacent violators el cual se explica a continuación.

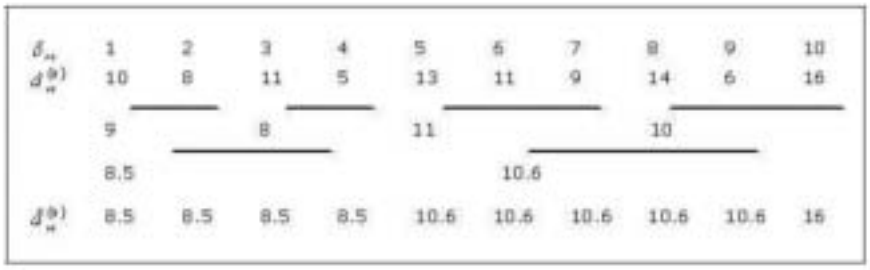


Figura 6.2: Ejemplo de Bloque de Violaciones

6.5.4. Algoritmo de Violación Adyacente

Esta aproximación para determinar las disparidades $\hat{d}_{rs}^{(0)}$ de $d_{rs}^{(0)}$ y δ_{rs} empieza por clasificar los valores δ_{rs} desde el más bajo al mas alto antes de compararlas a los valores correspondientes $d_{rs}^{(0)}$. Empezando con el que ocupa la posición mas baja de valores de δ_{rs}, los valores adyacentes $d_{rs}^{(0)}$ son comparados a cada δ_{rs} para determinar si están relacionados monótonamente a los $d_{rs}^{(0)}$. Mientras que $d_{rs}^{(0)} < d_{uv}^{(0)}$ cuando $\delta_{rs} < \delta_{uv}$ entonces $\hat{d}_{rs}^{(0)} = d_{rs}^{(0)}$. Cuando sea que un bloque de valores consecutivos de $d_{rs}^{(0)}$ sean encontrados que violan la propiedad requerida de monotonicidad los valores $d_{rs}^{(0)}$ son promediados con el mas reciente valor $d_{rs}^{(0)}$ no violado para obtener un estimador $\hat{d}_{rs}^{(0)}$. Este valor $\hat{d}_{rs}^{(0)}$ es entonces asignado a todos los puntos en un bloque particular. Este procedimiento es ilustrado en la figura6.2.

En este ejemplo los bloques de violaciones que están subrayados son promediados para obtener estimaciones de disparidades iniciales. Las nuevas estimaciones son revisadas en monotonicidad. Si los bloques de violaciones continúan el proceso de promediar continua. En el ejemplo anterior, dos pasos son requeridos para obtener un conjunto monótono de disparidades. En este caso las disparidades resultantes son constantes para cuatro de cinco disimilaridades consecutivas.

6.5.5. Uniones y Tipos de Monotonicidad

La condición de $\hat{d}_{rs}^{(0)} \leq \hat{d}_{uv}^{(0)}$ si $\delta_{rs} < \delta_{uv}$ es llamado monotonicidad débil mientras que la condición $\hat{d}_{rs}^{(0)} < \hat{d}_{uv}^{(0)}$ si $\delta_{rs} < \delta_{uv}$ es llamado monotonicidad fuerte. Hay que notar que las disparidades determinadas en el ejemplo anterior satisface los requerimientos de monotonicidad débil pero no los requerimientos de monotonicidad fuerte.

6.6. Tecnicas Auxiliares: Análisis de Conglomerados

Se hablo previamente, como empezando con una matriz de proximidad los objetos pueden ser conglomerados en una manera jerárquica y que los resultados pueden ser usado para obtener las proximidades derivadas entre los objetos. Estas proximidades derivadas pudieran entonces ser relacionadas a las proximidades originales para evaluar los procedimientos conglomerados.

Un análisis de conglomerado concentra sobre un ajuste riguroso las disimilaridades pequeñas o similaridades. En los primeros pasos del proceso jerárquico la proximidad grupal esta cercana a las proximidades originales. Cuando el conglomerado crece en tamaño por otro lado las proximidades grupales son mucho

menos comparables a las proximidades originales. De esta manera un análisis conglomerado jerárquico no es apto de proveer proximidades constantes a lo largo de la escala final. En contraste un proceso escalar como lo es componentes principales es apto para concentrarse sobre las disimilaridades grandes y da un trabajo muy pobre al ajustar las disimilaridades pequeñas. Por esta razón es muy útil el combinar los resultados obtenidos de un análisis de conglomerados y un análisis de escalas multidimensionales en la representación espacial de los objetos. Los resultados del procedimiento conglomerado jerárquico pueden mostrarse en un esquema, de esta manera se esta confirmando la proximidad de varios objetos.

Un procedimiento de gran utilidad para interpretar las soluciones MDS es buscar los agrupamientos de estímulos. Estos agrupamientos indicarán conjuntos de estímulos muy semejantes entre sí y diferentes a los demás, y pueden ser de utilidad si la finalidad principal del análisis es la clasificación. El análisis de conglomerados genera agrupamientos jerárquicos de los estímulos en función de su proximidad. Esta grupación se hace de tal modo que aquellos estímulos más similares entre sí formarán parte de un mismo conglomerado. A medida que la proximidad vaya disminuyendo, otros estímulos u otros conglomerados se irán uniendo a esta estructura jerárquica hasta que, finalmente, todos los estímulos pertenecerán a un único conglomerado. Si complementaros la información proporcionada por MDS con la información sobre agrupamiento proporcionada por el análisis de conglomerados, nos será más sencillo identificar grupos de estímulos con características semejantes, así como el número de grupos que existen.

6.7. Relación con el Análisis de Componentes Principales

Como se ha mencionado, si la matriz X de orden $(n \times p)$ se conoce, puede ser aproximada por un número pequeño de componentes principales $k \leq p$. Los componentes pueden ser usados para producir esquemas mostrando la relación entre los n objetos en X. En el contexto de escalas multidimensionales la matriz X es desconocida. Unicamente la matriz de disimilaridades D esta disponible. Si X es conocida entonces no es necesario generar la matriz A para determinar X.

6.7.1. Métrica MDS y Análisis de Componentes Principales

El análisis de coordenadas principales usa una matriz S de similaridad dada de orden $(n \times n)$ para derivar una representación espacial de n objetos. (S pudiera ser una matriz de covarianzas entre los n objetos u otro tipo de matriz XX'). Denotando los elementos de S por s_{rs}, $r, s = 1, 2, \ldots, n$, una nueva matriz C es obtenida calculando:

$$c_{rs} = s_{rs} - \bar{s}_{r\bullet} - \bar{s}_{\bullet s} + \bar{s}_{\bullet\bullet}$$

donde

$$\bar{s}_{r\bullet} = \frac{1}{2}\sum_{s=1}^{n} s_{rs}$$

$$\bar{s}_{\bullet s} = \frac{1}{n}\sum_{s=1}^{n} s_{rs}$$

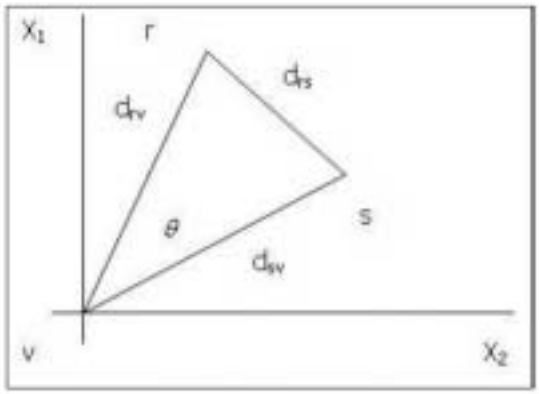


Figura 6.3: Relación de la ley del coseno entre tres puntos

y

$$\bar{s}_{\bullet\bullet} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} s_{rs}$$

Equivalentemente $C = (I - i_n\acute{i}_n)S(I - i_n\acute{i}_n)$ como en el caso de la construcción de una matriz semidefinida positiva basada en D. Análisis de componentes principales es entonces aplicado a C para determinar las coordenadas de los n objetos. Estableciendo $S = -\frac{1}{2}D^2$ donde D^2 es una matriz de distancias euclidianas cuadradas.

6.8. Una Derivación Alternativa de A

Dada la matriz D de disimilaridades de orden $(n \times n)$, una configuración espacial para los n objetos puede ser obtenida de la matriz semi-definida positiva A como se definió anteriormente. Los elementos de A fueron definidos como derivaciones de las medias de las filas y las columnas de D y por esta razón las filas y las columnas medias de los elementos de A son cero. Una aproximación alternativa para obtener una configuración espacial es definir los elementos de A usando un punto referencial alternativo. Se menciono anteriormente que la ley del coseno fue usada para relacionar la distancia euclidiana cuadrada a la medida coseno de similaridad. Esta aproximación puede ser también usada para obtener la matriz semidefinida positiva A.

Un punto referencial particular v es seleccionado de un conjunto de n puntos en un espacio de dimensión p. Para cada par de puntos, digamos r y s, las distancias euclidianas cuadradas que relacionan los tres puntos están dadas por d_{rs}^2,d_{rv}^2 y d_{sv}^2 respectivamente. Denotando por θ al ángulo entre los dos vectores formados del punto v a los puntos r y s respectivamente, la ley del coseno da la siguiente relación:

$$d_{rs}^2 = d_{rv}^2 + d_{sv}^2 - 2d_{rv}d_{sv}\cos\theta$$

La Figura 6.3 se ilustra la relación para las dos dimensiones X_1 y X_2. La matriz de orden $(n \times n)$ de elementos a_{rs},$r, s = 1, 2, \ldots, n$, esta dada por

$$a_{rs} = (d_{rv}^2 + d_{sv}^2 - d_{rs}^2) = 2d_{rv}d_{sv}\cos\theta$$

y se puede usar para obtener una matriz semidefinida positiva A. La matriz A puede ser usada para obtener una configuración espacial de los n puntos como ya se menciono.

6.9. El Problema de la Constante Aditiva

Si la matriz de disimilaridad D de orden $(n \times n)$ para n objetos es euclidiana, entonces existe un entero p tal que la configuración dimensional p puede ser determinada para los n objetos. En algunas aplicaciones las disimilaridades son estimadas de tal forma que la matriz D puede ser no euclidiana. Las disimilaridades pueden ser validas como distancias en escala intervalar, pero estas no pueden ser aceptadas como distancias de escala de razón. Las distancias intervalares son corregidas por una constante c, pero el origen es indefinido (ejemplo, temperaturas Celcius y Fahrenhit son escales intervalares pero no son escalas de razón puesto que le punto cero es arbitrario).

Las distancias verdaderas δ_{rs} pueden ser relacionadas a los valores de disimilaridad d_{rs} mediante la ecuación $\delta_{rs} = d_{rs} + c$, $r, s = 1, 2, \ldots, n$. Si el valor de c es suficientemente grande las disimilaridades observadas pueden ser no euclidianas. Para ilustrar esto recordemos que las distancias euclidianas deben satisfacer la desigualdad del triángulo dada por:

$$\delta_{rs} \leq \delta_{ru} + \delta_{su}$$

Mediante la substracción de una constante c de los tres términos en esta ecuación un punto puede ser alcanzado tal que $d_{rs} > d_{ru} + d_{su}$. La matriz de disimilaridad D no es euclidiana, la matriz A derivada de la relación:

$$a_{rs} = -\frac{1}{2}[\delta_{rs}^2 - \delta_{r\bullet}^{-}\delta_{\bullet s}^2 + \delta_{\bullet\bullet}^2]$$

$$r, s = 1, 2, \ldots, n$$

no será mas semidefinida positiva y por esta razón dará al menos un eugenvalor negativo. Si A no esta semidefinida positiva no será posible derivar p dimensiones que reproduzcan D. Es posible aun derivar dimensiones correspondientes a los eigenvalores positivos de A y por esta razón aproximar D. Si los eigenvalores negativos son relativamente pequeños la aproximación basada en los eigenvalores positivos podría ser adecuada.

Una aproximación alternativa es determinar una constante c que puede ser aumentada a todos los elementos fuera de la diagonal de D para asegurar que la matriz sea euclidiana. Si c es suficientemente grande, D será euclidiana, por otro lado lo que se requiere es el valor más pequeño de c que garantice que D es euclidiana. El objetivo de MDS es el de minimizar p y por esta razón c tiene que ser lo más pequeña posible. Algunas rutinas de calculo de MDS determinan una valor aproximado de c aunque es posible la determinación de c de manera precisa.

6.10. Aplicación de la Escala Métrica

Si tenemos el mapa exacto de distancias entre las ciudades principales de Europa podemos usar escalas métricas para producir un mapa que reproduzca las distancias exactas entre las ciudades. Por otro lado no se podrá reproducir un mapa que muestre la localización de las ciudades correctamente con respecto a su localización verdadera en coordenadas N-S y E-O. La localización derivada de las ciudades podrían moverse alrededor de las direcciones N-S y E-O también como el de rotar para obtener una correcta orientación. Si por otro lado podemos colocar dos ciudades en forma correcta, entonces las ciudades restantes serán automáticamente corregidas usando las distancias entre las ciudades. Son únicamente $(n-2)$ dimensiones independiente disponibles de las n ciudades.

Si en lugar de las distancias, tenemos horarios de una aerolínea, en los cuales están los tiempos de vuelo requeridos entre ciudades principales, el mapa de localización podría ser solo aproximado. Los tiempos de vuelos están basados también en factores como son: condiciones de vuelos y escalas. La matriz de tiempos de vuelos seguirá siendo euclidiana. Las coordenadas derivadas en dos dimensiones reflejaran simplemente algunos errores que involucran factores que no son distancias.

En experimentos que tienen que ver con humanos las disimilaridades son frecuentemente basadas en juicios que son objetos de medida de error. Las disimilaridades obtenidas son por lo tanto únicamente aproximaciones y en adición a la naturaleza de las dimensiones subyacentes es de cierta manera vaga. En el caso la solución MDS es usada para derivar una compresión acerca de las relaciones entre los objetos como se percibe en los sujetos del experimento. El análisis no da modelos que puedan hacer predicciones individuales precisas.

Adicionalmente a los problemas asociados con la medida de disimilaridad verdadera. Los procedimientos de MDS también intentan minimizar el número de dimensiones derivadas. Con diversos niveles de aproximación inherente en el procedimiento MDS, la intención de reproducir las disimilaridades precisas no parecen estar justificadas. Escalamiento no métrico busca únicamente el preserva una relación ordinal entre las disimilaridades originales y las distancias derivadas. Si dos objetos A y B se perciben más similares que los objetos C y D, entonces las distancias derivadas podrían también reflejar esta relación relativa. Si las diferencias en percepción similar son insignificantes, los procesos de escalamiento pueden no preservar las diferencias. Unicamente diferencias grandes en percepción de similaridad son preservadas mediante el proceso escalar.

6.11. Interpretación de la Configuración Obtenida

Una manera en que se interpretarán habitualmente los gráficos de MDS es a partir de la posición de los estímulos en las dimensiones: en qué se diferencian los estímulos situados a la derecha de los situados a la izquierda, o los situados arriba de los situados abajo. De modo que se utilizan las dimensiones como continuos que nos permiten resumir la información proporcionada por la representación espacial. Dado que se pueden rotar las soluciones proporcionadas por MDS (excepto las proporcionadas por el modelo INDSCAL), se puede utilizar como referencia cualquier par de ejes ortogonales alternativos a las dimensiones horizontal y vertical, si ello mejora la interpretabilidad de los resultados. La solución proporcionada por SPSS no presenta la orientación más interpretable, porque estamos acostumbrados a manejar un mapa utilizando unas coordenadas arbitrarias Norte-Sur y Este- Oeste.

Además de la rotación, existen otras transformaciones que se pueden realizar en las soluciones proporcionadas por MDS sin alterar para nada sus propiedades. Una de estas transformaciones es la reflexión. Es posible reflejar las coordenadas de los estímulos en cualquiera de las dimensiones, o en varias de ellas simultáneamente, si con ello se facilita la interpretación de los resultados. Para reflejar la configuración de estímulos en una dimensión determinada bastará con cambiar el signo a todas las coordenadas de esa dimensión (convirtiendo los valores negativos en positivos, y viceversa). La segunda transformación posible consiste en reescalar los estímulos, multiplicando los valores originales de las coordenadas por una constante. Esto es posible debido a que las distancias son valores en escala de razón, por lo que una transformación del tipo $x' = bx$ (donde x'son las nuevas coordenadas y x las originales) es perfectamente legítima. Una tercera

transformación posible (excepto en modelos vectoriales) es tratar el origen de coordenadas de su posición actual, sumando un valor constante a los valores de coordenadas originales ($x' = a + x$)

6.12. ALSCAL

ALSCAL es un programa de Escalamiento Multidimensional (MDS). Usa un planteamiento de mínimos cuadrados para escalar. Es capaz de hacer un gran número de análisis, es apropiado para cualquier tipo entre dos o tres tipos de escalas de medida: nominal, ordinal o de razón. ALSCAL esta disponible como un *stand-alone*[2] programm para computadoras con un compilador FORTRAN. También es distribuido como parte del sistema SPSS [16](Forrest).

ALSCAL realiza Escalamiento Métrico y No-Métrico con diferentes opciones. Puede analizar una o más matrices de disimilaridades o similaridades. El análisis representa las filas y columnas de la matriz de datos como puntos en un espacio Euclidiano. Si una fila y columna son similares, entonces los puntos estarán cercano, mientras que si las filas y columnas son disimilares estos puntos estarán alejados.

- ALSCAL realiza análisis métrico y no métrico: la escala multidimensional puede ser métrica o no métrica. El escalamiento métrico asume que las datos dis/similaridades son cuantitativos y que son medidos en un escala de medida intervalar o de razón. El escalamiento no-métrico asume que los datos son cualitativos, que los datos son de escala ordinal.

- ALSCAL analiza una o mas matrices de disimilaridad: ALSCAL puede analizar una o mas matrices de datos de dis/similaridad. La matrices pueden ser rectangulares o cuadradas, simétricas o asimétricas, condicional o incondicional y/o puede tener elementos faltantes. El programa permite el análisis de cualquier número de matrices, cada uno con cualquier número de filas o columnas.

- ALSCAL ejecuta escalamiento o expandido: ALSCAL puede analizar datos cuyas filas o columnas se relacionan al mismo conjunto de objetos o eventos (datos cuadrados los cuales pueden ser simétricos o asimétricos) o las cuales se relacionan a dos diferentes conjuntos de objetos o eventos (datos rectangulares). Cuando los dados son cuadrados el programa realiza un escalamiento multidimensional construyendo un espacio Euclidiano el cual tiene puntos para cada objeto/evento. Cuando los datos son rectangulares el programa realiza un escalamiento expandido, construyendo un espacio euclidiano el cual tiene puntos por cada objeto/evento de las filas y columnas. En ambos casos, la distancia entre los puntos corresponde a las dis/similaridades entre los objetos/eventos.

- ALSCAL realiza modelos de diferencia individuales: ALSCAL puede analizar datos que contiene varias matrices. Con varias matrices ALSCAL puede realizar diferencias replicadas o individuales de escalas multidimensionales o expandidas. Para escalamiento replicado o expandido el análisis construye un espacio Euclidiano justo como si fuese solo una matriz de datos. Para diferencias individuales escalamiento o expandido, los objetos/eventos son representados por puntos en un espacio euclidiano, mientras que las matrices son representadas por vectores o pesos en un espacio de diferencias individuales adicional.

[2] Aplicación independiente

- Disponibilidad: ALSCAL esta disponible como un ejecutable IBM-PC. Requiere un coprocesador matemático. El código de fuente FORTRAN, ejecutable, datos de prueba y los ejemplos son distribuidos en diskettes IBM-PC con una guía de usuario. El código fuente puede ser compilado en otras maquinas. ALSCAL esta disponible en SAS, SPSS y IMSL.

6.13. Procedimiento PROXSCAL

PROXSCAL (PROXimities SCALing[3]) de Commandeur y Heiser, es un programa creado en el departamento de Teoría de Datos de la Universidad de Leiden, el mismo que creó el módulo CATEGORIAS de SPSS. Es parte del paquete de MDS desde la versión 10 de SPSS para Windows. Este programa incorpora gran cantidad de novedades respecto a programas como ALSCAL o KYST. El programa se ha desarrollado a partir del programa SMACOF (Scaling by MAjorizing a COmplicated Function [4]), que utiliza el algoritmo de mayorización para minimización del Stress, además del método de mínimos cuadrados alternantes para la transformación de las puntuaciones.

PROXSCAL puede llevar a cabo gran número de análisis diferentes y utilizar varios modelos de MDS diferentes (entre ellos, el modelo métrico, no-métrico, INSCAL e IDIOSCAL). También permite aplicar restricciones al espacio de estímulos, bien fijando la posición de algunos estímulos en el mismo, bien restringiendo el espacio de tal modo que sea una combinación de variables externas. Las proximidades pueden ser transformadas de varias formas diferentes (monótona, potencial, spline, lineal y proporcional). Por último, y a diferencia de ALSCAL, PROXSCAL no lleva a cabo el análisis sobre los datos transformados (disparidades), sino directamente sobre las proximidades de entrada [37] (Real Deus 2001).

6.14. Comentarios

En general, el proposito de este análisis es detectar dimensiones subyacentes significativas que permita al investigador explicar las distancias entre los objetos estudiados. Con el Escalamiento Multidimensional se puede analizar toda clase de distancia o similaridades. Las similaridades pueden representar el porcentaje de acuerdo entre jueces, las percepciones de la gente, las similaridades entre marcas, etc. El algoritmo MDS cae dentro de una taxonomia, dependiendo del significado de la matriz resultante y puede ser:

- El Escalamiento Multidimensional Clásico también llamado Escalamiento Multidimensional Métrico asume que la matriz resultante es solo una matriz de distancias objeto a objeto. Análogamente al Análisis de Componentes Principales, un problema eigenvector es resulto para encontrar las localizaciones que minimiza distorsiones de la matriz de distancia. Su objetivo es encontrar una distancia euclidiana aproximada a una distancia dada. Puede ser generalizada para manejar problemas de distancias de tres formas (la generalización es conocida como DISTATIS).

- Escalamiento Multidimensional Métrico es un subconjunto del MDS clásico que asume una relación paramétrica conocida entre los elementos de la matriz de disimilaridades objeto a objeto y la distancia euclidiana entre los objetos.

[3]Proximidades escalares.

[4]Escalamiento por medio de mayorizar una función complicada (mayorizar en matemáticas es un orden parcial sobre vectores de números reales.

- Escalamiento Multidimensional Generalizado es un subconjunto de MDS clásico que permite a las distancias objetivo no ser euclidianas.
- Escalamiento Multidimensional No-métrico en contraste MDS, MDS no-métrico ambos encuentran una relación monótona no-paramétrica entre las disimilaridades en la matriz objeto-objeto y las distancias euclidianas entre los objetos y la localización de cada objeto en un espacio de dimensión menor. La relación se encuentra típicamente usando regresión isotónica.

Capítulo 7

Una Breve Descripción de SPSS

La Estadística es una ciencia matemática enfocada a la colección, el análisis, la interpretación y la presentación de los datos. Es aplicable a una gran variedad de disciplinas académicas, desde las ciencias físicas y sociales hasta las humanidades, también es usada en la toma de decisiones informadas en todas las áreas de negocios y en las gubernamentales.

Los métodos estadísticos pueden ser usados para resumir o describir una serie de datos; esto es llamada estadística descriptiva. Además, si existen patrones en los datos, éstos pueden ser modelados de forma tal que se pueden tomar en cuenta tanto el azar como la incertidumbre. Así como para delinear inferencias del proceso o de la población que se esta estudiando; esto es llamado inferencia estadística. Ambas, la estadística descriptiva y la inferencial pueden ser consideradas parte de la estadística aplicada. También existe una disciplina de estadística matemática, la cual se ocupa de la base teórica de la estadística.

La estadística aplicada es el uso de la estadística y la teoría estadística en situaciones reales. Cualquiera que cuente con observaciones empíricas, como un medio para conocer el universo donde están sumergidas, puede aplicar la estadística como herramienta de investigación. La estadística aplicada cuenta con una gran variedad de métodos para realizar análisis estadísticos. Aunque, hay métodos que por su complejidad requiere el uso de algún paquete estadístico, se podría decir que cualquier análisis estadístico, en nuestros días, se hace por medio de un paquete estadístico. Por lo que es necesario aprender el manejo de algún software estadístico.

La intención de este capítulo es describir de forma general el paquete SPSS. Primero se dará una descripción de los paquetes estadísticos existentes y sus interfaces. Para después detallar la constitución el software SPSS, un poco de su historia, sus ventanas principales, sus menús desplegables, su interfaz de usuario y sus principales características. Entre las principales características se presentaran las ventanas con las que cuenta y su manejo, su fichero de datos, sus procedimientos estadísticos y su galería de gráficas. Puesto que la versión que se usara en este trabajo esta en ingles los comandos se nombraran así, pero se dará su descripción en español.

7.1. Paquetes estadísticos

Existe un gran número de paquetes estadísticos que permiten realizar análisis estadísticos de manera sencilla y rápida. Algunos paquetes son fácilmente

operables, porque cuentan con interfaces gráficas de usuario (Graphic User Inferface, GUI). Algunos de los paquetes que cuentan con interfaces son: BioStat, EViews, JMP, R Commander y STATISTICA, por mencionar algunos. Otros requieren un entendimiento mas profundo de programación porque utilizan una línea interfaz o CLI de comandos, algunos de estos son: GAUSS, R, SAS y SPlus. Aunque ambos, la interfaz gráfica o la línea interfaz, permiten obtener resultados de procedimientos estadísticos estándar y de pruebas de significancia estadística, el manejo y el aprendizaje de estos no es igual. El manejo adecuado, así como el aprendizaje del manejo del paquete, dependerá de la profundidad a que se quiera llegar en el análisis estadístico. Esto es, si una persona desea realizar un análisis estadístico profundo, se sumergirá seriamente en el estudio de la estadística y deseara un análisis mas completo, por lo que es recomendable usar paquetes mas complejos como los de línea interfaz o CLI de comandos. Aquella persona que quiere un análisis estadístico y solo cuenta con una preparación elemental de estadística, deseara utilizar herramientas mas amigables como las que cuenta con interfaces gráficas de usuario GUI. Sin embargo no hay que olvidar que ambas interfaces, la CLI y la GUI, ejecutan procedimientos estadísticos estándar y que no difieren en gran manera.

Hay paquetes estadísticos que cuentan con una combinación de la dos. Por un lado la interfaz gráfica hace amigable el uso y comprensión de las técnicas estadísticas, por otro la opción del uso de la línea interfaz o CLI permite al usuario generar programas que le brinden un análisis especifico y personalizado. Entre estos paquetes está el SPSS (Statistical Package for the Social Science)1. El SPSS, al igual que los paquetes antes mencionados permite ejecutar procedimientos estadísticos de manera amigable como son: estadísticas descriptivas, estadísticas divariadas, predicción numérica, así como predicción de grupos indefinidos.

7.2. SPSS

El Paquete Estadístico para las Ciencias Sociales SPSS -Statistical Package for the Social Sciences- es un paquete de programas que sirve para manipulación, análisis y presentación de datos; este paquete es ampliamente usado en las ciencias sociales y en las del comportamiento. El programa de computadora SPSS fue puesto a disposición de público en el año 1968. Es usado en investigación de mercados, estudiosos de la salud, por compañías encuestadoras, por dependencias gubernamentales y de educación, entre otros. Además de los análisis estadísticos otros componentes del software básico son: manejo de datos (selección de casos, respaldo de archivos, creación de datos derivados) y documentación de datos (un diccionario metadata, que son datos que describen otros datos, es almacenado con datos).

Existen diversas presentaciones de SPSS. El programa central es llamado SPSS básico, existen un número de módulos complementarios que extienden el rango de entrada de datos, estadísticas o capacidades de reportar. Existen versiones de SPSS para Windows (98, 2000, ME, NT, XP), para plataformas mayores UNIX (Solaris, Linux, AIX) y para Macintosh. En el presente trabajo se describirá el más popular, SPSS para Windows, aunque la mayoría de las características son compartidas por otras versiones. Los análisis realizados en este trabajo esta hecho en SPSS 12.0 para Windows 2003. Aunque ésta versión de SPSS no es la última, los procedimientos realizados para los análisis descritos seguirán siendo apropiados, para versiones posteriores.

El presente trabajo usara los módulos SPSS Básico, Modelos Avanzados y Modelos de Regresión. Otros módulos adicionales son: SPSS Tablas, SPSS

Categorías, SPSS Tendencias, SPSS Análisis de Valores Faltantes, Validación de datos, Conjoint, Muestras Complejas entre otros.

1. SPSS Básico: provee métodos para la descripción de datos, inferencias simples para datos continuos y categóricos, así como regresión lineal simple. También provee técnicas para análisis de datos mulltivariados, específicamente para Análisis de Factores, Análisis de Conglomerados y Análisis Discriminante.

2. El modulo de Modelos Avanzados: incluye métodos para ajustar modelos lineales generales y modelos lineales mixtos, así como estimar datos de supervivencia.

3. El modulo de Modelos de Regresión: incluyen métodos que ajustan modelos de regresión no lineales, como lo es el Análisis de Regresión logística.

SPSS para Windows ofrece una hoja de cálculo electrónica para ingresar y visualizar archivos de datos –el editor de datos (Data Editor). Los resultados de los procesos estadísticos son mostrados en una ventana aparte, la ventana de resultados (Output Viewer). Estas ventanas toman la forma de tablas y/o gráficos que pueden ser manipulados interactivamente y pueden ser copiados directamente en otras aplicaciones. La interfase gráfica de usuario (GUI) hace a SPSS de fácil manejo, mediante procedimientos que pueden ser seleccionados de diversos menús disponibles.

7.3. Interfaz de SPSS

El paquete estadístico SPSS es un grupo de programas de computadora que están especializados en análisis estadísticos. Permite obtener resultados de procesos estandarizados estadísticos y pruebas de significancia estadística, que no requieren programación numérica de alto nivel. Su interfaz gráfica de usuario es un artefacto tecnológico de un sistema interactivo que posibilita, a través del uso y la representación del lenguaje visual, una interacción amigable con el sistema informativo, en el contexto de la interacción persona-ordenador.

La interfaz gráfica de usuario (Graphic User Iterface, GUI), es un tipo de interfaz de usuario que utiliza un conjunto de imágenes y objetos gráficos (íconos, gráficas, tipografía) para representar la información y acciones posibles en la interfaz (figura 7.1). La interfaz de SPSS realiza acciones mediante la manipulación directa, que facilita la interacción del usuario con la computadora. Esta interfaz surgió como una evolución de la línea de comandos de los primeros sistemas operativos y es pieza fundamental en el entorno gráfico.

Muchos de los componentes de SPSS son accesibles vía menús desplegables o pueden ser programados con una especificación 4GL "comando de lenguaje de sintaxis", que es un lenguaje de programación diseñado con un propósito específico en mente, en este caso un lenguaje estadístico. La programación del lenguaje de sintaxis tiene el beneficio de reproducir y manipular datos complejos y analizarlos. La interfase del menú desplegable también genera sintaxis, aunque por default, es invisible para el usuario. Los programas pueden correrse interactivamente o ignorarse. Además un lenguaje macro puede ser usado para escribir subrutinas de lenguaje de comandos y scripts SAX Basic pueden accesar la información del diccionario de datos y manipular la ventana de resultados que se genera por default.

SPSS esta restringido a una estructura de archivos internos, a un tipo de datos, a un procesamiento de datos y una correspondencia de archivos, los cuales simplifican considerablemente la programación. El conjunto de datos de SPSS

Figura 7.1: Interfaz Grafica de Usuario GUI

siempre tienen la estructura de una tabla bidimensional, donde las filas típicamente representan los casos y las columnas representan variables. Únicamente dos tipos de datos están definidos: numérico y texto (ó string). Todos los procesamientos de datos ocurren secuencialmente caso por caso a través del archivo. Los archivos pueden ser correspondidos uno a uno y uno a muchos, pero no muchos a muchos.

7.4. Ventanas de SPSS

Son tres ventanas principales la que componen la interfaz gráfica de usuario de SPSS, la ventana de editor de datos, la ventana de resultados y la ventana de sintaxis. La primera ventana es el Data Editor -Editor de Datos- (figura 7.2). En esta ventana es donde se pueden ingresar datos, hacer manejo de los datos, transformar los datos, generar estadísticas y gráficos, así como tener una visualización de las variables, entre otras opciones. Está compuesta por dos ventanas: Data View y Variable View.

La segunda ventana es la de Output –Resultados- (figura 7.3). Esta es la ventana donde se muestran los resultados de los análisis estadísticos ejecutados. Los resultados pueden ser editados, cambio de fuente, tamaño, presentación y color de los gráficos, o bien pueden ser copiados en otros documentos para su edición y manejo.

La tercera y última ventana, de las ventanas más importantes de SPSS, es la ventana de Syntax –Sintaxis- (figura 7.4). Esta ventana guarda los comandos empleados en el proceso del análisis, se pueden editar comandos que no existen en las ventanas de dialogo de SPSS.

7.4.1. Ventana Vista de Datos

Como se había mencionado antes el Data Editor (Editor de Datos) cuenta con dos ventanas. Por default Data View (Vista de Datos), la cual permite ingresar los datos y visualizarlos. La otra ventana es Variable View (Vista de Variables), la cual permite especificar los tipos de variables así como visualizarlos.

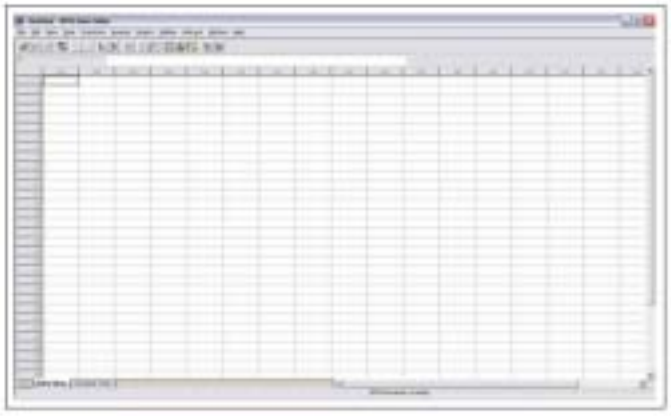


Figura 7.2: Ventana Editor de Datos (Data Editor)

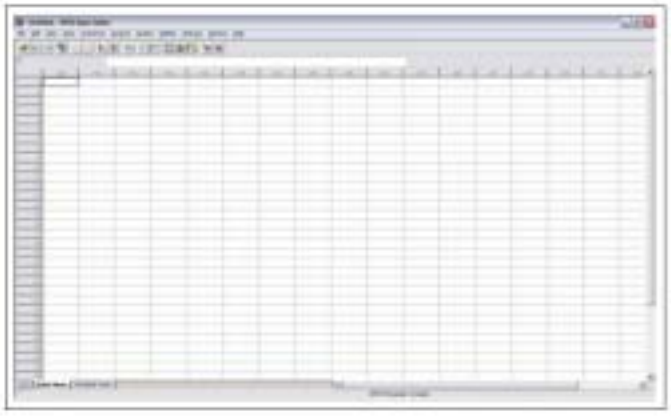


Figura 7.3: Ventana de Resultados (Output)

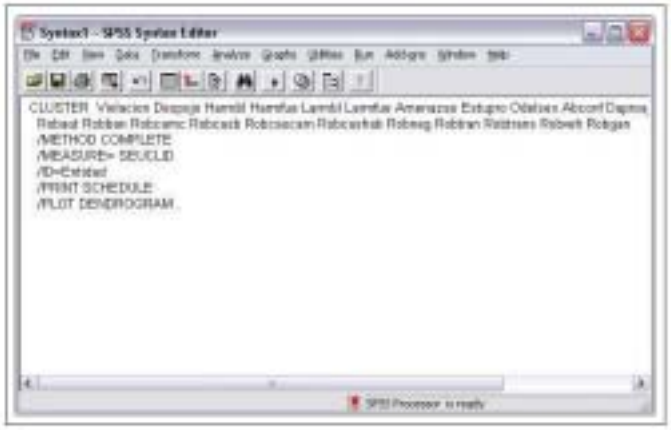


Figura 7.4: Ventana de Sintaxis (Syntax)

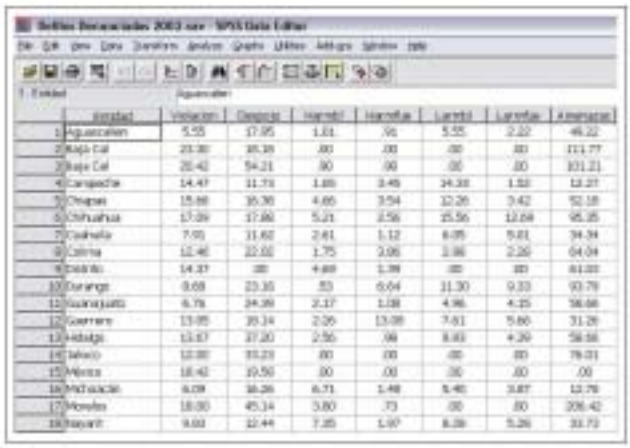


Figura 7.5: Ventana Vista de Datos

Figura 7.6: Barra de Herramientas Pricipal de SPSS

Estas dos ventanas pueden accesarse mediante la selección de las pestañas que se encuentran del lado izquierdo de la pantalla. Desde ésta ventana los valores de los datos pueden ser ingresados. Para la mayoría de los análisis, SPSS asume que las filas representan casos y que las columnas representan variables. Por default SPSS alinea los datos numéricos a mano derecha de la celda y el texto (string) al lado izquierdo de la celda. Además por default SPSS usa un punto para indicar un valor numérico faltante y las celdas de las variables string son dejadas simplemente vacías. La apariencia de Data View (figura 7.5) puede ser manipulada desde el menú desplegable View. Éste puede ser usado para cambiar la fuente de las celdas, quitar líneas, y para hacer visible las etiquetas de valor.

7.4.2. Ventana Variable

La ventana Variable View sirve para definir las variables. Cada definición de variable ocupa una fila de esta ventana. Tan pronto como se han ingresado datos en la ventana de Data View, el nombre asignado de las columnas ocupa una fila en la ventana Variable View. Hay 10 características a especificar en las columnas de la ventana Variable View, que se presentan a continuación.

7.5. Barra de Herramientas de SPSS

La barra de herramienta de SPSS (figura 7.6) es una componente de la interfase gráfica del programa que ofrece las siguientes opciones:

7.5.1. Guardar y Recuperar Archivos de Datos

Para guardar y recuperar archivos de datos se tiene que accesar al menú desplegable después seleccionar File (Archivo) en la barra de menú, al igual que se hace en windows. Los datos mostrados en el editor de datos pueden ser salvados usando los comandos Save (Salvar) y Save As... (Salvar Como). En la forma

Característica	Nombre	Descripción
Name	Nombre seleccionado de la variable	Este puede ser de ocho caracteres alfanuméricos pero tiene que empezar con una letra. El guío bajo se puede usar, pero signos como () y (&) no pueden ser usados.
Type	El tipo de variable	SPSS provee un tipo de variable por default una vez que los valores han sido ingresados en una columna de la ventana Data View. Ésta opción ofrece un número de tipos de datos incluyendo varios formatos de datos numéricos, fechas, o tipos de cambio. El tipo de variable puede ser cambiada seleccionando la respectiva entrada y dando un clic en el símbolo de los tres puntos, que aparece del lado derecho de la celda.
Width	El ancho actual de las entradas de los datos	El ancho de las variables numéricas por default es ocho. El ancho puede ser aumentado o reducido, dando un clic en el símbolo de los tres puntos, que aparece del lado derecho de la celda.
Decimals	El número de decimales	El número de dígitos a la derecha después del punto decimal, que serán desplegados en las entradas de los datos. Dando un clic en el símbolo de los tres puntos, que aparece del lado derecho de la celda.
Label	El etiqueta asociada a los nombres de la variables	En contraste con la variable nombre, ésta no esta confinada a ocho caracteres y espacios que pueden usarse. Es buena idea asignar etiquetas a las variables porque ayudan a recordar el significado de las variables.
Values	Los códigos de la variables asociadas a las etiquetas	Para variables categóricas un código entero debe ser asignado a cada categoría y la variable a definir tiene que ser numérica.
Missing	Los códigos de los valores faltantes	SPSS reconoce un punto como indicador de un valor faltante. Si se utilizan otros códigos, estos tienen que ser declarados (dando un clic en el símbolo de los tres puntos, que aparece del lado derecho de la celda).
Columns	El ancho de la columna de la variable	El ancho de la celda por default es ocho, de la ventana Data View. La celda puede ser cambiada, dando un clic en el símbolo de los tres puntos, que aparece del lado derecho de la celda.
Align	La alineación de las entradas de las variables	SPSS por default alinea las variables numéricas del lado derecho de la celda y las variables string a la izquierda. De igual manera puede ser cambiada dando un clic en el símbolo de los tres puntos, que aparece del lado derecho de la celda.
Measure	Escala de medida de la variable	Por default la selección de SPSS depende de los tipos de datos. Esto es para variables tipos numéricas, la escala de medida seleccionada es la de intervalo. Para variables tipo string la escala de medida seleccionada es la nominal. La tercer opción, y última, es la ordinal que es para variables categóricas, ésta no es asignada por default por lo que es necesario declararla. Es muy importante una buena asignación de las variables, por que ésta tiene implicaciones en los métodos estadísticos que serán aplicables. El orden jerárquico, que maneja SPSS, de esta clasificación es: nominal ¡ordinal¡escalar

Cuadro 7.1: Características de la Ventana Variable de SPSS

Característica	Nombre	Descripción
File	Archivo	Se usa para abrir, crear, guardar archivos así como para convertir y exportar archivos de otras bases de datos.
Edit	Data	Se usa para deshacer y rehacer procedimientos, cortar, copiar, pegar, borrar y buscar variables.
View	Vista	Se usa para generar cambios en la imagen de un archivo de datos.
Data	Datos	Se usa para definir variables, copiar propiedades de los datos, definir fechas, insertar variables, insertar casos, para localizar un caso, transponer, reestructurar, etc.
Transform	Transformar	Se usa para manipular variables o archivos, recodificar variables, reemplazar valores faltantes, entre otras opciones.
Analyze	Analizar	Se usa para seleccionar y generar procedimientos estadísticos.
Graphs	Gráficas	Se usa para generar gráficas.
Utilities	Utilidades	Se utiliza para ver las características de las variables, comentar archivos de datos, definir conjuntos de variables, se puede ver como está el editor de menú.
Add-on	Incorporar	Se usa para adicionar módulos de SPSS y para acceder a los servicios que presta SPSS.
Windows	Ventanas	Se usa para arreglar, seleccionar y controlar los atributos de las diferentes ventanas de SPSS.
Help	Ayuda	Se usa para accesar a la ayuda, a los tutoriales, a la sintaxis de referencia, entre otros.

Cuadro 7.2: Opciones de la Barra de Herramientas de SPSS

Figura 7.7: Menú desplegable de Archivo

común de Windows por medio del control de salvar, de la barra de herramientas, se salvaran los datos bajo el mismo nombre. Si se desea guardar el archivo con un diferente nombre, se usa Save As... SPSS reconoce una gran variedad de formatos, por ejemplo: a los archivos de SPSS se les da la extensión *.sav, ASCII (*.dat), Exlcel (*.xls), o dBASE (*.dvf) están disponibles. Los archivos con diferente terminación, a la de SPSS, los convierte al formato *.sav, previas indicaciones.

7.5.2. Los Menús Estadísticos

Los menús desplegables disponibles después de seleccionar Data, Transform, Analyze y Graphs de la barra de menús proveen procedimientos concernientes con diferentes aspectos de un análisis estadístico. Estos permiten la manipulación del formato de los datos que serán usados para el análisis (Data), generación de nuevas variables (Transform), realizar procedimientos estadísticos (Analyze) y generar gráficas (Graphs). La mayoría de las selecciones del menú estadístico abre cajas de dialogo (figura 7.8). Estas ventanas pueden ser usadas para selec-

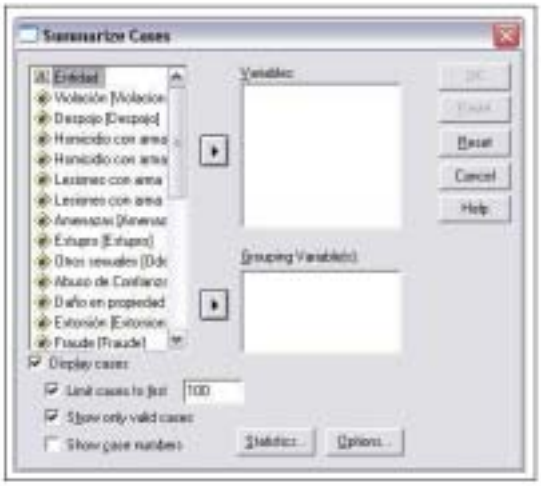


Figura 7.8: Ventanas de Diálogo

cionar variables y opciones de análisis. Un dialogo principal para procedimientos estadísticos tiene diferentes componentes:

Lista de variables de origen es una lista de variables, de la ventana Data View, que pueden ser usadas en el análisis requerido. Únicamente los tipos de variables que serán permitidos para el procedimiento son mostrados en la lista de origen. Las variables de tipo "string" no son permitidas. Un icono-signo cerca del nombre de las variables indica el tipo de variable. Un signo # es usado para indicar variables numéricas y *A* indica que la variables es de tipo string.

Lista de variables objetivo es la lista de variables que serán incluidas en el análisis. Botones de comandos son botones que pueden ser presionados para instruir al programa a realizar una acción.

7.6. Procedimientos Estadísticos

El efectuar una variedad de análisis estadísticos con SPSS es el punto focal de este paquete, bajo el menú desplegable de Analyze (Analizar) (figura 7.9) existen una gran variedad de métodos estadísticos. Estos métodos estadísticos que pueden ser usados para resumir o describir una colección de datos (Estadística Descriptiva) o bien para modelar patrones en los datos así como para formular hipótesis acerca de la población que se estudia (Estadística Inferencial). Partiendo desde la premisa de que los procedimientos estadísticos pueden ser usados para resumir o describir una colección de datos, algunos procedimientos que forman parte de SPSS se describirán brevemente a continuación.

El comando Reports ofrece cuatro procedimientos que permiten dar un resumen-informe de las variables.

- OLAP Cubes (Cubos OLAP) calcula totales, medias y otras estadísticas univariadas para datos continuos. Un nivel separado en una tabla es creado para cada categoría de cada variable.
- Case Sumaries (Resúmenes de casos) genera resúmenes estadísticos por variables. Todos las variables son tabuladas, se puede escoger el orden en que se desplegaran los procedimientos estadísticos seleccionados y se generara una síntesis resumen para cada variable cruzada de todas las categorías.

Figura 7.9: Menú desplegable Analizar

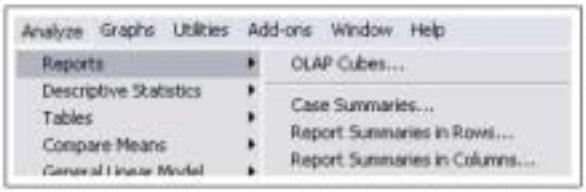


Figura 7.10: Opciones del comando Reports (Reportes)

- Los otros dos procedimientos son: Report Summaries in Rows y Report Summaries in Columns, el primero genera resúmenes-reportes por filas y el último genera resúmenes-reportes por columnas (figura 7.10).

7.7. Clasificación

La clasificación estadística (Classify)es un procedimiento estadístico en el cual los individuos son colocados en grupos basados en información cuantitativa sobre una o más características inherentes en los objetos y basados en un conjunto de vectores respuestas de objetos nombrados previamente. Existen muchos métodos para clasificar, pero estos resuelven uno de tres problemas matemáticos relacionados.

Lo primero es encontrar un mapa espacial de características (el cual es típicamente un vector multidimensional) a un conjunto de valores. Esto es equivalente a particionar las características en regiones, después asignar una etiqueta a cada región. Tales algoritmos típicamente no son confiables a menos que un post-proceso sea aplicado. A otro conjunto de algoritmos primero para resolver este problema primero se le aplica un conglomerado no supervisado a las características, después se intenta etiquetar cada uno de los conglomerados o regiones.

Bajo este comando se puede accesar a cuatro opciones de clasificación: Conglomerados en dos fases, Conglomerado k-medias, Conglomerado Jerárquico y Discriminante (figura 7.11).

- Análisis de Conglomerados de dos fases (TwoStep Cluster Analysis) es-

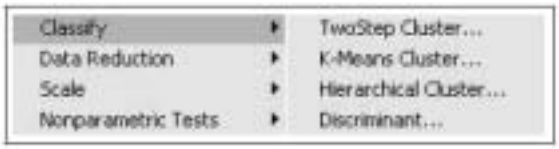


Figura 7.11: Opcines del Comando Clasificación

te procedimiento es una herramienta exploratoria diseñada para revelar agrupamientos naturales (conglomerados) dentro de un conjunto de datos que de otra manera no son aparentes. El algoritmo empleado en este procedimiento tiene varias componentes deseables que lo diferencian de las técnicas de conglomerados tradicionales:

- Manejar variables categóricas y continuas, mediante la suposición de que las variables son independientes, una distribución multinomial-normal conjunta puede ser estimada sobre variables categóricas y continuas.
- Selecciona automáticamente el número de conglomerados por medio de comparar los valores de un modelo-seleccionado de diferentes soluciones de conglomerados, el procedimiento puede determinar automáticamente de forma óptima el número de conglomerados.
- Escalabilidad por medio de la construcción de un árbol de conglomerados-características que resumen los puntajes, el algoritmo dos fases permite analizar un conjunto grande de datos.

- Análisis de conglomerados de k medias (K-Means Cluster Analysis) este procedimiento pretende identificar grupos relativamente homogéneos de casos basados en la característica seleccionada, usando un algoritmo que puede manejar un numero grande de casos. Aunque, este algoritmo requiere que se especifique el numero de conglomerados. Se puede especificar inicialmente conglomerados centrales si se tiene información previa de los datos. Se puede seleccionar uno de dos métodos para clasificar casos, ya sea actualizando los conglomerados centrales iterativamente o únicamente clasificándolos. Se puede salvar los miembros de los conglomerados, información de las distancias, y los conglomerados centrales. Opcionalmente se puede especificar una variable cuyos valores son usados para nombrar los resultados. También se puede pedir un análisis de la varianza del estadístico F. mientras que otras estadísticas son oportunistas (los procedimientos intentan formar grupos que difieren), los tamaños relativos de los estadísticos proveen información de cada una de las contribuciones de las variables para la separación de los grupos.

- Análisis de conglomerados jerárquicos (Hierarchical Cluster Analysis) este procedimiento intenta identificar grupos homogéneamente relativos de casos (o variables) basados en las características seleccionados, usando un algoritmo que empieza con cada caso (o variable) en un conglomerado separado y combina conglomerados hasta que únicamente queda uno. Se puede analizar variables brutas o se puede escoger de una gran variedad de transformaciones estandarizadas. Distancias o medidas de similaridades son generadas por procedimientos de proximidades. Estadísticas son mostradas en cada fase para ayudar a seleccionar la mejor solución.

- Análisis de Discriminante (Discriminant Analysis) este análisis es útil para situaciones en donde se quiere construir un modelo predictivo de un

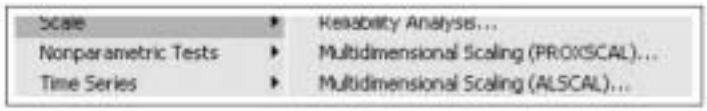


Figura 7.12: Opciones del Comando desplegable de Escalas

grupo basado sobre características observadas de cada caso. Los procedimientos generados una función discriminante (o para mas de dos grupos, un conjunto de funciones discriminantes) basados en combinaciones lineales de las variables predoctoras que proveen el mejor discriminante entre dos grupos. Las funciones son generadas de una muestra de casos para cuyos miembros del grupo es conocido; las funciones pueden entonces ser aplicadas a nuevos casos con medidas para las variables predictoras para un grupo nuevo de variables.

7.8. Escalas

Este comando (Scale) contempla tres opciones: análisis de fiabilidad, escalamiento multidimensional (PROXSCAL) y escalamiento multidimensional (figura 7.127.26).

- Análisis de Fiabilidad (Reliability Analysis) permite estudiar las propiedades de escalas de medida y los objetos que los componen. El procedimiento calcula un numero de medidas usadas comúnmente de fiabilidad escalar también provee información de las relaciones entre individuos en la escala. Coeficientes de correlación interclase pueden ser usados para calcular estimadores fiables ínter índices.

- Escalamiento Multidimensional PROXSCAL (Multidimensional Scaling - PROXSCAL-) este método multivariado intenta encontrar la estructura en un conjunto de medidas de proximidad entre objetos. Esto se logra, asignando observaciones a localidades específicas en un espacio de menor dimensión conceptual tal que las distancias entre los puntos del espacio coincidan con las disimilaridades o similaridades lo más cerca posible. El resultado es una representación de los objetos en un espacio de dimensión más pequeño, el cual en muchos casos ayuda a entender los datos.

- Escalamiento Multidimensional (Multidimensional Scaling) este método pretende encontrar una estructura en un conjunto de medidas de distancias entre objetos o casos. Esto se logra mediante la asignación de observaciones a una localidad en un espacio conceptual (usualmente dos o tres dimensiones) tal que la distancia entre puntos en el espacio coincidan con las disimilaridades lo mas cercano posible. En muchos casos, las dimensiones de este espacio conceptual pueden ser interpretadas y usadas para una comprensión posterior de los datos. Si se las variables medidas se tomaron objetivamente, escalas multidimensionales pueden ser usadas como un técnica de reducción de datos (el procedimiento de escala multidimensionales calculara las distancias de datos multivariados, si es necesario). Escalas multidimensionales puede ser aplicada para posicionar disimilaridades subjetivamente entre objetos o conceptos. Además, el procedimiento de escalas multidimensionales puede manejar disimilaridades de múltiples fuentes, como son índices múltiples o múltiples respuestas de cuestionarios.

Capítulo 8

La Seguridad Pública en México: Un Análisis Conglomerado y de Escalas Multidimensionales en SPSS

El tema de seguridad pública en el país esta presente en la agenda gubernamental hoy en día, junto con los temas de pobreza y la escasez de agua. En el ámbito social la seguridad y la justicia han pasado a ser objeto de análisis y crítica constante. La seguridad pública es una de las exigencia más sentidas de la ciudadanía. La dimensión y complejidad de la seguridad pública, aunada a las tendencias de incremento en las cifras delictivas, hacen necesaria una reconceptulización de los mecanismos de estudio desde un enfoque multidisciplinario y multifactorial. El enfoque multidisciplinario del conocimiento científico brindará en mayor grado la objetividad de los resultados y al sustento de teorías. Uno de los factores más complejos de este fenómeno es la percepción de la población sobre seguridad pública

La percepción ciudadana sobre hechos relacionados con la violencia está influenciada por los medios de comunicación. La cobertura periodística sobre la delincuencia en México, hacia finales de las década de 1980, comenzó a recolocarse como el Tema de Temas. Los medios de comunicación empezaron a cubrir notas rojas, antes ausentes, en los principales espacios noticiosos y de opinión de la radio y televisión, así como en los encabezados de las primeras planas de prensa, al convertirse la seguridad pública en el asunto de mayor preocupación para millones de personas. Esto por supuesto no fue un hecho fortuito, fue el producto de las notables transformaciones registradas en los últimos tiempos en todos los órdenes de la vida nacional. La consolidación de la corrupción, la impunidad y la aparición de poderosos grupos de interés dentro de las estructuras de la seguridad pública en México fue parte de un fenómeno identificado por analistas como un virtual proceso de gangsterización de alto impacto para el Estado Mexicano. Después del 2 de julio del 2000 se abrieron cauces para la transición y el desmantelamiento de aparatos de poder. A la caída del viejo régimen emergieron actores y factores antisistemicos que, como las organizaciones de crimen organizado, crecieron en los últimos años amparados de la capacidad corruptiva.

La percepción del público sobre seguridad pública es claramente negativa, como es posible observar en diferentes encuestas realizadas a lo largo del país. Estas percepciones sin embargo muchas veces son el resultado directo del bombardeo informativo de notas rojas y de la carencia de información confiable que deriva en opiniones negativas y en muchas ocasiones erróneas. La percepción a nivel nacional de la seguridad pública se ha visto ensombrecida por la falta de información confiable de nueva creación que emplee instrumentos científicos para su análisis. Por lo que la motivación del presente trabajo es aportar una metodología para el análisis de la información, comparando los índices de delincuencia entre entidades federativas, conjuntamente con la generación de una nueva perspectiva para una futura realización de instrumentos similares para crear una conciencia real del problema mediante un análisis multivariado apropiado de estos factores.

8.1. Seguridad Pública en México

La violencia que está padeciendo la población Mexicana se traduce en un problema de seguridad pública, tanto por la dimensión que ha adquirido la muerte por dichas causas como por los efectos materiales y emocionales que ocasiona. Su origen se encuentra en factores históricos, demográficos, psicológicos, económicos, biológicos y sociales entre otros. Estos factores determinantes de la creciente inseguridad abarcan el campo de acción de diferentes disciplinas. A ello se debe que su conocimiento sistemático se convierta en una necesidad prioritaria para los mexicanos.

La inseguridad en México se ha venido deteriorando desde hace 50 años, en un proceso constante y acumulativo, pero no es un incidente instantáneo y único. Las causas generatrices de ésta, han señalado los especialistas, son más de medio centenar. Estas causas están divididas en cinco grandes grupos: las históricas, de naturaleza económica, índole social, de orden político-administrativas y de naturaleza cultural. El primer grupo está ligado a frustraciones ancestrales, composición étnica de la población, predisposición congénita, condiciones geográficas y alteraciones climatológicas. El segundo está conformado al desarrollo económico desigual, desempleo o subempleo, falta de expectativas profesionales, insuficiencia retributiva del salario y los nuevos patrones de consumo. El tercero esta formado por factores tales como la deficiente planeación urbana, sobrepoblación, mala canalización del ocio y la disgregación familiar. Bajo el grupo político-administrativos se tienen contemplados los factores de corrupción, incompetencia política, abandono presupuestal, abandono administrativo, falta de voluntad política para combatir la delincuencia, la insuficiencia de centros de readaptación social, benevolencia de las penas, deficiente legislación y los procedimientos en materia penal, tortuosos e incomprensibles. El último y no menos importante son los factores de naturaleza cultural: bajo nivel educativo, cultura de impunidad, crisis de valores, promoción de la violencia a través de los medios de comunicación [39].

La seguridad pública[1] forma parte esencial del bienestar de una sociedad. Un estado de derecho genera las condiciones que permiten a un individuo realizar

[1]En la Constitución Política se reconocen los derechos fundamentales de todo individuo. Estos valores son tutelados en el ámbito de todas la actividades estatales incluídas en la función de seguridad pública. Para entender de manera precisa lo que el concepto de seguridad pública implica para el Estado Mexicano es necesario referirse a los reordenamientos que han tenido por objeto estructurar los sistemas de seguridad pública dentro de nuestra sociedad. Este reordenamiento, que se publicó en el Diario Oficial de la Federación el 11 de noviembre de 1995, conceptualiza a la seguridad pública como "la función a cargo del Estado que tiene como fines salvaguardar la integridad y derechos de las personas, así como preservar las libertades, el orden y la paz públicos".

sus actividades cotidianas con la confianza de que su vida, su patrimonio y otros bienes jurídicos tutelados están exentos de todo peligro daño o riesgo. Ante la realidad de un estado que no cumple con una de sus principales funciones, la de suministrar seguridad, los ciudadanos centran una gran parte de su esfuerzo en la defensa de sus bienes y derechos.

Hasta este punto es evidente que la violencia social requiere de diversas investigaciones acerca de su origen y causalidad y no es terreno exclusivo de una sola disciplina y no debe circunscribirse a posiciones teóricas únicas y definidas, como teorías sociales o de área médica, la epidemiología o la aplicación de métodos con alto riesgo estadístico teórico. Es decir, no se debe adecuar la realidad a las necesidades metodológicas de la teorías, si no más bien analizar los factores en forma multidimensional que exija un trabajo conjunto de especialistas de diferentes áreas del conocimiento, sin olvidar la incorporación de nuevas metodologías e instrumentos de análisis cualitativo y cuantitativo que ofrecen la aplicación de modelos matemáticos.

Además, y a diferencia de otras épocas, los medios de comunicación tienen un papel muy importante en la difusión de hechos vinculados con la violencia que en la mayoría de los casos influyen en la percepción el público. Por esto, es necesario realizar investigaciones científicas que garanticen la objetividad de los resultados. Puesto que la información generada será fundamental para describir el comportamiento delictivo en cada una de las zonas analizadas es necesario el uso de un índice adecuado que sea comparable entre las diversas poblaciones, para después proponer explicaciones de incidencia delictiva de acuerdo a las características regionales que presenta el fenómeno violento es las áreas estudiadas.

Una de las formas primarias con que se mide el fenómeno delictivo en nuestro país es contabilizando el número de denuncias presentadas ante el Ministerio Público. Otra manera es por medio del número de consignaciones judiciales y sentencias ejecutadas. Una más es la de dividir el número de delitos conocidos entre un determinado número de habitantes. A este resultado se le conoce como índice de criminalidad de una ciudad o país. Estas formas de medir la criminalidad son limitadas y proporcionan una interpretación errónea del problema delictivo al usar cifras parciales, lo que imposibilita un diseño adecuado de políticas de atención en la materia. Una consecuencia negativa adicional es la erosión de la confianza entre autoridades y ciudadanos.

La unidad de Análisis sobre Violencia Social del Instituto de Investigaciones Sociales de la UNAM realiza actividades encaminadas al fortalecimiento de investigaciones y publicaciones que aporten datos confiables y proporcionen análisis novedosos. Para ello aprovechan las encuestas de Seguridad Pública mediante técnicas multivariantes. Todo esto, como parte del desarrollo de investigaciones que cuenten con elementos de análisis estadístico que enriquezcan el conocimiento que se tiene sobre violencia social.[21].

La estadística multivariada describe una colección de procedimientos que involucran observaciones y análisis de más de una variable estadística al mismo tiempo. Una distinción entre univariada y multivariada es que la estadística univariada únicamente tiene una variable dependiente mientras que la estadística multivariada tiene dos o más variables dependientes. Algunas técnicas multivariadas permiten crear grupos de objetos o variables similares entre sí, a partir de sus características medidas; mientras que otras técnicas que están dirigidas a explorar las dimensiones subyacentes, formando percepciones acerca de las disimilaridades o similarilaridades entre objetos, por mencionar algunas. Aplicaciones de estas técnicas son las investigaciones que van dirigidas a obtención de clasificaciones. El Análisis de Conglomerados tiene como objetivo clasificar el conjunto de objetos en un número reducido de grupos basados en las semejanzas

entre ellos. Proporcionando un mapa de los objetos en un espacio reducido en el que la posición del objeto refleja su grado de disimilaridad o similaridad percibida con otros objetos, esta técnica es conocida como Escalas Multidimensionales.

8.2. Descripción de los datos

Los datos que serán usados en este análisis surgen del Sistema de Información Delictiva: La Estadística de Seguridad Pública en México del Instituto Nacional de Ciencias Penales del 2004. Estos datos se pueden ver en las tablas 8.1 y 8.2. El Sistema proporciona información estadística delincuencial y de seguridad pública en México, para cada una de las entidades federativas y el Distrito Federal. La información que se presenta es solamente la que es denunciada y no toma en cuenta las cifras negras de delito o no denunciadas. Además se incluyen cifras sobre población estimada por entidad federativa de la CONAPO para el año 2002.

Entidad Federativa			Homicidio		Lesiones		Otros delitos			Patrimonial			
	Violación	Despojo	Homicidio con arma blanca	Homicidio con arma de fuego	Lesiones con arma blanca	Lesiones con arma de fuego	Amenazas	Estupro	Otros delitos sexuales	Abuso de Confianza	Daño en propiedad ajena	Extorsión	Fraude
Aguascalientes	55	178	10	9	55	22	488	9	162	470	1757	13	926
Baja California	609	417	0	0	0	0	2921	140	1312	934	13666	90	1793
Baja California Sur	91	217	4	4	0	0	451	29	141	291	1634	22	648
Campeche	105	77	12	25	104	11	89	23	74	33	270	0	79
Chiapas	645	617	192	146	505	141	2149	217	410	873	4187	110	2352
Chihuahua	548	522	167	82	499	407	3058	51	532	998	4859	49	1571
Coahuila	191	259	63	27	145	121	829	24	299	352	4758	0	954
Colima	71	114	10	22	17	13	365	17	50	93	410	0	344
Distrito Federal	1298	0	424	126	0	0	5514	25	2073	2613	20859	289	5831
Durango	132	331	8	101	172	142	1427	39	136	739	1809	17	1310
Guanajuato	381	1102	106	53	248	209	2872	63	364	1078	10291	52	2168
Guerrero	422	541	73	423	246	188	1011	84	190	497	2800	75	1056
Hidalgo	321	804	60	23	233	103	1377	73	256	407	2257	21	839
Jalisco	797	2028	0	0	0	0	5046	190	578	1599	8189	253	6676
México	2533	2421	0	0	0	0	0	209	0	2373	14448	0	3515
Michoacán	255	683	281	62	226	162	595	135	269	598	2643	162	2085
Morelos	294	666	62	12	0	0	3371	63	221	521	2759	108	1313
Nayarit	95	112	71	19	80	51	326	49	91	102	382	0	297
Nuevo León	362	1171	58	32	0	0	2257	104	1160	937	6231	32	4653
Oaxaca	317	893	0	0	0	0	2993	128	613	419	3587	101	1716
Puebla	696	1491	230	87	829	227	4766	151	466	1069	5865	0	4012
Querétaro	210	316	7	9	47	21	587	17	136	409	2548	8	1138
Quintana Roo	419	544	19	22	371	43	415	40	349	593	2856	6	1468
San Luis Potosí	404	0	73	28	0	0	3415	91	954	1107	6196	13	2718
Sinaloa	204	465	391	36	223	234	1164	85	214	460	907	92	814
Sonora	223	350	73	44	0	0	466	87	264	235	2793	17	604
Tabasco	277	442	0	0	0	0	2429	152	376	511	2714	25	1519
Tamaulipas	561	159	98	33	190	54	2262	94	412	574	2655	9	1819
Tlaxcala	225	98	9	2	4	3	109	0	140	68	864	0	215
Veracruz	1209	1787	140	167	901	252	4126	149	608	1555	7929	71	2921
Yucatán	293	0	15	8	345	49	3786	30	272	1233	7088	0	1611
Zacatecas	180	327	32	19	136	121	439	42	47	342	1959	0	1223
Total	14373	19017	2688	1620	5573	2563	61043	2620	13215	24073	151477	1635	60122

Figura 8.1: Averiguaciones previas iniciadas (Denuncias) ante Agencias del Ministerio Público del fuero común 2002. FUENTE: Sistema de Información Delectiva: La estadística de seguridad pública en México

Entidad Federativa	Privación de la libertad	Robo										
	Secuestro	Robo en Autobús	Robo de Banco	Robo de Camión de Carga	Robo de Casa de Bolsa	Robo de Casa de Cambio	Robo en Casa habitación	Robo en Negocios	Robo a Transeúnte	Robo a Transportistas	Robo de Vehículo	Robo de Ganado
Aguascalientes	2	7	2	2	0	0	777	630	500	4	477	29
Baja California	22	0	15	0	0	0	4089	1566	0	0	23387	0
Baja California Sur	0	0	0	0	0	0	1735	437	9	0	832	40
Campeche	0	5	1	3	0	0	174	107	183	6	43	82
Chiapas	12	0	6	0	0	0	8225	1756	664	284	1203	515
Chihuahua	4	3	15	6	0	22	5852	3704	717	78	7314	309
Coahuila	2	0	3	0	0	0	2704	2401	1283	0	583	81
Colima	1	7	0	7	1	0	428	150	50	1	163	12
Distrito Federal	144	0	83	0	0	0	6753	12348	20960	10150	34475	0
Durango	9	1	0	3	0	0	1331	319	43	9	519	165
Guanajuato	11	0	16	0	0	0	2318	1571	286	127	2847	292
Guerrero	28	19	3	0	0	0	1445	629	3428	357	1799	100
Hidalgo	4	21	0	21	0	0	1543	820	424	33	1330	139
Jalisco	16	0	6	0	0	0	6345	6931	4488	719	10721	466
México	82	0	31	0	0	0	9743	3628	31869	922	33414	473
Michoacán	24	0	12	0	0	0	2337	1566	845	57	1667	203
Morelos	6	34	9	18	0	0	1066	976	1392	0	1633	94
Nayarit	11	0	0	0	0	3	917	379	1	9	272	116
Nuevo León	3	0	20	2	0	0	5569	4396	3526	0	2063	161
Oaxaca	6	0	11	0	0	0	1555	470	738	53	246	141
Puebla	10	28	30	49	0	0	2279	1816	2764	112	3088	137
Querétaro	1	0	2	0	0	0	1252	1148	266	71	1394	98
Quintana Roo	0	0	3	0	1	0	2972	1708	674	31	637	41
San Luis Potosí	2	0	1	0	0	0	1988	972	799	37	1150	0
Sinaloa	12	0	5	0	0	0	1352	544	0	0	2284	217
Sonora	7	0	6	0	0	0	556	406	0	0	857	61
Tabasco	0	0	0	0	0	0	1543	368	4950	0	662	315
Tamaulipas	4	0	8	0	0	0	4220	1752	93	0	2954	166
Tlaxcala	0	2	2	9	0	0	392	379	108	5	457	66
Veracruz	10	23	2	18	0	0	3312	2239	1414	14	1411	504
Yucatán	0	0	0	0	0	0	3477	693	2118	0	302	29
Zacatecas	0	11	0	2	0	3	1008	462	153	1	900	168
Total	433	161	300	140	2	28	85078	57471	84474	13085	141007	5220

Figura 8.2: Averiguaciones previas iniciadas (Denuncias) ante Agencias del Ministerio Público del fuero común 2002. FUENTE: Sistema de Información Delectiva: La estadística de seguridad pública en México

Los datos consisten de 25 tipos de delito denunciados: violación, despojo, homicidio con arma blanca, homicidio con arma de fuego, lesiones con arma blanca, lesiones con arma de fuego, amenazas, estupro, otros delitos sexuales, abuso de confianza, daño en propiedad ajena, extorsión, fraude, secuestro, robo en autobús, robo en banco, robo de camión de carga, robo de Casa de Bolsa, robo de Casa de Cambio, robo en casa habitación, robo a negocios, robo a transeúnte, robo a transportistas, robo de vehículo y robo de ganado. Los datos reportados en el Sistema de Información delictiva son los brutos y debido a que se pretende formar grupos que se puedan comparar entre sí, se tienen que estandarizar las variables para que sean comparables. Se calcula un índice delictivo o de criminalidad por cada 100,000 hab. y población total de la entidad federativa, tabla 8.3. Es decir el número de delitos de una entidad entre la población total de la entidad por 100,000 hab.

$$i_d = \frac{N_{delitos}}{N_{personas}} 100,000_{hb}$$

donde

- i_d Índice de delitos denunciados
- $N_{delitos}$ Número de delitos denunciados por entidad federativa y por tipo de delito
- $N_{personas}$ Número de personas por entidad federativa

Los índices de delitos denunciados que se usaran durante el análisis de seguridad pública, se muestran en las tablas 8.4 y 8.5

Entidad Federativa	1997	1998	1999	2000	2001	2002
Aguascalientes	870379	894348	918977	944285	978189	991555
Baja California	2203516	2294336	2388903	2487367	2578210	2613430
Baja California Sur	388881	402265	411981	424041	439611	445616
Campeche	639803	656334	673292	690689	715907	725586
Chiapas	3692542	3767206	3843279	3920892	4063089	4118600
Chihuahua	2855057	2919546	2985482	3052907	3163870	3207094
Coahuila	2195072	2228981	2263211	2298070	2381476	2414012
Colima	505020	517598	529905	542627	562289	569971
Distrito Federal	8492681	8530036	8567555	8605239	8912675	9034466
Durango	1418139	1429233	1439410	1448661	1500955	1521463
Guanajuato	4447473	4518191	4590038	4663032	4829808	4896806
Guerrero	2934055	2981002	3030329	3079649	3190295	3233807
Hidalgo	2125146	2161335	2198147	2235591	2315894	2347538
Jalisco	5997170	6103549	6211815	6322002	6549412	6638901
México	12011252	12362662	12724090	13096606	12569229	12754627
Michoacán	3849007	3894008	3939579	3985667	4128283	4184595
Morelos	1437076	1475447	1514045	1555296	1611100	1630117
Nayarit	898412	903227	910151	920185	953568	966596
Nuevo León	3596856	3674273	3753356	3834141	3973469	4027754
Oaxaca	3307159	3350451	3394217	3438765	3562075	3610749
Puebla	4778038	4870454	4972499	5076686	5258788	5338646
Querétaro	1287438	1325269	1364214	1404306	1454789	1474667
Quintana Roo	736744	792202	926225	874963	907154	919546
San Luis Potosí	2206135	2236793	2267859	2299380	2382190	2414748
Sinaloa	2432047	2465469	2501419	2536844	2629084	2665002
Sonora	2090792	2132082	2174086	2216969	2297757	2329148
Tabasco	1765187	1805428	1848634	1891829	1960364	1987147
Tamaulipas	2591310	2644194	2698157	2753222	2852950	2891936
Tlaxcala	897192	918499	940313	962646	997349	1010976
Veracruz	6697088	6766965	6837592	6909975	7157495	7255293
Yucatán	1563473	1594435	1626010	1658210	1718254	1741730
Zacatecas	1329926	1337773	1345667	1353610	1402105	1421344
Total general	92225166	93933305	95690699	97483412	100997604	102377645

Figura 8.3: Poblacion Total por entidad federativa. FUENTE: Para 1997,1998 y 1999 proyecciones propias con datos del INEGI. Recuento de poblacion 1995 y censo general de población 2000. Para 2001 y 2002: proyecciones de población 2005-2030 del Consejo Nacional de Población CONAPO

Entidad Federativa			Homicidio		Lesiones		Otros delitos			Patrimonial			
	Violación	Despojo	Homicidio con arma blanca	Homicidio con arma de fuego	Lesiones con arma blanca	Lesiones con arma de fuego	Amenazas	Estupro	Otros delitos sexuales	Abuso de Confianza	Daño en propiedad	Extorsión	Fraude
Baja California	23.30	18.18	0.00	0.00	0.00	0.00	111.77	5.36	50.20	35.74	522.91	3.44	68.61
Baja California Sur	20.42	54.21	0.90	0.90	0.00	0.00	101.21	6.51	31.64	65.30	366.68	4.94	145.42
Campeche	14.47	11.73	1.65	3.45	14.39	1.52	12.27	3.17	10.20	4.55	37.21	0.00	10.89
Chiapas	15.66	16.38	4.66	3.54	12.26	3.42	52.18	5.27	9.85	21.20	101.66	2.67	57.11
Chihuahua	17.09	17.88	5.21	2.56	15.56	12.69	95.35	1.59	16.59	31.12	151.51	1.53	48.99
Coahuila	7.91	11.62	2.61	1.12	6.05	5.01	34.34	0.99	12.39	14.58	197.10	0.00	39.52
Colima	12.46	22.02	1.75	3.06	2.98	2.28	64.04	2.98	8.77	14.56	71.93	0.00	60.35
Distrito Federal	14.37	0.00	4.89	1.39	0.00	0.00	61.09	0.28	22.95	28.92	230.87	3.20	64.54
Durango	8.68	23.18	0.53	6.64	11.30	9.33	93.79	2.56	8.94	48.57	110.50	1.12	86.10
Guanajuato	6.76	24.39	2.17	1.08	4.96	4.15	58.66	1.29	8.05	22.02	210.20	1.06	44.18
Guerrero	13.05	18.14	2.26	13.08	7.61	5.66	31.26	2.60	5.88	15.37	86.58	2.32	32.65
Hidalgo	13.67	37.20	2.56	0.90	9.93	4.39	58.66	3.11	10.99	17.34	96.14	0.89	35.48
Jalisco	12.00	33.23	0.00	0.00	0.00	0.00	76.01	2.86	8.71	24.09	123.35	3.81	100.56
México	18.42	19.58	0.00	0.00	0.00	0.00	0.00	1.52	0.00	17.25	105.04	0.00	25.56
Michoacán	6.09	16.26	6.71	1.48	5.40	3.87	12.78	3.23	6.28	14.29	63.16	3.87	48.63
Morelos	18.00	45.14	3.80	0.73	0.00	0.00	206.42	3.86	13.53	31.90	168.57	6.61	80.40
Nayarit	9.63	12.44	7.35	1.97	6.28	5.28	33.73	5.07	9.41	10.55	31.24	0.00	30.73
Nuevo León	8.99	31.87	1.44	0.79	0.00	0.00	56.04	2.58	29.60	23.26	154.70	0.79	115.52
Oaxaca	8.78	26.65	0.00	0.00	0.00	0.00	82.89	3.54	16.98	11.60	99.34	2.80	47.52
Puebla	13.06	29.39	4.31	1.63	15.55	4.26	89.41	2.83	9.12	20.05	110.02	0.00	75.26
Querétaro	14.24	23.84	0.47	0.61	3.19	1.42	39.81	1.15	9.22	27.74	172.78	0.54	77.17
Quintana Roo	45.57	69.73	2.07	2.39	40.35	4.68	45.13	4.35	37.95	64.49	310.59	0.65	159.10
San Luis Potosí	16.73	0.00	3.02	1.16	0.00	0.00	141.42	3.77	39.51	45.84	256.59	0.54	112.56
Sinaloa	7.65	18.85	14.67	1.31	8.37	8.78	43.68	3.19	8.03	17.26	34.03	3.45	30.54
Sonora	9.57	16.42	3.13	1.09	0.00	0.00	20.01	3.74	11.39	10.09	119.92	0.73	25.93
Tabasco	13.94	24.47	0.00	0.00	0.00	0.00	122.24	8.15	18.92	25.72	136.58	1.26	76.44
Tamaulipas	19.40	6.01	3.39	1.14	6.57	1.87	78.22	3.25	14.25	19.85	91.81	0.31	62.90
Tlaxcala	22.26	10.13	0.09	0.20	0.40	0.30	10.78	0.00	13.05	6.73	36.00	0.00	21.27
Veracruz	16.66	26.41	1.93	2.30	12.42	3.47	56.87	2.05	8.38	21.43	109.29	0.98	40.26
Yucatán	16.82	0.00	0.86	0.46	19.87	2.81	217.37	1.72	15.62	70.79	406.95	0.00	92.49
Zacatecas	12.66	24.44	2.25	1.34	9.57	9.51	30.09	2.95	3.31	24.06	130.79	0.00	86.05
Total	14.04	20.24	2.63	1.58	5.44	2.50	59.63	2.56	12.91	23.51	147.96	1.60	58.73

Figura 8.4: Índice de delitos denunciados (Delitos denunciados por entidad federativa por cada 100,000 habitantes 2002).

Entidad Federativa	Privación de la libertad	Robo										
	Secuestro	Robo en Autobús	Robo en Bancos	Robo de Camión de Carga	Robo de Casa de Bolsa	Robo de Casa de Cambio	Robo en Casa habitación	Robo en Negocios	Robo a Transeúnte	Robo a Transportistas	Robo de Vehículo	Robo de Ganado
Baja California	0.94	0.00	0.57	0.00	0.00	0.00	187.45	59.92	0.00	0.00	694.89	0.00
Baja California Sur	0.00	0.00	0.00	0.00	0.00	0.00	380.35	98.07	2.02	0.00	186.71	8.98
Campeche	0.00	0.69	0.14	0.41	0.00	0.00	23.98	14.75	25.22	0.83	5.93	11.30
Chiapas	0.29	0.00	0.15	0.00	0.00	0.00	79.33	42.64	16.12	6.90	29.21	12.50
Chihuahua	0.12	0.09	0.47	0.19	0.00	0.69	182.47	115.49	22.36	2.43	228.06	9.63
Coahuila	0.08	0.00	0.12	0.00	0.00	0.00	112.01	99.45	53.15	0.00	24.15	3.36
Colima	0.18	1.23	0.00	1.23	0.18	0.00	75.09	26.32	8.77	0.18	28.60	2.11
Distrito Federal	1.59	0.00	0.92	0.00	0.00	0.00	74.85	136.68	232.00	112.35	381.59	0.00
Durango	0.59	0.07	0.00	0.20	0.00	0.00	87.48	20.97	2.83	0.59	34.11	10.84
Guanajuato	0.22	0.00	0.33	0.00	0.00	0.00	47.35	32.09	4.21	2.59	58.15	5.96
Guerrero	0.87	0.59	0.09	0.00	0.00	0.00	44.60	19.45	106.00	11.04	55.63	3.09
Hidalgo	0.17	0.89	0.00	0.89	0.00	0.00	65.73	34.93	18.06	1.41	56.78	5.92
Jalisco	0.24	0.00	0.09	0.00	0.00	0.00	95.57	104.40	67.60	10.83	161.49	7.02
México	0.60	0.00	0.23	0.00	0.00	0.00	70.80	26.36	230.24	6.70	242.93	3.44
Michoacán	0.57	0.00	0.29	0.00	0.00	0.00	55.85	37.47	20.22	1.36	39.84	4.85
Morelos	0.37	2.08	0.55	1.10	0.00	0.00	65.27	59.76	85.24	0.00	99.99	5.76
Nayarit	1.14	0.00	0.00	0.00	0.00	0.31	94.87	39.21	0.10	0.93	28.14	12.00
Nuevo León	0.07	0.00	0.50	0.05	0.00	0.00	138.27	109.14	87.54	0.00	51.22	4.00
Oaxaca	0.17	0.00	0.30	0.00	0.00	0.00	43.07	13.02	20.44	1.61	6.81	3.91
Puebla	0.19	0.53	0.56	0.92	0.00	0.00	42.75	34.07	51.85	2.10	56.43	2.57
Querétaro	0.07	0.00	0.14	0.00	0.00	0.00	84.90	77.85	18.04	4.81	94.53	0.65
Quintana Roo	0.00	0.00	0.33	0.00	0.11	0.00	323.20	185.74	73.30	3.37	69.27	4.46
San Luis Potosí	0.08	0.00	0.04	0.00	0.00	0.00	82.33	40.25	39.09	1.53	47.62	0.00
Sinaloa	0.45	0.00	0.19	0.00	0.00	0.00	50.73	20.41	0.00	0.00	85.70	8.14
Sonora	0.30	0.00	0.26	0.00	0.00	0.00	23.87	17.43	0.00	0.00	36.79	2.62
Tabasco	0.00	0.00	0.00	0.00	0.00	0.00	77.65	18.52	249.10	0.00	33.31	15.85
Tamaulipas	0.14	0.00	0.28	0.00	0.00	0.00	145.92	60.59	3.22	0.00	102.15	5.74
Tlaxcala	0.00	0.20	0.20	0.89	0.00	0.00	38.77	37.49	10.68	0.49	45.20	6.53
Veracruz	0.14	0.32	0.03	0.25	0.00	0.00	45.65	30.85	19.49	0.19	19.45	6.95
Yucatán	0.00	0.00	0.00	0.00	0.00	0.00	199.63	51.27	121.80	0.00	17.34	1.67
Zacatecas	0.00	0.77	0.56	0.14	0.00	0.21	70.92	32.50	10.76	0.07	63.32	11.82
Total	0.42	0.16	0.29	0.14	0.00	0.03	83.10	56.14	82.51	12.78	137.73	5.10

Figura 8.5: Índice de delitos denunciados (Delitos denunciados por entidad federativa por cada 100,000 habitantes 2002).

8.3. Primer ejemplo

8.3.1. Análisis de Escalas Multidimensionales de Entidades Federativas con SPSS

El Análisis de Escalas Multidimensionales MDS es una técnica exploratoria en donde las proximidades son valores que indican la cercanía, objetiva o subjetiva, entre dos o más objetos, obteniendo un mapa de estímulos en un espacio de varias dimensiones. MDS es un conjunto de técnicas estadísticas relacionadas que son frecuentemente usadas para la visualización de los datos para explorar similaridades o disimilaridades en los datos. Un algoritmo MDS comienza con una matriz de similaridades objeto-objeto, después asigna un lugar a cada objeto en un espacio de menor dimensión. Es apropiado para graficar o para una visualización en 3D.

El procedimiento básico de MDS es:

1. Formular el problema: seleccionar una medida de distancia; existen varias formas de calcular la distancia.

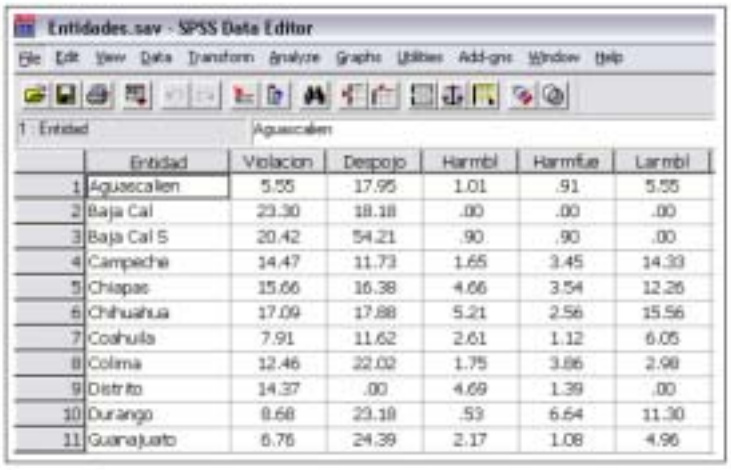
2. Seleccionar los grupos a los que se desea aplicar la técnica de MDS.

3. Ejecutar el programa estadístico para MDS. Existen generalmente dos opciones: Métrica MDS (el cual maneja datos a nivel intervalo o de razón) y No-Métrica MDS (el cual maneja datos a nivel ordinal). El investigador decide el número de dimensiones que desee crear. Entre más dimensiones es mejor el ajuste estadístico, pero más difícil es la interpretación de resultados.

4. Mapear los resultados y definir las dimensiones. El programa mapea los resultados. El mapa localiza los puntos por medio de coordenadas (usualmente en un espacio de dos dimensiones). Las proximidades entre los objetos indican qué tan similares son entre ellos. Los resultados se tienen que interpretar.

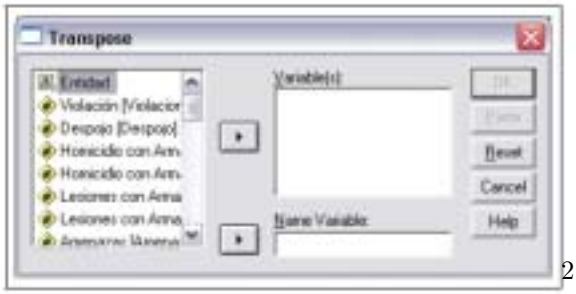
5. Prueba de resultados para validar.

La aplicación del procedimiento de Escalas Multidimensionales MDS servirá para conocer las ventanas de diálogo del programa, así como el modo de proceder para llevar a cabo el análisis e interpretar los resultados. El ejemplo que se realizará es una tabla que contiene diversos tipos de delitos denunciados ante agencias del ministerio público de las 32 entidades federativas de la República Mexicana (Aguascalientes, Baja California Sur, Baja California Norte, Campeche, Chiapas, Chihuahua, Coahuila, Colima, Distrito Federal, Durango, Guanajuato, Guerrero, Hidalgo, Jalisco, México, Michoacán, Morelos, Nayarit, Nuevo León, Oaxaca, Puebla, Querétaro, Quintana Roo, San Luis Potosí, Sinaloa, Sonora, Tabasco, Tamaulipas, Tlaxcala, Veracruz, Yucatán y Zacatecas). Para cada estado se presentan 25 tipos de delitos denunciados ante ministerio público: violación (Violacion), despojo (Despojo), homicidio con arma blanca (Harmbl), homicidio con arma de fuego (Harmfue), lesiones con arma blanca (Larmbl), lesiones con arma de fuego (Larmfue), amenazas (Amenazas), estupro (Estupro), otros delitos sexuales (Odelsex), abuso de confianza (Abconf), daño en propiedad ajena (Daproaj), extorsión (Extorsion), fraude (Fraude), secuestro (Secuestro), robo en autobús (Robaut), robo en banco (Robban), robo de camión de carga (Robcamc), robo de Casa de Bolsa (Robcasb), robo de Casa de Cambio (Robcascam), robo en casa habitación (Robcashab), robo a negocios (Robneg), robo a transeúnte (Robtran), robo a transportistas (Robtrans), robo de vehículo (Robveh) y robo de ganado (Robgan).



	Entidad	Violacion	Despojo	Harmbl	Harmfue	Larmbl
1	Aguascalien	5.55	17.95	1.01	.91	5.55
2	Baja Cal	23.30	18.18	.00	.00	.00
3	Baja Cal S	20.42	54.21	.90	.90	.00
4	Campeche	14.47	11.73	1.65	3.45	14.33
5	Chiapas	15.66	16.38	4.66	3.54	12.26
6	Chihuahua	17.09	17.88	5.21	2.56	15.56
7	Coahuila	7.91	11.62	2.61	1.12	6.05
8	Colima	12.46	22.02	1.75	3.86	2.98
9	Distrito	14.37	.00	4.69	1.39	.00
10	Durango	8.68	23.18	.53	6.64	11.30
11	Guanajuato	6.76	24.39	2.17	1.08	4.96

Figura 8.6: Ventana del Editor de Datos con las variables de los 32 estados



2

Figura 8.7: Opción para transponer datos

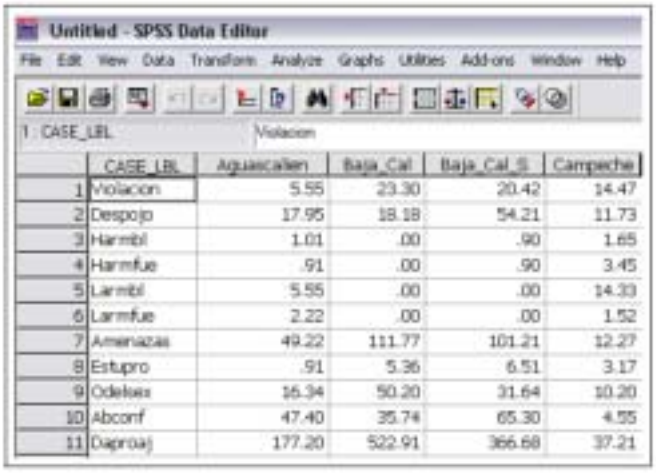
Una vez que se tiene la base de datos en SPSS, se verifica la estructura de las columnas, etiquetas, las escalas y los valores perdidos.

Procedimento

Dado que los datos originales no son proximidades, se tendrá que obtener a partir de ellos un coeficiente de disimilaridad entre estados. Un procedimiento sencillo consiste en obtener la matriz de correlaciones entre los países, y transformar luego éstas en disimilaridades. Para obtener una matriz de correlaciones (ver apéndice A) entre países, primero se debe transponer la matriz de datos que se muestra en la figura 8.6, de forma que los estados pasen a ser las columnas de la nueva matriz, mientras que los tipos de delito denunciados se encuentren en las filas. Para ello se utilizara el FLIP de SPSS. Las opciones dentro de la barra de herramientas que serán útiles para este propósito son: *Data-Transpose*

Esto abrirá una ventana de dialogo, donde es posible definir las variables que se quieren usar en el análisis. Las variables se especifican bajo la lista **Variables(s)** con la ayuda de los botones "mover variables de un lugar a otro". Se seleccionaran todas las variables respecto el tipo de delitos denunciados bajo la opción variables y la entidad se posiciona en **Name Variable** y se da aceptar ver figura 8.7. La posición de las variables se muestra en la figura 8.8.

A continuación, podemos calcular la matriz de correlaciones a SPSS y guardar el archivo, al que llamaremos $\backslash Entidades_2.sav$". La sintaxis correspondiente sería la siguiente:



	CASE_LBL	Aguascalien	Baja_Cal	Baja_Cal_S	Campeche
1	Violacion	5.55	23.30	20.42	14.47
2	Despojo	17.95	18.18	54.21	11.73
3	Harmbl	1.01	.00	.90	1.65
4	Harmfue	.91	.00	.90	3.45
5	Larmbl	5.55	.00	.00	14.33
6	Larmfue	2.22	.00	.00	1.52
7	Amenazas	49.22	111.77	101.21	12.27
8	Estupro	.91	5.36	6.51	3.17
9	Odelaix	16.34	50.20	31.64	10.20
10	Abconf	47.40	35.74	65.30	4.55
11	Daproaj	177.20	522.91	366.68	37.21

Figura 8.8: Vista de los datos transpuestos

```
CORRELATIONS
/VARIABLES =Aguascalien Baja_Cal Baja_Cal_S Campeche Chiapas
Chihuahua Coahuila Colima Distrito Durango Guanajuato Guerrero
Hidalgo Jalisco México Michoacán Morelos Nayarit Nuevo_Leon
Oaxaca Puebla Querétaro Quintana San_Luis Sinaloa
Sonora Tabasco Tamaulipas Tlaxcala Veracruz Yucatán Zacatecas
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE
/MATRIX=OUT (Entidades_2.sav).
```

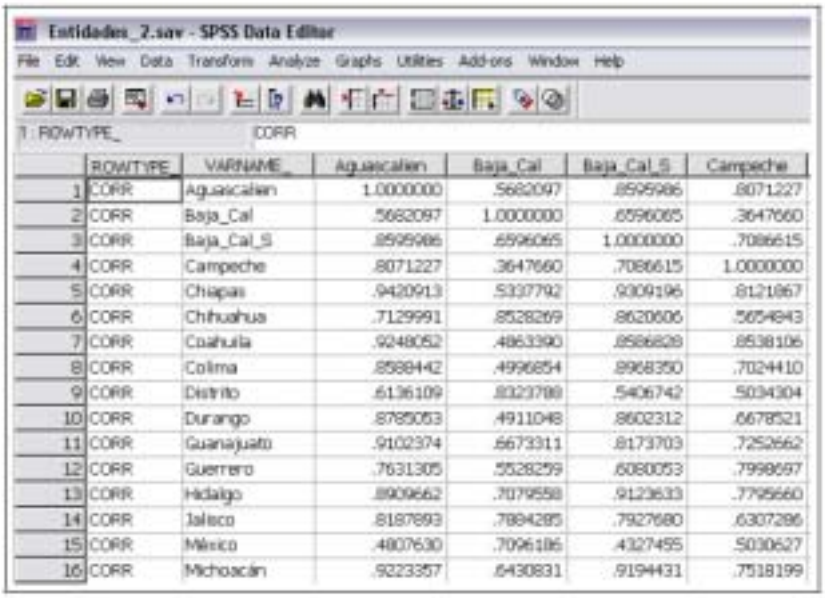
El archivo "$Entidades_2.sav$" contendrá varias filas con estadísticos descriptivos, seguidas de la matriz de correlaciones. Si eliminamos estas filas innecesarias el archivo tendrá el aspecto que muestra la figura 8.9.

El editor de datos muestra, a la izquierda, dos variables de cadena (que sólo contienen texto) creadas por SPSS. Ambas son variables especiales, y se identifican porque sus nombres terminan en un carácter de subrayado. La primera de ellas contiene información sobre el tipo de datos del archivo. En este caso, aparece la etiqueta "CORR" que indica que es una matriz de correlaciones. La segunda variable contiene el nombre de los 32 estados. Como se puede ver en la figura 8.23, existen correlaciones altas entre algunos estados, como la existente entre Aguascalientes y Chiapas (.942), o entre Chiapas e Hidalgo, y correlaciones casi nulas, como la existente entre Baja California Norte y Campeche (.364). Sin embargo, resulta muy difícil interpretar los resultados directamente a partir de la matriz de correlaciones. Se utilizará SPSS para obtener una representación espacial de los estados e interpretar a partir de ellas las relaciones existentes entre los mismos.

Para ello, todavía debemos llevara a cabo una última transformación. Aunque algunos programas de MDS aceptan matrices cuadradas de correlaciones o de varianzas-covarianzas como entrada en SPSS sólo se acepta como entrada una matriz de disimilaridades. Debemos, por tanto, llevar a cabo una transformación de las correlaciones para convertirlas en disimilaridades. Una de las más conocidas es la siguiente:

$$d_{ij} = \sqrt{2(1 - r_{ij})}$$

La sintaxis correspondiente a la transformación deseada será la siguiente:



	ROWTYPE_	VARNAME_	Aguascalien	Baja_Cal	Baja_Cal_S	Campeche
1	CORR	Aguascalien	1.0000000	.5682097	.8595986	.8071227
2	CORR	Baja_Cal	.5682097	1.0000000	.6596065	.3647660
3	CORR	Baja_Cal_S	.8595986	.6596065	1.0000000	.7086615
4	CORR	Campeche	.8071227	.3647660	.7086615	1.0000000
5	CORR	Chiapas	.9420913	.5337792	.9309196	.8121867
6	CORR	Chihuahua	.7129991	.8528269	.8620606	.5654943
7	CORR	Coahuila	.9248052	.4863390	.8586828	.8538106
8	CORR	Colima	.8588442	.4996854	.8968350	.7024410
9	CORR	Distrito	.6136109	.8323788	.5406742	.5034304
10	CORR	Durango	.8785053	.4911048	.8602312	.6678521
11	CORR	Guanajuato	.9102374	.6673311	.8173703	.7252662
12	CORR	Guerrero	.7631305	.5528259	.6080053	.7998697
13	CORR	Hidalgo	.8909662	.7079558	.9123633	.7795660
14	CORR	Jalisco	.8187893	.7884285	.7927680	.6307286
15	CORR	México	.4807630	.7096186	.4327455	.5030627
16	CORR	Michoacán	.9223357	.6430831	.9194431	.7518199

Figura 8.9: Matriz de correlaciones entre países

```
COMPUTE Aguascalien =SQRT(2*(1-Aguascalien)).
COMPUTE Baja_Cal =SQRT(2*(1-Baja_Cal)).
COMPUTE Baja_Cal_S =SQRT(2*(1-Baja_Cal_S)).
COMPUTE Campeche =SQRT(2*(1-Campeche)).
COMPUTE Chiapas =SQRT(2*(1-Chiapas)).
COMPUTE Chihuahua =SQRT(2*(1-Chihuahua)).
COMPUTE Coahuila =SQRT(2*(1-Coahuila)).
COMPUTE Colima =SQRT(2*(1-Colima)).
COMPUTE Distrito =SQRT(2*(1-Distrito)).
COMPUTE Durango =SQRT(2*(1-Durango)).
COMPUTE Guanajuato =SQRT(2*(1-Guanajuato)).
COMPUTE Guerrero =SQRT(2*(1-Guerrero)).
COMPUTE Hidalgo =SQRT(2*(1-Hidalgo)).
COMPUTE Jalisco =SQRT(2*(1-Jalisco)).
COMPUTE México =SQRT(2*(1-México)).
COMPUTE Michoacán =SQRT(2*(1-Michoacán)).
COMPUTE Morelos =SQRT(2*(1-Morelos)).
COMPUTE Nayarit =SQRT(2*(1-Nayarit)).
COMPUTE Nuevo_Leon =SQRT(2*(1-Nuevo_Leon)).
COMPUTE Oaxaca =SQRT(2*(1-Oaxaca)).
COMPUTE Puebla =SQRT(2*(1-Puebla)).
COMPUTE Querétaro =SQRT(2*(1-Querétaro)).
COMPUTE Quintana =SQRT(2*(1-Quintana)).
COMPUTE San_Luis =SQRT(2*(1-San_Luis)).
COMPUTE Sinaloa =SQRT(2*(1-Sinaloa)).
COMPUTE Sonora =SQRT(2*(1-Sonora)).
COMPUTE Tabasco =SQRT(2*(1-Tabasco)).
COMPUTE Tamaulipas =SQRT(2*(1-Tamaulipas)).
COMPUTE Tlaxcala =SQRT(2*(1-Tlaxcala)).
COMPUTE Veracruz =SQRT(2*(1-Veracruz)).
COMPUTE Yucatán =SQRT(2*(1-Yucatán)).
```

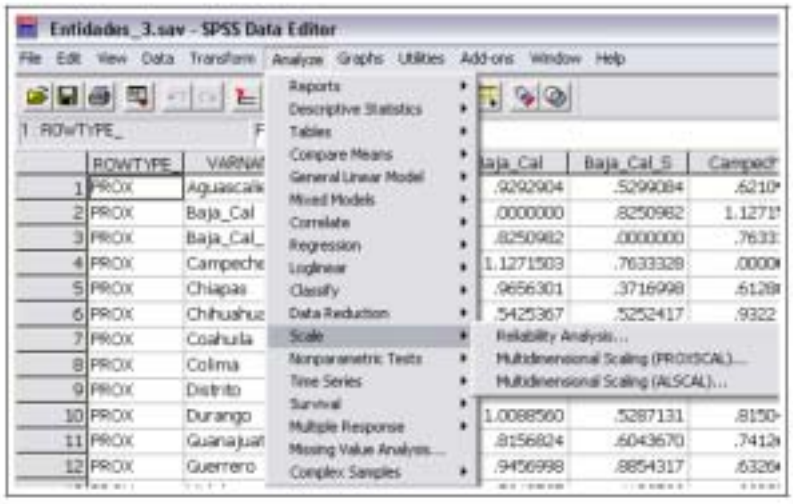


Figura 8.10: Opciones para el análisis de escalas multidimensionales por el método ALSCAL

```
COMPUTE Zacatecas =SQRT(2*(1-Zacatecas)).
COMPUTE rowtype_="PROX".
EXECUTE.
```

Los primeros treinta y dos comandos efectuarán la transformación en los valores de la correlación para cada país. Por su parte, el penúltimo comando sirve para indicar a SPSS, a través de la variable rowtype que los datos son ahora proximidades ("PROX") y no correlaciones ("CORR"). Ahora es posible ejecutar el procedimiento de escalamiento en SPSS.

Una vez que ya se tengan los datos arreglado de esta manera se procede al análisis. Para llevar a cabo el análisis se usara el Escalamiento Multidimensional Clásico o Escalamiento Multidimensional Métrico por medio del algoritmo ALSCAL para obtener una solución apropiada. Una vez que se tiene la base en orden los datos se puede pasar a la ejecución del procedimiento estadístico. Para esto se tiene que elegir **Analyze** de la barra de menúes, se selecciona la categoría **Scale- Multidimensional Scaling** (ALSCAL). Como se muestra en la figura 8.10.

Esto abrirá una ventana de dialogo, donde es posible definir las variables que se quieren usar en el análisis de escalas multidimensionales. Las variables se especifican bajo la lista **Variables(s**) con la ayuda de los botones "mover variables de un lugar a otro". Se seleccionarán las variables estado que se desean analizar respecto el tipo de delitos denunciados, esto porque nos interesa comparar los estados y ver si existe algún tipo de agrupamiento natural. Adicionalmente a esto, en esta ventana permite especificar si existe más de una matriz de entrada. Así como indicar si los datos de entrada son disimilaridades o no. En caso de que sea necesario calcular las distancias a partir de los datos de perfil existe la opción crear las distancias a partir de los datos **(Create distance from data)**.

Este cuadro de diálogo inicial nos permite especificar, en primer lugar, las variables que serán objeto de análisis. En segundo lugar, nos permite especificar si existe más de una matriz de entrada. Finalmente, también nos permite indicar si los datos de entrada son disimilaridades (distancias), o si es necesario calcular distancias a partir de los datos. En el caso de que nuestros datos de entrada sean disimilaridades, existe un botón, etiquetado como **Shape** , que permitirá espe-

Figura 8.11: Opciones del cuadro de diálogo principal para MDS.

Figura 8.12: Ventana de diálogo de modelos.

cificar la forma de la matriz de datos. En caso de que sea necesario calcular las distancias a partir de datos de perfil existe otro botón, etiquetado como Measure, que nos permitirá especificar la medida de distancia que deseamos. También aparecen en el cuadro de diálogo dos botones etiquetados **Model** y **Options**, figura 8.11.

Lo que primero se tiene que hacer es introducir en la casilla etiquetada Variables las 32 variables correspondientes a los estados. La opción por defecto del cuadro de diálogo es que nuestros datos son distancias, y que nuestra matriz es cuadrada simétrica. Dado que es así, no cambiaremos nada en la casilla etiquetada como **Distances**. Se pulsa a continuación el botón **Modelo**. Aparecerá el cuadro de diálogo que se muestra en la figura 8.12.

Este cuadro de diálogo permite especificar el nivel de medida de los datos, el modelo de MDS a emplear, la condicionalidad de los datos y el número de dimensiones que queremos que tenga la solución. En este paso sólo efectuaremos un cambio. En la casilla etiquetada **Level of Measurement** especificaremos **Interval** como se muestra en la figura 8.12. A continuación pulsaremos el botón **Continue**. Una vez de nuevo en el cuadro de diálogo anterior, pulsaremos el botón **Opciones**. Aparecerá el cuadro de diálogo mostrado en la figura 8.12.

En este cuadro de diálogo se puede pedir que el procesamiento incluya alguna información adicional en el listado de salida, así como también cambiar los criterios de convergencia del análisis. También aquí sólo efectuaremos un cambio. En la casilla etiqueta **Display** solicitaremos **Group plots**, tal y como se muestra en la figura 8.13. A continuación se pulsará el botón **Continue**, lo que

Figura 8.13: Ventana de diálogo de opciones.

```
Iteration history for the 2 dimensional solution (in squared distances)

                Young's S-stress formula 1 is used.

            Iteration      S-stress      Improvement

                 1          .18852
                 2          .14148         .04712
                 3          .13995         .00146
                 4          .13904         .00910

                    Iterations stopped because
              S-stress improvement is less than   .001000

          Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data
(disparities)
          in the partition (row, matrix, or entire data) which
          is accounted for by their corresponding distances.
           Stress values are Kruskal's stress formula 1.

              For  matrix
   Stress  =   .15412      RSQ =  .92697
```

Figura 8.14: Primera parte de los resultados de MDS algoritmo ALSCAL.

devolverá al cuadro de diálogo inicial. En éste se pulsara el botón **Ok**, con lo que se realizará el análisis. El siguiente paso es interpretar la solución proporcionada por el programa.

Si ahora si se analiza el editor de resultados de SPSS, se encontrará primero, un listado con el resumen del procedimiento. En este resumen aparece, en primer lugar, un listado con el proceso de convergencia hacia la solución del análisis como se muestra en la figura 8.14.

En este listado se muestra la minimización de un índice de ajuste denominado S-stress mediante un proceso iterativo, que se detiene al no conseguir una mejora superior al 0.001. A continuación se muestran los valores de dos índices de ajuste: Stress y RSQ. Ambos indican el ajuste de la solución proporcionada, pero en sentido inverso. El Stress es un indicador de "maldad" de ajuste, por tanto éste será mejor cuanto más próximo sea su valor a cero. En este caso, su valores 0.154. Por su parte, la RSQ es un indicador de bondad de ajuste, mayor cuanto más próximo sea su valor a uno. En este caso, su valor es 0.926. Ambos índices indican, pues, un ajuste bueno para nuestros datos.

A continuación se muestran las coordenadas de los 32 estados en las dos dimensiones. La primera y segunda columnas identifican a los 32 estados mientras que la tercera y la cuarta columna representan respectivamente sus coordenadas en el eje horizontal y el vertical. Por conveniencia las coordenadas se han normalizado, de modo que, para una dimensión dada, la medida de las coordenadas siempre vale cero, y su desviación típica siempre vale 1. La representación gráfica de estas coordenadas es mostrada en la figura 8.15.

En la configuración formada se puede apreciar que aquellos países entre lo

Figura 8.15: Segunda parte de los resultados de MDS algoritmo ALSCAL.

que existían altas correlaciones se encuentran próximos entre sí, como es el caso de Querétaro, Zacatecas, Aguascalientes, Chiapas, Hidalgo, Nuevo León, Michoacán, Quintana Roo, Baja California Sur. Por otra parte, aquellos países cuyas correlaciones fueron casi nulas se encuentran alejados entre sí, como es el caso de Estado de México, Distrito Federal, Tabasco y Guerrero. Se observan tres grupos grandes y dos pequeños. Un grupo es el antes mencionado, un segundo grupo estaría formado por Jalisco, Sinaloa, Tlaxcala, Chihuahua, Nayarit y Tamaulipas. El tercer grupo esta formado por Campeche, Morelos, Oaxaca, Yucatán, San Luis Potosí, Veracruz, Puebla, Guanajuato, Coahuila y Durango. Tabasco y Guerrero parecen aislados, separados de ambos grupos, asímismo los estados de Estado de México, Distrito Federal y Baja California Norte se encuentran separados de los mencionados grupos, como se puede observar en la figura 8.16. Un segundo examen revelará que los dos de los grupos grandes se alinean a los largo de la primera dimensión positiva (horizontal), mientras que los estados Distrito Federal y Baja California Norte se alinean a lo largo de la primera dimensión, también, pero en sentido opuesto (negativa).

Si examinamos los índices totales de delitos denunciados veremos que el grupo mas grande esta conformado por estados con índices menores de delitos, el segundo grupo esta formado por estados con índices de delitos mayores y los estados que representan los puntos aislados son aquellos en donde los índices de delitos son muy altos y diferentes que el resto de la republica mexicana.

Antes se comentó que los dos índices de ajuste proporcionados por el listado de MDS (Stress y RSQ) indican un buen ajuste de los datos a las distancias derivadas a partir de la matriz de coordenadas. Este ajuste también puede apreciarse visualmente. En la figura 8.17 se presenta un diagrama de dispersión que relaciona las proximidades contenidas en la matriz de datos de entrada con las distancias existentes entre los estados representados en la solución. En el diagrama de dispersión no se muestran los valores originales de las disimilaridades, sino una transformación lineal de éstas, llamadas disparidades. Como puede verse, los puntos 300 (que representan todos los pares posibles de distancias entre 32 estados-estímulos) se encuentran casi alineados, lo que indica que el ajuste

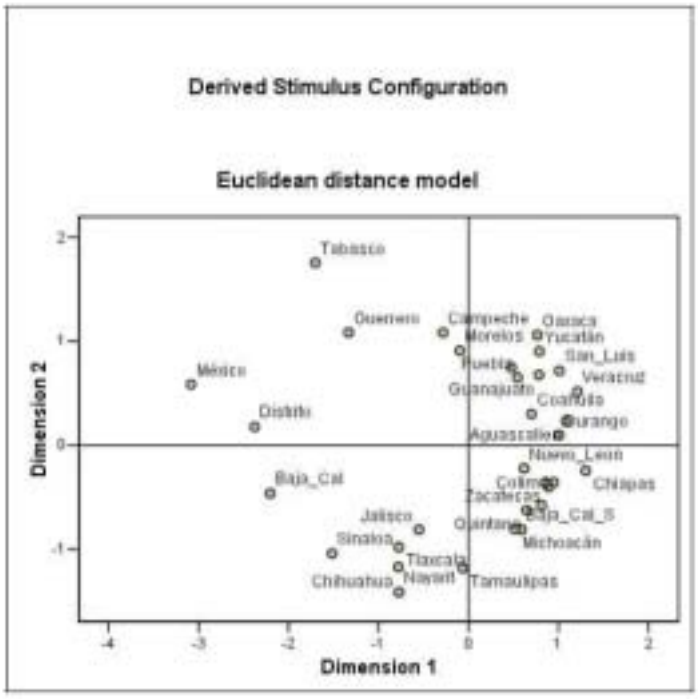


Figura 8.16: Representación en dos dimensiones de las coordenadas de los treinta y dos entidades federativas..

encontrado es bueno, en el mismo sentido que lo indicaban los valores de Stress y RSQ.

Este ejemplo ha servido para ilustrar que es posible utilizar MDS con datos que inicialmente parecen pensados para otro tipo de técnicas más tradicionales, como el análisis factorial. También sirvió para ver una de las formas de obtener las solución recurriendo a la formación existente en la matriz de datos original. Sin embargo, la técnica posee muchas más posibilidades.

Como antes se había mencionado, un procedimiento de gran utilidad para interpretar las soluciones MDS es buscar los agrupamientos de estímulos. Estos agrupamientos indicarán conjuntos de estímulos muy semejantes entre sí y diferentes a los demás, y pueden ser de utilidad si la finalidad principal es la clasificación como en este caso.

8.3.2. Análisis de Conglomerados de entidades federativas con SPSS

Se sabe que el Análisis de Conglomerados tiene como objetivo principal particionar un conjunto de individuos de acuerdo con ciertas características, tal que se puedan formar grupos o conglomerados similares en forma general. El Análisis de Conglomerados es una técnica de interdependencia, es decir que no hace ninguna distinción entre variables dependientes e independientes. El conjunto entero de relaciones interdependientes es examinado y se reduce el número de observaciones o casos mediante la agrupación de ello en un conjunto pequeño de conglomerados (Capítulo 5).

El procedimiento básico del Análisis de Conglomerados es:

1. Formular el problema: seleccionar los individuos a los que se desea aplicar la técnica de conglomerados así como codificar las variables si es necesario.

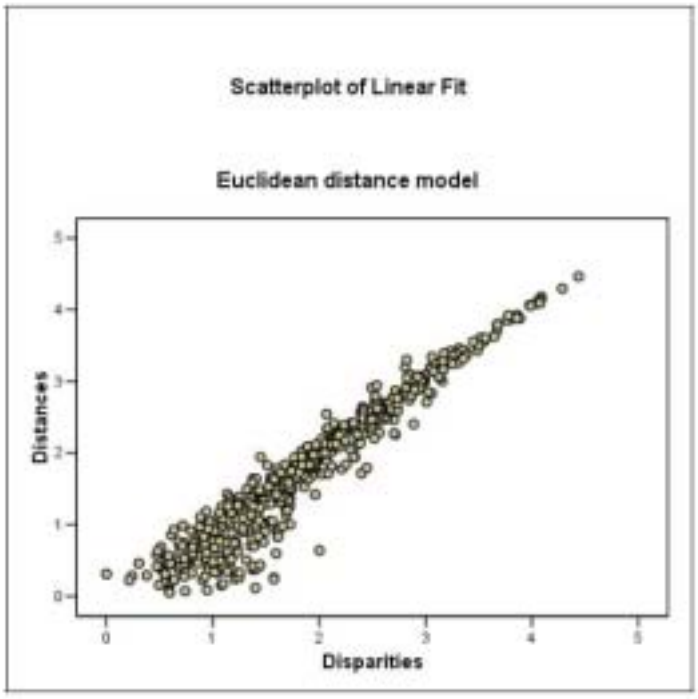


Figura 8.17: Gráfico de ajuste lineal entre datos (disparidades) y distancias.

2. Seleccionar una medida de distancia: existen varias formas de calcular la distancia.
3. Seleccionar un procedimiento de conglomerado: una vez que se selecciono la medida de distancia es necesario elegir el algoritmo con que se construirán los conglomerados
4. Decidir el número de conglomerados
5. Interpretar los conglomerados: sacar conclusiones con la ayuda de mapas perceptúales y dendogramas son útiles.
6. Evaluar fiabilidad y validez: repetir el análisis pero usando diferentes medidas de distancia, repetir el análisis con el uso de diferentes técnicas de conglomerado, dividir los datos aleatoriamente en dos y analizar cada parte separadamente.

Antes de realizar cualquier procedimiento estadístico es necesario revisar la estructura de los datos así como las variables que se analizaran, algunos puntos que se tienen que revisar son:

- Número de variables más de 15 menos de 35, en este caso se tienen 25 variables.
- Tipo de variables (consistencia).
- Número de casos 32

Una vez que se tiene la base de datos en SPSS, se verifica la estructura de las columnas, etiquetas, las escalas y los valores perdidos.

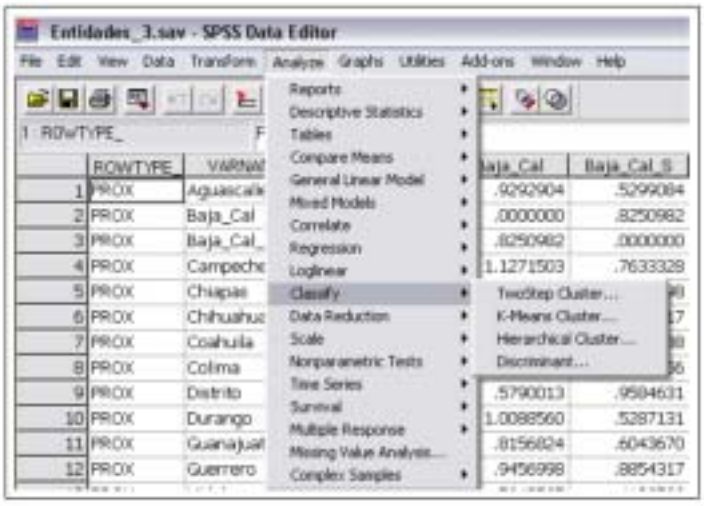


Figura 8.18: Comando de procedimientos estadísticos de clasificación para un Conglomerado Jerárquico.

Procedimiento

Se usarán dos métodos de Análisis de Conglomerados para obtener una solución apropiada. El primero será el **Hierarchical Cluster** (Conglomerados jerárquicos) que ayudara a determinar el número de conglomerados, porque permite visualizar las posibles estructuras por medio de un dendograma de los conglomerados. El segundo será el **K-Means Cluster** (Conglomerado de k-medias) donde se especificaran el número de conglomerados en que se desea dividir la muestra1. Una vez que se tiene la base en orden los datos se puede pasar a la ejecución del procedimiento estadístico. Para esto se tiene que elegir **Analyze** de la barra de menús, se selecciona la categoría **Classify- Hierarchical Cluster** (figura 8.18).

Esto abrirá una ventana de dialogo, donde es posible definir las variables que se quieren usar en el análisis de Conglomerados. Las variables se especifican bajo la lista **Variables(s)** con la ayuda de los botones "mover variables de un lugar a otro". Se seleccionaran las variables de tipos delitos denunciados que se desean analizar respecto a las entidades federativas. Puesto que esta etapa del análisis se enfocara al conglomerado por casos (entidades federativas) en lugar de variables, se selecciona **Cases** bajo la opción **Cluster** (figura 8.19) y se marca el cuadrado de **Statistics** bajo la opción **Display**.

Cuando se pulsa el botón Statistics que abrirá otra ventana de subdiálogo que permitirá seleccionar tablas y diagramas como se puede apreciar en la figura 8.20.

La opción **Agglomeration Schedule** muestra los casos o conglomerados combinados en cada etapa, las distancias entre los casos o los conglomerados que se van combinando, así como el último nivel del proceso de aglomeración en el que cada caso se una al conglomerado correspondiente. La opción **Proximity matrix** proporciona la matriz de distancias entre los elementos. La opción **Cluster Membership** muestra el conglomerado al cual se asigna cada caso en una o varias etapas de la combinación de los conglomerados.

En este caso únicamente seleccionaremos, dentro de la ventana de subdiálogo **Statistics**:**Agglomeration Schedule- None- Continue**

Recordemos que lo único que nos interesa en esta fase del proceso es tener una idea estructural de los posibles conglomerados en que dividiremos a la

Figura 8.19: Declaración de un Análisis de Conglomerados Jerárquico.

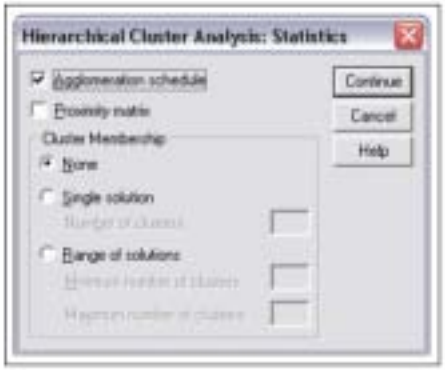


Figura 8.20: Especificando resultado de un Análisis de Conglomerado Jerárquico.

Figura 8.21: Solicitando un Dendograma.

población, cuando se generen los conglomerados por medio de k-medias. Esta representación gráfica de los conglomerados la podemos obtener por medio de un dendograma.

Para solicitar el dendograma, se selecciona el botón **Plots** (figura 8.21), que abrirá una ventana de subdiálogo con diferentes características para declarar. El primer recuadro **Dendogram** realiza un gráfico de árbol que es la representación visual de los pasos de una solución de conglomeración jerárquica, para solicitarlo procederemos de la siguiente manera:**Plots- Dendograma- None- Continue**

Como sabemos, es necesario elegir el método con el que se desea hacer el conglomerado. La determinación del método de conglomerado, como sabemos, es una decisión de investigador. Pero recordemos que SPSS cuenta con los siguientes métodos:Between-groups linkage (Vinculación entre grupos), Within group linkage (Vinculación entre grupos), Nearest neighbor (Vinculación Simple), Furthest neighbor (Vinculación Completa), Centroid clustering (Agrupación de Centroides), Median clustering (Agrupación de Medias) y Ward's method (Método de Ward) [ver Capitulo 4].

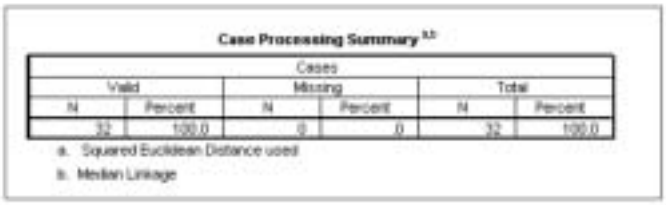
Como método de aglomeración seleccionaremos uno de los más sencillos, denominado el "vecino más cercano" (También conocido como Single Linkage-Vinculación Simple), que simplemente genera los distintos conglomerados uniendo sucesivamente los dos estímulos con mayor proximidad. Cuando se trata de unir un estímulo a un conglomerado, la proximidad entre ambos se calcula con el menor valor de proximidad entre ese estímulo y cualquiera de los estímulos ya incluídos en el conglomerado. El análisis nos proporciona una estructura de aglomeración que se resume en el dendograma.

El botón **Method** (figura 8.22) abre una ventana de subdiálogo que determina el método que se ha de emplear. El método jerárquico involucra, esencialmente, dos etapas principales: la primera es convertir los datos en una matriz de proximidad cuyos elementos son similaridades o disimilaridades entre las marcas que se están analizando, que después usa para combinar las marcas.

SPSS ofrece una variedad de medidas de proximidad bajo la parte de Measure dentro de la ventana de subdialogo Method. Las medidas son enlistadas de acuerdo al tipo de datos que pueden ser usados por ellos. Como todas las variables (tipos de delitos) que se usan en este análisis están medidas en una escala de intervalo [Capítulo 2], lo más lógico y apropiado es seleccionar la distancia euclidiana, aunque otras opciones están disponibles como menú desplegable. Puesto que la medida en que los delitos están medidos es por cada 100,000 habitantes no es necesario estandarizar las distancias.

El proceso de selección del Método de Análisis de Conglomerados Jerárquicos será el siguiente:**Method- Cluster Method: Nearest neighbor (Vincula-**

Figura 8.22: Especificación de la Medida de proximidad y el método de conglomeración.



Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
32	100.0	0	0	32	100.0

a. Squared Euclidean Distance used
b. Median Linkage

Figura 8.23: Resumen de Procesamiento de los Casos.

ción Simple)- Measure (Interval-Euclidean Distance)- Continue- Ok

Los resultados del análisis de conglomerado por el método de Vinculación simple **(Single Linkage)** que se muestran en la ventana de resultados de SPSS son los siguientes:

Primero se muestra una tabla que contiene el resumen del procesamiento de los casos **(Case Processing Summary)** los subíndices señalan la distancia empleada y el tipo de método de vinculación del conglomerado (figura 8.23). En este caso se uso la distancia Euclidiana **(Euclidean distance)** y el método de vinculación **(Single Linkage)**.

En la figura 8.24, la columna bajo el nombre Conglomerados que se combinan (Cluster Combined) parte del Historial de Conglomerados (Agglomeration Schedule) muestra cuáles estados son combinados en cada etapa del procedimiento de conglomerado. Primero el estado 5 (Chiapas) se une al estado 8 (Colima) puesto que en este por medio de esta estrategia, los individuos más alejados son lo que más pronto de unen. A este método se le conoce también como El Vecino más Alejado. La distancia se muestra bajo la columna de nombre Coeficientes (Coefficients). Después, el estado 20 (Oaxaca) se une al 30 (Veracruz) y así sucesivamente. SPSS usa el número del primer estado en un conglomerado para nombrar el conglomerado. Por ejemplo, en el paso cuarto el estado 13 (Hidalgo) es unido al conglomerado consistente del estado 5 y 8 (nombrado "conglomerado 5" en esta etapa). Cuando los conglomerados son unidos el valor del coeficiente depende del tipo de método de vinculación usado. Aquí con el método de Vinculación Completa, la distancia entre "conglomerado 5" y el estado 13 (Hidalgo) es de 50.981 delitos cometidos puesto ya que esta es la distancia más grande entre el estado 13 y cualquiera de los miembros del "conglomerado 5" (la distancia al estado 8).

En la figura 8.24, la columna bajo el nombre Conglomerados que se combinan **(Cluster Combined)** parte del Historial de Conglomerados **(Agglomeration**

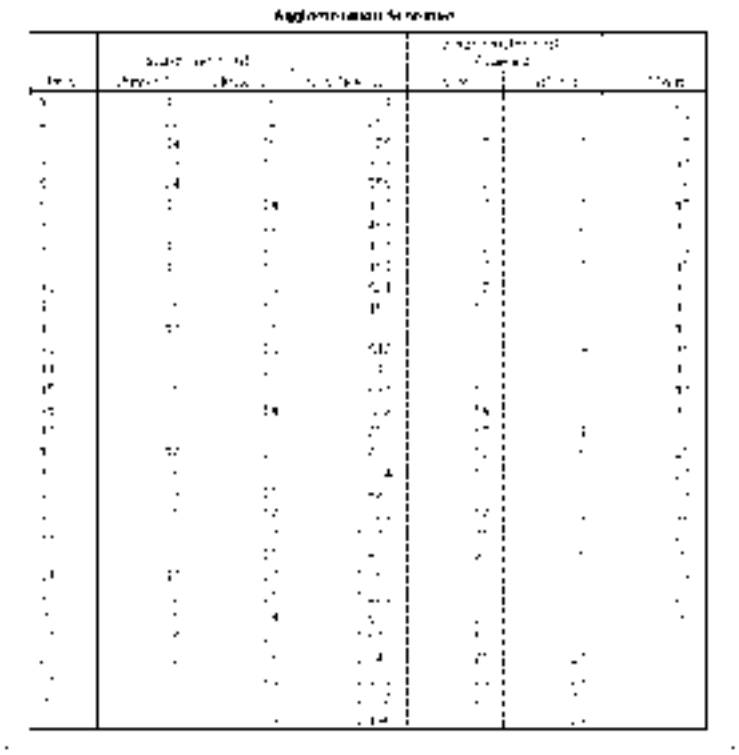


Figura 8.24: Historial de Conglomerados del Método de Vinculación Completa.

Schedule) muestra cuáles estados son combinados en cada etapa del procedimiento de conglomerado. Primero el estado 11 (Guanajuato) se une al estado 26 (Sonora) puesto que por medio de esta estrategia, los individuos más cercanos son lo que más pronto de unen. A este método se le conoce también como Vecino más Cercano. La distancia se muestra bajo la columna de nombre Coeficientes **(Coefficients)**. El siguiente el estado 22 (Querétaro) se une al 32 (Zacatecas) y así sucesivamente. SPSS usa el número del primer estado en un conglomerado para nombrar el conglomerado. Por ejemplo, en la etapa cinco el estado 31 (Yucatán) es unido al conglomerado consistente del estado 24 (San Luis Potosí) y 30 (Veracruz, nombrado "conglomerado 24" en esta etapa). Cuando los conglomerados son unidos el valor del coeficiente depende del tipo de método de vinculación usado. Aquí con el método del vecino más cercano, la distancia entre "conglomerado 24" y el estado 31 (Yucatan) es de 0.39 delitos cometidos puesto ya que ésta es la distancia más cercana entre el estado 31 y cualquiera de los miembros del "conglomerado 24" (la distancia al estado 30).

Las columnas bajo el título Etapa en la que el conglomerado aparece por primera vez **(Stage Cluster First Appers)** muestra las etapas en la cual un conglomerado o estado es unido por primera vez a su forma actual. Por ejemplo el "conglomerado 24" unido en la etapa 5 fue construido en la etapa 3. Finalmente la columna próxima etapa **(Next Stage)** muestra cuando un conglomerado construido en la etapa actual será incluido en otra combinación o etapa. Por ejemplo "conglomerado 24" como esta construido en la etapa 5 (24, 30 y 31) no será usado hasta la etapa 6 donde será unido con el estado 24 (San Luís Potosí).

Es más fácil de seguir este procedimiento de vinculación de grupos e individuos (estados) en un dendograma, el cual es un diagrama de árbol que despliega

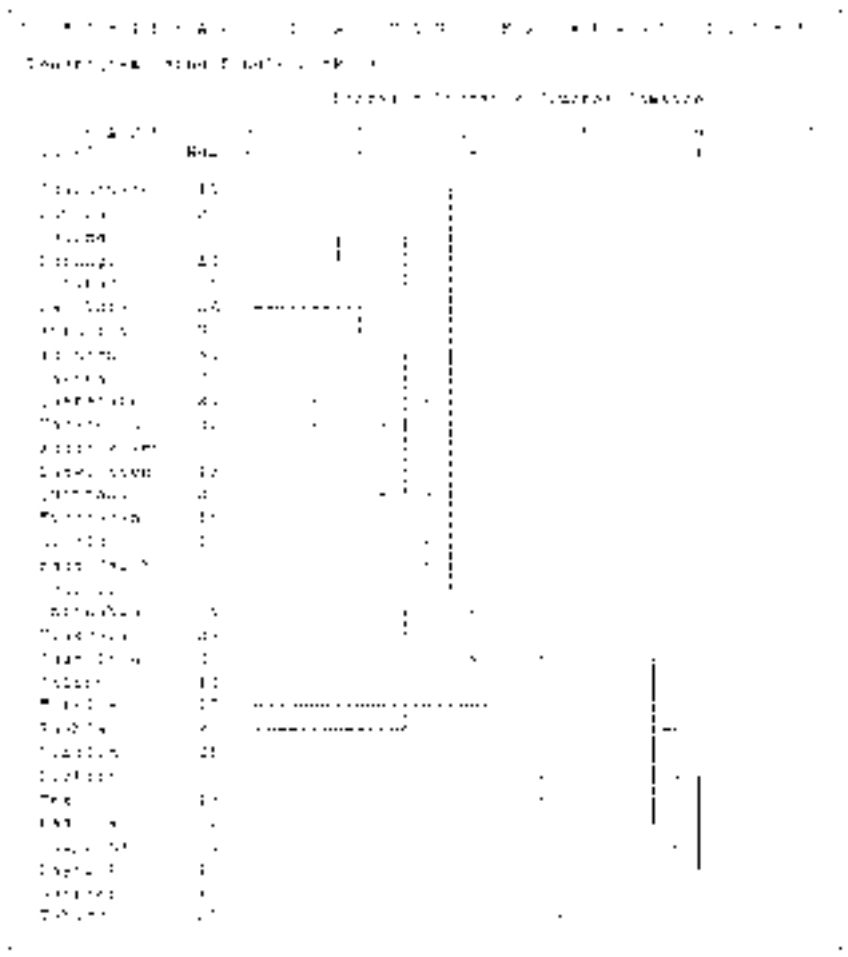


Figura 8.25: Dendograma usando el método de vinculación el Vecino más cercano

las series de fusiones del proceso de conglomeración de individuos a un solo grupo. El dendograma en este caso se muestra en la figura 8.25. SPSS reescala las distancia en un rango de 0 a 25.

Como se mencionó al inicio del análisis se correrán dos métodos de conglomerados, el primero Hierarchical Cluster (jerárquico) que ayuda a determinar un número viable de grupos y el segundo es el K-Means Cluster (k- medias) donde ya se hace la separación en sí. El primer método fue con una muestra y con eso fue suficiente, pero en con el segundo se tiene que hacer con todos los datos por lo que se tiene que realizar los siguientes pasos para tener en cuenta toda la muestra nuevamente:**Data- Select Cases- All Cases- Ok**

Una vez que están todos los casos se puede pasar a la ejecución del procedimiento estadístico. Para esto se tiene que elegir **Analyze** de la barra de menús, se selecciona la categoría **Classify- K-Means Cluster** (figura 8.26). Esto abrirá una ventana de dialogo, donde es posible definir las variables que se quieren usar en el análisis de Conglomerados. Las variables se especifican bajo la lista **Variables(s)** con la ayuda de los botones "mover variables de un lugar a otro". Se seleccionaran los atributos que se desean observar respecto a los consumidores. El número de conglomerados tiene que ser definido por el analista, en este caso seleccionaremos 7.

El método que se seleccionará es **Iterate** y **classify**. La opción bajo el nom-

Figura 8.26: Declaración de un Análisis de Conglomerados de K-Medias.

Figura 8.27: Guardar nueva variable del conglomerado de pertenencia de cada caso.

bre de **Save** permite generar una nueva variable con el número de conglomerado correspondiente por caso, es decir guarda la selección del conglomerado de pertenencia. El procedimiento completo es de la siguiente manera:**Save- Cluster membership- Continue** figura 8.27.

Para continuar con el análisis se tiene que ir al botón **Options** donde se establecen estadísticos como son: centros de conglomerados iniciales, tabla de ANOVA e información de los conglomerados para cada caso. Para este ejemplo solicitaremos únicamente los centros de conglomerados iniciales, figura 8.28.**Inicial cluster centres- Continue- Ok**

La salida del conglomerado resultante es mostrada a continuación. La tabla de centros de conglomerados iniciales (**Inicial cluster centers**) muestra los valores iniciales usados por el algoritmo. La tabla historial de iteraciones **(Iteration History)** indica que el algoritmo ha convergido y los centros de los conglomerados **(Final Cluster Centers)** y el número de casos en cada conglomerado **(Number of cases in each Cluster)** describen la solución final de conglomerados.

En el método de conglomeración de k-medias los centros de conglomerados

Figura 8.28: Estadísticas centros de conglomerados iniciales

Figura 8.29: Opción Sumaries Cases

son determinados por la medias de las variables de los estados pertenecientes a los conglomerados (también conocidos como centroides).La conformación de la conglomeración por estados podemos obtenerla en SPSS por medio de las siguientes opciones:**Analize-Reports- Case Summaries**

Las variables que se seleccionaran son únicamente el número de caso del conglomerado y la entidad federativa, porque lo que nos interesa es observar el comportamiento por estado, figura 8.29.

El resultado se muestra en la figura 8.30.

Con esta información y con el dendograma se puede tener una idea más concisa de que estados se parecen entre sí y cómo pueden ser agrupados. Recordemos que esta información esta basada en el número de delitos cometidos denunciados ante el ministerio público y que a partir de estos datos se obtuvieron coeficientes de disimilaridad entre estados. Mediante la obtención de la matriz de correlación entre los estados, y transformar luego éstas en disimilaridades.

8.3.3. Conclusiones

Del dendograma y del trazado de una línea vertical sobre éste se definió el número de grupos de estados. De aquí se desprende que los nueve grupos que se forman por ser similares son:

- Grupo 1: Aguascalientes, Baja California Sur, Chiapas, Coahuila, Colima, Hidalgo, Michoacán, Nuevo León, Queretaro, Quintana Roo y Zacatecas.
- Grupo 2: Baja California Norte y Sinaloa
- Grupo 3: Distrito Federal y Estado de México.
- Grupo 4: Nayarit
- Grupo 5: Durango, Puebla, Guanajuato, Oaxaca, Morelos, San Luis Potosí, Sonora, Veracruz y Yucatán.
- Grupo 6: Chihuahua, Jalisco, Tamaulipas y Tlaxcala.
- Grupo 7: Campeche, Guerrero y Tabasco.

La estructura de este agrupamiento se debe, sin duda, al grado de seguridad de cada estado. Es decir al número de delitos cometidos denunciados. Por

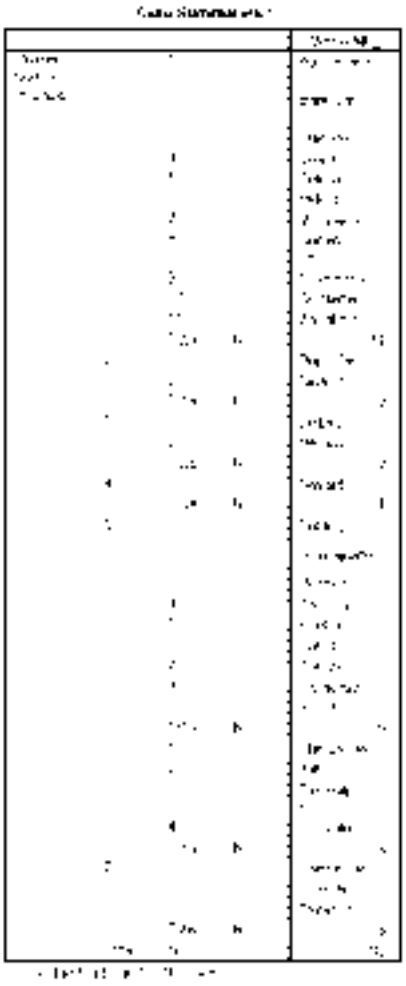


Figura 8.30: Opción Sumaries Cases

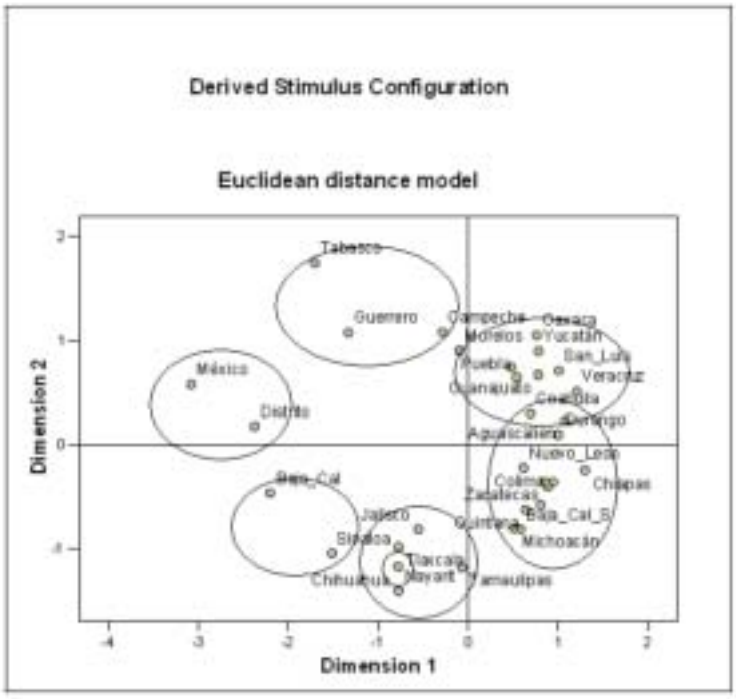


Figura 8.31: Espacio de estímulos para las 32 estidades federativas, junto con algunas estructuras de agrupamiento del análisis de conglomerados

ejemplo, de acuerdo al índice de delitos el grupo 2, que está formado por el estado de Baja California Norte representa el estado con mayor número de delitos cometidos por lo que es el estado con mayor inseguridad. Contrariamente a lo que se tiene pensado, de que la Ciudad de México es el estado de la republica mexicana con más inseguridad. El grupo 7, formado por Nayarit es uno de los estados con menor cantidad de delitos, así como el estado de Campeche que se encuentra aislado de la representación en dos dimensiones de las coordenadas de los treinta y dos entidades federativas que es el estado que tiene menos numero de delitos registrados.

La información complementaria proporcionada por el análisis de conglomerados puede incorporarse a la configuración de estímulos en dos dimensiones, para apreciar más claramente la estructura de agrupamiento. Esto es lo que se representa en la figura 8.31, que combina el espacio de estímulos y común con la solución proporcionada por el análisis de conglomerados.

8.4. Segundo Ejemplo

Un segundo ejemplo del uso de las técnicas multivariadas de análisis de conglomerados y escalas multidimensionales, sobre la misma base de información, sería efectuar el análisis sobre los tipos de delitos denunciados ante ministerio público. La idea principal de este ejemplo será reducir el número de variables (tipos de delitos) para que un análisis posterior se efectué con las variables más significativas y que el resultado final sea el más adecuado. A lo largo de este segundo ejemplo sólo se presentarán algunos comandos así como los resultados del análisis de conglomerados y escalamiento multidimensional. En este ejemplo sólo se mencionaran los comandos que se emplearan durante el análisis sin presentar las ventanas.

8.4.1. Análisis de Escalas Multidimensionales de tipos de delitos

Al igual que el primer ejercicio, y dado que los datos originales no son proximidades, se tendrá que obtener a partir de los datos un coeficiente de disimilaridades entre tipos de delitos. Primero se tiene que obtener una matriz de correlaciones entre los tipos de delitos. A diferencia del primer ejemplo no se transponen los datos, para obtener la matriz de correlaciones es necesaria la sintaxis siguiente:

```
CORRELATIONS
/VARIABLES =Violacion Despojo Harmbl Harmfue Larmbl Larmfue Amenazas Estupro Odelsex Abc
Extorsion Fraude Secuestro Robaut Robban Robcamc Robcasb Robcascam Robcashab Robneg Robt
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE
/MATRIX=OUT (Delitos_1.sav).
```

El archivo "$\backslash Delitos_1.sav$" contendrá varias filas con estadísticos descriptivos, seguidas de la matriz de correlaciones. Se eliminarán estas filas innecesarias del archivo, para poder seguir con el análisis. Puesto que es muy difícil interpretar directamente los resultados a partir de la matriz de correlaciones es necesario llevar a cabo una última transformación. La sintaxis correspondiente a la transformación deseada será la siguiente:

```
COMPUTE Violacion =SQRT(2*(1-Violacion)).
COMPUTE Despojo =SQRT(2*(1-Despojo)).
COMPUTE Harmbl =SQRT(2*(1-Harmbl)).
COMPUTE Harmfue =SQRT(2*(1-Harmfue)).
COMPUTE Larmbl =SQRT(2*(1-Larmbl)).
COMPUTE Larmfue =SQRT(2*(1-Larmfue)).
COMPUTE Amenazas =SQRT(2*(1-Amenazas)).
COMPUTE Estupro =SQRT(2*(1-Estupro)).
COMPUTE Odelsex =SQRT(2*(1-Odelsex)).
COMPUTE Abconf =SQRT(2*(1-Abconf)).
COMPUTE Daproaj =SQRT(2*(1-Daproaj)).
COMPUTE Extorsion =SQRT(2*(1-Extorsion)).
COMPUTE Secuestro =SQRT(2*(1-Secuestro)).
COMPUTE Robaut =SQRT(2*(1-Robaut)).
COMPUTE Robban =SQRT(2*(1-Robban)).
COMPUTE Robcamc =SQRT(2*(1-Robcamc)).
COMPUTE Robcasb =SQRT(2*(1-Robcasb)).
COMPUTE Robcashab =SQRT(2*(1-Robcashab)).
COMPUTE Robneg =SQRT(2*(1-Robneg)).
COMPUTE Robtran =SQRT(2*(1-Robtran)).
COMPUTE Robtrans =SQRT(2*(1-Robtrans)).
COMPUTE Robveh =SQRT(2*(1-Robveh)).
COMPUTE Robgan =SQRT(2*(1-Robgan)).
COMPUTE rowtype_="PROX".
EXECUTE.
```

Los primeros 25 comandos efectuarán la transformación en los valores de la correlación para cada tipo de delito. Por su parte, el penúltimo comando sirve para indicar a SPSS, a través de la variable rowtype, que los datos son ahora proximidades ("PROX") y no correlaciones ("CORR"). Ahora es posible ejecutar el procedimiento de escalamiento en SPSS.

Para llevar a cabo el análisis se usara el algoritmo PROXSCAL para obtener una solución apropiada. De la barra de menú principal se selecciona la opción **Analyze** en donde se busca el procedimiento estadístico **Scale- Multidimensional Scaling (Proxscal)**.

Este procedimiento abrirá una ventana de diálogo en donde es posible definir el formato de los datos y el número de fuentes que se quieren usar en el análisis de escalas multidimensionales. Como los datos que se están empleando son proximidades, el formato de los datos son proximidades. El procedimiento a seguir será el siguiente: primero **Format Data- The data are proximidades**, segundo **Number of Sources- One matriz source** y por último **One source-The proximities are in a matriz across columns**. En esta ventana de dialogo seleccionamos el botón: **Define** para seleccionar las proximidades que se quiere usar en el análisis. Las variables se especifican bajo la lista **Proximitie(s)** con la ayuda de los botones "mover variables de un lugar a otro". Se seleccionaran las variables estado que se desean analizar respecto el tipo de delitos denunciados, esto porque nos interesa ver si existe algún tipo de agrupamiento natural. Se pulsa a continuación el botón **Model**. Esta ventana permite especificar el nivel de medida de los datos, tipo de proximidades, forma de la matriz y el número de dimensiones que queremos que tenga la solución. En este paso sólo efectuaremos un cambio. En la casilla etiquetada **Level of Measurement** especificaremos **Interval**. A continuación pulsaremos el botón **Continue**. Una vez de nuevo en el cuadro de diálogo anterior, pulsaremos el botón **Restrictions-No restrictions-Continue**. Después de ésto se selecciona la opción **Options** donde se tiene que seleccionar lo siguiente: **Options- Inicial Configuration (Simples)-Continue**. Lo que sigue es la selección del diagrama y las opciones a seguir son las siguientes: **Plots- Commun Space-Continue**. Finalmente se regresará a la ventana principal, para esto se pulsara el botón **Ok**, para solicitar que se ejecute el análisis bajo los criterios anteriores. El siguiente paso es interpretar la solución proporcionada por el programa.

En el editor de resultados de SPSS, se encontrará un resumen del procedimiento. En este resumen aparece, en primer lugar, un listado de los casos procesados, mostrando todos los pares posibles de distancia entre los delitos estados (figura 8.32. En la configuración formada se puede apreciar que aquellos delitos entre los que existen altas correlaciones se encuentran próximos entre sí, como es el caso de: Otros delitos sexuales y estupro. Otro grupo: robo de camión de carga, robo de auto, robo a casa habitación y despojo. Estos resultados servirán para analizar los dos tipos de análisis multivariado.

8.4.2. Análisis de Conglomerados de tipos de delitos

A continuación se mostrará el procedimiento del análisis de conglomerados para los tipos de delito. Se usaran dos métodos de Análisis de Conglomerados, para obtener una solución apropiada. El primero será el **Hierarchical Cluster** (Conglomerados jerárquicos) que ayudará a determinar el número de conglomerados, porque permite visualizar las posibles estructuras por medio del dendograma de los conglomerados. El segundo será es el **K-Means Cluster** (Conglomerado de k-medias) donde se especificarán el número de conglomerados en que se desea dividir la muestra. Una vez que se tiene la base en orden los datos se puede pasar a la ejecución del procedimiento estadístico. Para esto se tiene que elegir Analyze de la barra de menúes, se selecciona la categoría **Classify- Hierarchical Cluster**.

Esto abrirá una ventana de diálogo, donde es posible definir las variables que se quieren usar en el análisis de Conglomerados. Las variables se especifican bajo la lista **Variables(s)** con la ayuda de los botones "mover variables de un

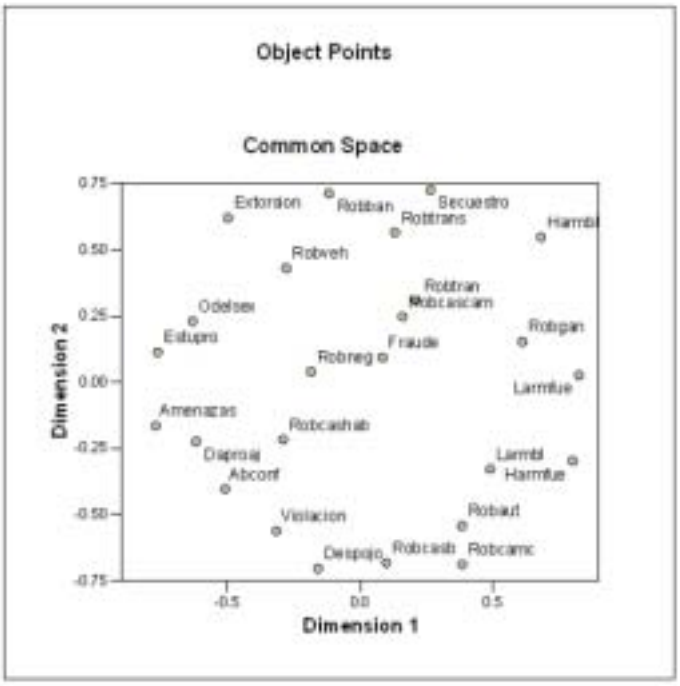


Figura 8.32: Representación en dos dimensiones de las coordenadas de los 25 tipos de delito

lugar a otro". Se seleccionaran las variables de tipos delitos denunciados que se desean analizar respecto a las entidades federativas. En este ejemplo el análisis se enfocara al conglomerado por casos (entidades federativas) en lugar de a las variables. El procedimiento será el siguiente: se selecciona Cases bajo la opción **Cluster** y se marca el cuadrado de **Statistics** bajo la opción **Display**. Cuando se pulsa el botón **Statistics** que abrirá otra ventana de subdialogo que permitirá seleccionar tablas y diagramas.

La opción **Agglomeration Schedule** muestra los casos o conglomerados combinados en cada etapa, las distancias entre los casos o los conglomerados que se van combinando, así como el último nivel del proceso de aglomeración en el que cada caso se une al conglomerado correspondiente. La opción **Proximity matrix** proporciona la matriz de distancias entre los elementos. El **Cluster Membership** muestra el conglomerado al cual se asigna cada caso en una o varias etapas de la combinación de los conglomerados.

En este caso únicamente seleccionaremos, dentro de la ventana de subdiálogo Statistics:**Agglomeration Schedule- None- Continue**. Recordemos que lo único que nos interesa en este ejemplo es tener una idea estructural de los posibles conglomerados en que dividiremos a la población, cuando se generen los conglomerados por medio de k-medias. Esta representación gráfica de los conglomerados se puede obtener por medio de un dendograma.

Para solicitar el dendograma, se selecciona el botón **Plots**, que abrirá una ventana de subdiálogo con diferentes características para declarar. El primer recuadro **Dendogram** realiza un gráfico de árbol que es la representación visual de los pasos de una solución de conglomeración jerárquica, para solicitarlo procederemos de la siguiente manera: **Plots- Dendograma- None- Continue**.

Como método de aglomeración seleccionaremos uno de los más sencillos, denominado el "vecino más cercano", que simplemente genera los distintos conglomerados uniendo sucesivamente los dos estímulos con mayor proximidad.

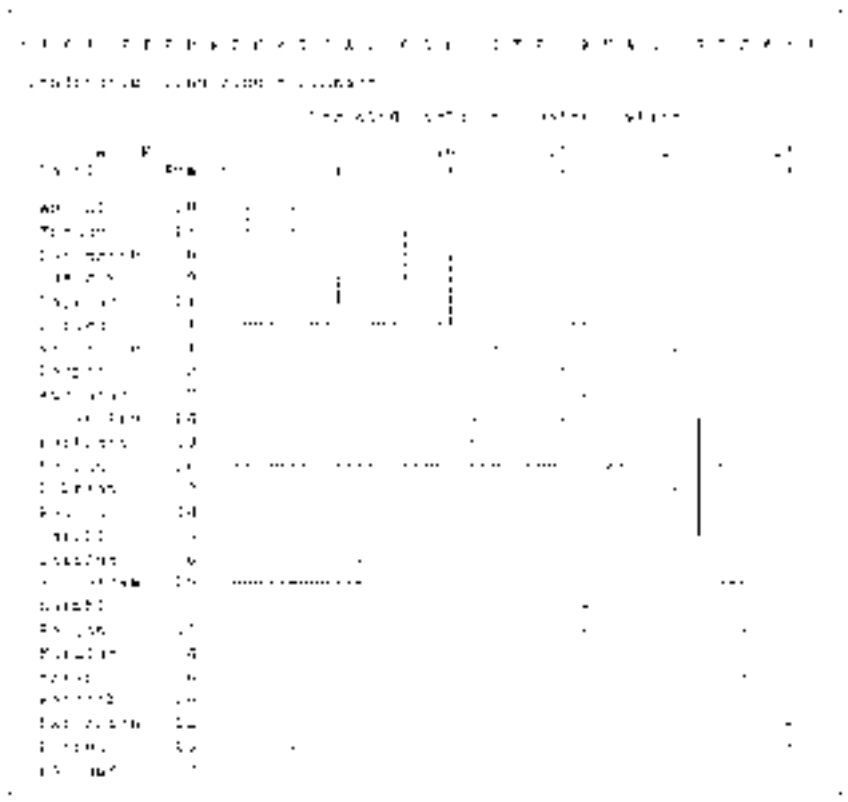


Figura 8.33: Dendograma

Cuando se trata de unir un estímulo a un conglomerado, la proximidad entre ambos se calcula con el menor valor de proximidad entre ese estímulo y cualquiera de los estímulos ya incluídos en el conglomerado. El análisis proporciona una estructura de aglomeración que se resume en el dendograma.

El botón **Method** abre una ventana de subdiálogo que determinará el método que se ha de emplear. El método jerárquico involucra convertir los datos en una matriz de proximidad cuyos elementos son similaridades o disimilaridades entre las marcas que se están analizando, para después usarlos para combinar las marcas.

SPSS ofrece una variedad de medidas de proximidad bajo la parte de **Measure** dentro de la ventana de subdialogo **Method**. Puesto que la medida en que los delitos están medidos es por cada 100,000 habitantes no es necesario estandarizar las distancias.

El proceso de selección del Método de Análisis de Conglomerados Jerárquicos será el siguiente: **Method- Cluster Method: Nearest neighbor (Vinculación Simple)- Measure (Interval-Euclidean Distance)- Continue- Ok**

Los resultados del análisis de conglomerado por el método de Vinculación simple (Single Linkage) que se muestran en la ventana de resultados de SPSS se puede ver en la figura 8.33.

El dendograma sera útil para decidir el número de conglomerados en un conjunto de datos. Por medio del trazado de una línea vertical a lo largo de los casos y contando el número de intersecciones, de esos números se eligen lo que parezcan más viables para el caso. Una vez que se determina el número de conglomerados que se desean, se corre el método de k-medias. De acuerdo a la línea que se trazó sobre el dendograma parece ser conveniente formar 7 grupos

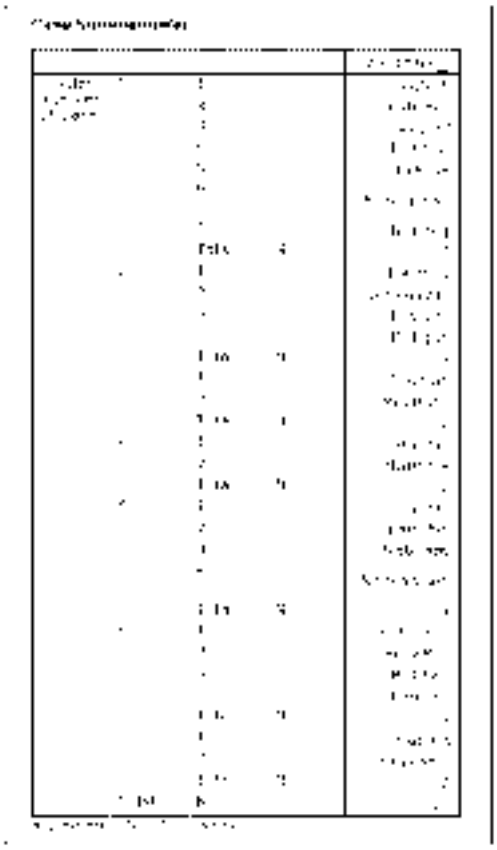


Figura 8.34: Resumen de casos

de estados.

Con la información del resumen de casos (figura 8.34) y con el dendograma se puede tener una idea más objetiva de cuáles tipos de delitos se parecen entre sí y como pueden ser agrupados. Recordemos que esta información esta basada en el número de delitos cometidos denunciados ante el ministerio público y que a partir de estos datos se obtuvieron coeficientes de disimilaridad entre estados, mediante la obtención de la matriz de correlación entre los estados y la transformación de éstas en disimilaridades.

8.4.3. Conclusiones

Del dendograma y del trazado de una línea vertical sobre éste se definió el número de grupos de estados. De aquí se desprende que son nueve grupos que se forman por ser similares entre sí, la conformación de éstos son:

Grupo 1: Violación, Otros delitos sexuales, abuso de confianza, daño a propiedad ajena, fraude, robo a casa habitación y robo a negocio.

Grupo 2: Despojo, amenaza, estupro y robo de ganado.

Grupo 3: Robo a transeúnte y robo a transportistas.

Grupo 4: Homicidio con arma blanca y homicidio con arma de fuego.

Grupo 5: Lesiones con arma blanca, Lesiones con arma de fuego, Robo de Casa de Bolsa y Robo de Casa de Cambio.

Grupo 6: Extorsión, secuestro, robo a banco y robo a vehículo

Grupo 7: Robo en autobús y robo de camión de carga.

La estructura de este agrupamiento se debe, sin duda, al grado de correlación entre las variables (tipo de delitos). Con estos grupos podemos generar nuevas variables y reducir las variables para tener un mejor análisis. Es importante señalar que este tipo de variables pueden ser agrupadas sin ninguna dificultad por un estudioso de ciencias penales. Pero si no se tiene conocimiento de esta área, estas dos técnicas de análisis multivariado son muy buenas opciones para encontrar similitudes entre los datos y generar grupos que sean similares entre sí, de tal manera que el analista no use su juicio para este propósito y solo enfoque en ejecutar los procesos. De esta manera se evita que juicios de valor por parte del analista interfieran con los resultados.

Capítulo 9

Conclusiones

El objetivo de esta tesis fue presentar dos métodos de análisis multivariado: escalamiento multidimensional y análisis de conglomerados, con la idea de proporcionar herramientas de análisis estadístico que enriquezcan el conocimiento que se tiene sobre inseguridad. La inseguridad que se encuentra directamente relacionada a los delitos que se cometen en una entidad no puede ser medida por medio de cifras brutas sino que se tiene que ir más allá. Estas dos herramientas del análisis multivariado representan una alternativa de análisis estadístico que se encuentra fuertemente fundamentada en estructuras de proximidad antes de que una medida de proximidad sea escogida.

El primer análisis es muy útil, sobre todo a nivel exploratorio. Esto se debe a que el concepto clave son los datos de entrada, a las que se le denomina proximidades. Las proximidades son valores que indican la cercanía, objetiva o subjetiva, entre dos o más objetos. Sin embargo, la medida de proximidad va más allá de la mera cercanía física en el espacio. También pueden utilizarse otras medidas de proximidad objetivas o la proximidad subjetiva entre estímulos. Las proximidades pueden ser de distinto tipo, tales como medidas de similaridad o disimilaridad, correlaciones y muchas otras. Pero también, es posible obtener proximidades a partir de datos que nos son proximidades calculando una medida de similaridad o disimilaridad entre las filas o las columnas de una matriz. En este caso, hablamos de coeficientes de similaridad o disimilaridad. La representación espacial de las proximidades se hace de tal modo que si dos estímulos son valorados como muy parecidos (o como poco diferentes) se encontrarán a poca distancia uno de otro, y viceversa. La representación de los estímulos es un espacio de pocas dimensiones facilita la interpretación de las proximidades al mostrarlas en forma visual en lugar de numérica, capturando lo esencial de la información original y reduciendo el error o "ruido" existente en los datos. Se hace posible, de este modo, observar la "estructura oculta" en los datos, bien sea en forma de agrupaciones significativas de estímulos, bien en forma de dimensiones a lo largo de las cuales interpretar las diferencias entre los estímulos, o de alguna otra de las formas posibles de interpretar una solución de MDS.

Es importante notar que un procedimiento de gran utilidad para interpretar las soluciones MDS es tratar de buscar agrupamientos de estímulos. Estos agrupamientos indicarían conjuntos de estímulos muy semejantes entre sí y diferentes a los demás, que pueden ser de gran utilidad sí la finalidad principal del análisis es la clasificación. El análisis de conglomerados genera agrupamientos jerárquicos de los estímulos en función de su proximidad. Esta agrupación se hace de tal modo que aquellos estímulos más similares entre sí formarán parte de un mismo conglomerado. A medida que la proximidad vaya disminuyendo, otros estímulos u otros conglomerados se irán uniendo a esta estructura jerárquica hasta

que, finalmente, todos los estímulos pertenezcan a un único conglomerado. Si se complementa la información proporcionada por MDS con la información sobre agrupamientos proporcionada por el análisis de conglomerados, será más sencillo identificar grupos de estímulos con características semejantes, así como el número de grupos que existen. Más aún la información complementaria proporcionada por el análisis de conglomerados puede añadirse a la configuración de estímulos en dos dimensiones, para que se pueda apreciar de forma más clara la estructura de agrupamiento.

Un trabajo posterior podría ser generar un análisis de factores primero y posteriormente realizar los dos análisis multivariados mencionados en este trabajo, para obtener un mejor resultado en la clasificación de grupos entre entidades federativas de la república mexicana. Lo anterior, con base a obtener un grupo menor de tipos de delitos que sean significativos al análisis y no cargar con el número tan grande de variables que comparados con los casos es mucho mayor. Siempre es preferible tratar de reducir el número de variables antes de hacer un análisis. Pero, puesto que la idea original de este trabajo fue la presentación teórica y practica de únicamente estos dos análisis, no se realizó el análisis de factores. Los resultados de este análisis permitirán conocer en que variables se puede fijar el analista para determinar que estados son parecidos entre sí en cuanto a los delitos que se comenten en su entidad, bien podríamos decir el nivel de inseguridad "real" de la República Mexicana. El manejo de herramientas estadísticas más avanzadas permitirá al investigador tener una idea objetiva del comportamiento de delitos en una entidad y no crea ideas subjetivas del problema.

Apéndice A

Cómo obtener proximidades a partir de una matriz rectangular

Las matrices rectangulares en dos vías y dos modos, que contienen datos multivariados o "de perfil" del tipo sujetos x atributos ó estímulos x atributos, son fuentes de datos habituales en investigación. Es el tipo de matrices que uno utiliza como datos de entrada en la mayoría de los análisis estadísticos, como análisis factorial o análisis de regresión múltiple. Estas matrices de datos suelen obtenerse midiendo a cada sujeto o a cada estímulo en una serie de atributos. También es posible obtener matrices rectangulares en tres vías y tres modos, del tipo estímulos x atributos x sujetos. Estas matrices suelen obtenerse pidiendo a una muestra de sujetos que evalúen una serie de estímulos en un mismo conjunto de atributos.

Este tipo de datos no son apropiados para ser utilizados directamente en MDS, pero es posible transformarlos en proximidades entre las entidades que formen las filas (o las columnas) de la matriz rectangular. En el caso de las matrices en tres vías y tres modos, los datos pueden transformarse en proximidades entre las entidades que formen las filas (por ejemplo, estímulos) o las columnas (por ejemplo, atributos) de la matriz. De este modo, obtendríamos una matriz. De este modo, obtendríamos una matriz de proximidades para cada una de las entidades que conforman la tercera vía de la matriz (por ejemplo, sujetos).

Para generar una matriz de proximidades a partir de una matriz rectangular, es necesario tener en cuenta que las proximidades se generarán de distinta forma para datos medidos en escala nominal, ordinal o de intervalo.

A.1. Datos en escala de intervalo

Una forma rápida y sencilla de obtener datos de proximidad a partir de datos de perfil es a partir de la matriz de correlaciones entre estímulos (columnas) o entre individuos (filas) de la matriz rectangular. La matriz de correlaciones es sólo un caso especial de la familia de matrices de productos escalares, que a su vez están muy relacionadas con las matrices de distancias, de modo que es posible transformar las correlaciones en distancias de forma sencilla, mediante la transformación (Cox,1982).

$$d_{ij} = \sqrt{2(1 - r_{ij})}$$

	Atributos			
	A	B	C	D
Estímulo 1	2	4	5	8
Estímulo 2	3	6	7	7

Cuadro A.1: Evaluación de los dos estimulos en función de cuatro atributos

	Atributos			
	A	B	C	D
Estímulo 1	2	4	5	8
Estímulo 2	3	6	7	7
Diferencia	-1	-2	-2	1

Cuadro A.2: Diferencias entre puntuaciones de los estímulos 1 y 2

Otra forma muy utilizada de obtener proximidades a partir de datos de perfil es calcular la distancia euclidiana entre los estímulos.

$$\delta_{ij} = \sqrt{\sum_{a=1}^{m}(x_{ia} - x_{ja})^2}$$

Donde i y j son los estímulos y x_{ia} y x_{ja} son las puntaciones de ambos estímulos en el atributo a. Esta es la distancia que SPSS calcula por defecto cuando crea proximidades a partir de datos de perfil. El cálculo de la distancia euclidiana entre dos estímulos es sencillo. Un ejemplo hipotético seria: que un sujeto tiene que evaluar dos estímulos en función de cuatro atributos (llamados A, B, C y D) en una escala del 1 al 10. Supongamos que las respuestas del sujeto son las que se muestran en el cuadro A.1

Para calcular la distancia euclidiana entre los estímulos 1 y el estímulo 2 se tiene que hallar la diferencia entre sus puntuaciones en los cuatro atributos en que fueron evaluados se muestran en el cuadro A.2.

Para evitar el problema del signo de las diferencias, elevaremos éstas al cuadrado y las sumaremos, para obtener una medida global de las diferencias entre los dos estímulos en los cuatro atributos considerados, cuadro A.3.

La suma de estas diferencias nos da e valor 10, que es el valor de la distancia al cuadrado entre los estímulos 1 y 2. La raíz cuadrada de este valor será, por tanto, la distancia entre ambos estímulos. Este valor es de (10)1/2= 3.1622.

Hay varios aspectos que es necesario tener en cuenta en cuanto a la distancia euclidiana. El primero de ellos es que, como puede apreciarse, la distancia se calcula como la raíz cuadrada de una suma de diferencias al cuadrado. Si los distintos atributos están medidos en escalas diferentes, esto provocará que algunas diferencias cuadráticas sean especialmente grandes, lo que dará mayor peso a algunos de los atributos en la estimación de la distancias euclidiana. Para evitar este problema, es recomendable estandarizar las escalas utilizadas para

	Atributos			
	A	B	C	D
Estímulo 1	2	4	5	8
Estímulo 2	3	6	7	7
Diferencia	-1	-2	-2	1
Diferencia al cuadrado	1	4	4	1

Cuadro A.3: Diferencias al cuadrado entre puntuaciones de los estímulos 1 y 2

todos los atributos antes de calcular las distancias. El segundo aspecto se refiere a la existencia de correlaciones importantes entre algunos de los atributos, lo que podría indicar que todos ellos son, en realidad, distintas medidas de un mismo atributo subyacente. Al calcular la distancia euclidiana entre estímulos a partir de las puntuaciones en los atributos, el atributo subyacente mencionado tendrá una influencia superior a los demás, puesto que contribuirá al valor de la distancia varias veces, mientras que los demás atributos lo harán sólo una vez.

A.2. Datos en escala ordinal

Aunque existen varias medidas de asociación para datos en escala ordinal (ver capítulo 3), quizá la más conocida y utilizada de ellas sea el coeficiente de correlación de rangos *rho* de Spearman (ρ), que se calcula a partir de las diferencias cuadráticas entre los rangos obtenidos por los estímulos en los distintos atributos:

$$\rho = \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Una vez obtenidos los valores de r pueden ser transformados fácilmente en proximidades utilizando la fórmula de transformación de correlaciones en distancias ya mostrada, o alguna otra transformación similar.

A.3. Datos en escala nominal

Como es sabido, una escala nominal se encuentra dividida en un número n de categorías mutuamente excluyentes y exhaustivas, de tal modo de cada caso es asignado a una y sólo una de esas categorías. El grado de asociación entre dos estímulos se medirá luego mediante una tabla de contigecia $2xn$ donde se comparará el número de casos asignados a una categoría determinada (frecuencias obsevadas) con el número de ellos que sería de esperar si las categorías fuesen indepndientes entre sí para cada uno de los estímulos (frecuencias esperadas). Un caso especial de los datos en escla nominal lo constituyen las variables dicotómicas o binarias, donde únicamente se expresa si un individuo posee o no posee un determinado atributo. La tabla de contingencia resultante es una tabla $2x2$, donde cada casilla indica el cruce de casos presente y ausentes para dos estímulos dados. Así pues, existen dos formas diferentes de utilizar datos en escala nominal para transformarlos en proximidades:

1. Datos de recuento, donde la matriz rectangular de entrada contiene el número de veces en cada variable (fila) ha sido asignada a una categoría (columna) determinada.

2. Datos binarios, donde la matriz rectangular contiene únicamente dos valores (generalmente 1 y 0), que indican si una determinada variable (fila) posee o no un determinado atributo (columna).

A.3.1. Datos de recuento

Cuando la matriz de entrada es una matriz de frecuencias para distintos estímulos en una serie de categorías, el grado de asociación entre estímulos (o entre categorías) se suele calcular mediante medidas basadas en chi-cuadrado (χ^2). Las dos más empleadas son la distancia chi-cuadrado y la distancia phi-cuadrado.

La distancia chi-cuadrado entre dos estímulos X e Y se calcula como la suma de las diferencias entre frecuencias observadas y esperadas para cada estímulo.

		Estímulo Y		
		Presente	Ausente	
Estímulo X	Presente	a	b	(a+b)
	Ausente	c	d	(c+d)
		(a+c)	(b+d)	N=(a+b+c+d)

Cuadro A.4: Tabla de contingencia de casos presentes y ausentes

Cada diferencia, a su vez, se calcula como la raíz cuadrada de la suma, para las i categorías de la variable, de las diferencias cuadráticas entre las frecuencias observadas (X_i, Y_i) y las esperadas $(E(X_i), E(Y_i))$, divididas por sus frecuencias esperadas. Formalmente

$$d(X,Y) = \sqrt{\sum_i \frac{(X_i - E(X_i))^2}{E(X_i)} + \sum_i \frac{(Y_i - E(Y_i))^2}{E(Y_i)}}$$

Por su parte, phi-cuadrado es igual a chi-cuadrado dividido por el número de casos. Formalmente

$$d(X,Y) = \frac{\chi^2}{N} = \sqrt{\frac{\sum_i \frac{(X_i - E(X_i))^2}{E(X_i)} + \sum_i \frac{(Y_i - E(Y_i))^2}{E(Y_i)}}{N}}$$

A.3.2. Datos binarios

Cuando la matriz de entrada contiene únicamente datos binarios que indican la presencia o ausencia de una determinada característica en el estímulo, la proximidad entre estímulos se determina a partir de una tabla de contingencia $2x2$ con cuatro casillas (a,b,c,d) que indican el cruce de casos presentes y ausentes, tabla A.4.

Existe un número impresionante de medidas de asociación y proximidad para datos dicotómicos. Una de las medidas de asociación más conocida es el coeficiente de correlación ϕ de Person:

$$\phi = \frac{(ad - bc)}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$$

Al igual que son los coeficientes de correlación anteriores, podemos transformar luego en distancias el valor de las correlaciones entre estímulos utilizando alguna de las fórmulas existentes al afecto.

Bibliografía

[1] Abdi, H. (2007). The Binomial Distribution: Binomial and Sign Tests. In N.J. Salkind (Ed.): *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage. [Online]. Disponible en: `http://www.utdallas.edu/~herve/Abdi-Binomial2007-pretty.pdf` [Cited January 17,2007]

[2] Arango, A.(2004). *Sistema de Información Delictiva*. México: Instituto Nacional de Ciencias Penales.

[3] Bolch, B. W. & Huang, C. J. (1974). *Multivariate Statistical Methods for Business and Economics*. New Jersey: Prentice-Hall.

[4] Brace, C. L., Nelson, A. R., Seguchi, N., Oe,H., Sering, L., Qifeng, P., Yongyi, L. & Tumen, D. (2001). Old World sources of the first New World human inhabitants: A comparative craniofacial view. *National Academy of Science*. [Online]. 31 Jul. US Disponible en: `http://www.pnas.org/cgi/content/full/171305898v1`

[5] Brillinger, D. (2000). Time Series: a stretch of values on the same scale indexed by a time-like parameter. *Int. Encyc. Social and Behavioral Science*. [Online]. California, Berkeley. Disponible en: `http://www.stat.berkeley.edu/~brill/Papers/encysbs.pdf`

[6] Burke, A., Abeles, E. & Chen, B. (2004). The Response of the Auto Industry and Consumers to Changes in the Exhaust Emission and Fuel Economy Standards (1975-2003): A Historical Review of Changes in Technology, Prices and Sales of Various Classes of Vehicles. *Institute of Transportation Studies*. [Online]. 1 Jun,. California, U.S. Disponible en `http://repositories.cdlib.org/itsdavis/UCD-ITS-RR-04-4`

[7] Campbell, N. R. (1957). *Foundations of Science: The philosophy of Theory and Experimentation*. New York: Dover.pp 68

[8] Campbell, S. K.(2002). *Flaws and Fallacies in Statistical Thinking*. New York: Dover.pp 13

[9] Chatfield, C. & Collins, A. (1980). *Introduction to Multivariate Analysis*. London: Chapman and Hall.

[10] Cox, T. & Cox, M. (2001). *Multidimensional Scaling*. New York: Chapman & Hall/CRC.

[11] Cramer, J.S. (2003). The origins and development of the logit model. *Cambridge Resources*. [Online]. Disponible en: `http://www.cambridge.org/resources/0521815886/1208\ default.pdf`

[12] Everitt, B. S. (1993). *Cluster Analysis*. New York: John Wiley & Sons Inc.

[13] Everitt, B. S. & Dunn, G.(2000). *Applied Multivariate Data Analysis*. London: Arnold.

[14] Everitt, B. S. & Landau, S.(2004). *A Handbook of Statistical Analyses using SPSS*. Boca Raton: Chapman & Hall.

[15] Farr, W. A. & Hard, J. S. (1987). Multivariate analysis of climate along the southern coast of Alaska some forestry implications. *Department of Agriculture, Forest Service, Pacific Northwest Research Station*. [Online]. Portland, U.S. Disponible en: `http://www.treesearch.fs.fed.us/pubs/8990`

[16] Forrest W. Young. ALSCAL:Software for Multidimensional Scaling. [Online]. Disponible en: `http://forrest.psych.unc.edu/research/alscal.htm` [Cited January 19,2007]

[17] Gordon, A. (1990). *Classification*. New York: Chapman & Hall/CRC.

[18] Hernandez, L. (2001). *Técnicas de Taxonomia Numérica*. España: La Muralla.

[19] Hsieh, W. W. (2001). Nonlinear Canonical Correlation Analysis of the Tropical Pacific Climate Variability Using a Neural Network Approach. *Journal of Climate*,14, 2528–2539. [Online]. Jun. Washington, U.S. Disponible en: `http://ams.allenpress.com/perlserv/?request=get-document\&doi=10.1175\%2F1520-0442(2001)014\%3C2528:NCCAOT\%3E2.0.CO\%3B2`

[20] Jeansonne, A. (2002). Loglinear Models. [Online].(Update 16 Sep) Disponible en: `http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log\%20Linear\%20Models.htm` [Citado el 16 de Enero del 2007]

[21] Jimenez, R. (2002). *Los desafíos de la seguridad pública en México. Percepción negativa de la seguridad pública: Ciudad de México y República Mexicana*. [e-book]. México: Instituto de Investigaciones Jurídicas de la UNAM. Serie Doctrina Jurídica, Num.120. Disponible en: `http://www.bibliojuridica.org/libros/1/419/14.pdf` [Citado en 10 de febrero del 2007]

[22] Jobson, J. (1992). *Applied Multivariate Data Analysis*. New York: Springer-Verlag.

[23] Johnson, R. & Wichern, D. (1988). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

[24] Kaufman, Leonard & Rousseeuw (1990). *Finding Groups in Data: An introduccion to Cluster Analysis*.New York: John Wiley & Sons

[25] Kendall, Sir M.(1980). *Multivariate Analysis*. Bristol, Great Britain: Charles Griffin & Company LTD.

[26] Kishino, H. & Waddell, P. J. (2000). Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data.*Genome informatics*. Workshop on Genome Informatics, 11, 83-95p. [Online]. Tokyo, Japan. Disponible en: `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Journals\&term=\%22Genome+Inform+Ser+Workshop+Genome+Inform\%22[Title+Abbreviation]`

[27] Krzanowski, W. J & Marriot F.H.C. (1995). *Multivariate Analysis Part 2: Classification, covariance structures and repeated measurementes*. New York: Arnold.

[28] Lebart, L., Morineau, A. & Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: correspondence analysis and related techniques for large matrices*. United States: John Wiley & Sons.

[29] Manly, B. F. J. (1994). *Multivariate Statistical Methods*. London: Chapman & Hall.

[30] Marriott, F. H. C. (1974). *The Interpretation of Multiple Observations*. London: Academic Press.

[31] Martínez, R.(1999). *El Análisis Multivariante en la Investigación Científica*. Madrid: La Muralla.

[32] Mary J. Allen & Wendy M. Yen (2002). *Introduction to Measurement Theory*. Long Grove: Weveland Press.

[33] Morrison, D. F. (1990).*Multivariate Statistical Methods*. New York: McGraw-Hill.

[34] National Council of Teachers of Mathematics (1970). *Medida*. México: Trillas.

[35] Peñaloza, P. y Garza, M. (2002). *Los Desafíos de la Seguridad Pública en México*. [e-book]. México: Instituto de Investigaciones Jurídicas de la UNAM. Serie Doctrina Jurídica, Num.120. Disponible en:`http://www.bibliojuridica.org/libros/libro.htm?l=419`

[36] Prescher, A; Meyers, A. & Graf von Keyserlingk, D. (2005). Instrumentation of multi-centric craniofacial collections. *Durham Anthropology Journal*,12, 2-3. [Online]. Durham, Uk. Disponible en: `http://www.dur.ac.uk/anthropology.journal/vol12/iss2-3/prescher/prescher.html`

[37] Real Deus, J. E. (2001). *Escalamiento Multidimensional*. Madrid: La Muralla.

[38] Richerme, M. (2001). *Eleven Multivariate Analysis Techniques: Key Tools In Your Marketing Research Survival Kit*. Texas: Decision Analyst.

[39] Romero,J.(2002). *La Seguridad Pública en México*.[e-book]. México:Instituto de Investigaciones Jurídicas de la UNAM.Serie Doctrina Jurídica, Num.120. Disponible en: `http://www.bibliojuridica.org/libros/1/419/20.pdf`

[40] Saint-Arnaud, S. & Bernard, P. (2003). Convergence or Resilience? A Hierarchical Cluster Analysis of the Welfare Regimes in Advanced Countries. Current Sociology,51, 499-527 . [Online]. Toronto, Canada. Disponible en: `http://csi.sagepub.com/cgi/content/abstract/51/5/499`

[41] Siegel, S. (1980). *Estadística no paramétrica aplicada a las Ciencias Sociales*. México: Trillas.

[42] Siromoney, G., Bagavandas, M. & Govindaraju, S. (1980). An application of component analysis to the study of South Indian sculptures. *Computers and Humanities*, Vol. 14, pp. 29-37. [Online]. India. Disponible en: `http://www.cmi.ac.in/gift/Iconometry/icon\ southindian.htm`

[43] SPSS (2006). *SPSS Advanced Models*. [Online]. Disponible en: `http://www.spss.com/advanced\ models/data\ analysis.htm` [Cited January 15,2007]

[44] SPSS (2006). *SPSS Complex Samples*. [Online]. Available from: http://www.spss.com/complex\ samples/ [Cited January 15,2007]

[45] Stamovlais, D., Tsaparlis, G., Kamilatos, C., Papaoikonomou, D. & Zarotiadou, E.(2004) Conceptual Understanding versus Algorithmic Problem Solving: A Principal Component Analysis of a National Examination. *The Chemical Educador*, 9,398-405. [Online]. 24 Nov. US. Disponible en: http://chemeducator.org/bibs/0009006/960398gt.htm

[46] Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, 9. [Online]. 22 Sep. Disponible en: http://www.amstat.org/publications/jse/v9n3/stanton.html

[47] StatSoft, Inc. (2006). Electronic Statistics Textbook. [Online]. Tulsa, OK: StatSoft. Disponible en:http://www.statsoft.com/textbook/stathome.html.

[48] Stevens, S. S. (1946). *On the theory of scales of measurement*. Science, 103, 677-680.

[49] Stevens, S. S. (1951). *Mathematics, measurement and psychophysics*. In S.S. Stevens (Ed.), Handbook of experimental psychology. New York: Adisson Wesley.

[50] Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1), 65-72. [Online]. Disponible en: http://www.spss.com/research/wilkinson/Publications/Stevens.pdf [citado el 15 de julio 2006]

[51] Visauta, B. (1998). *Análisis estadístico con SPSS para Windows*. España: McGraw Hill.