UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"Análisis de Conglomerados para Datos Direccionales"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

INGRITH RIVERA CABRERA

DIRECTORA DE TESIS: Mat. MARGARITA ELVIRA CHÁVEZ CANO

2006

Por ser mi mejor amigo, Dios te dedico mi tesis.

Dios, por tu eterna compañía y por tu infinito amor ¡mil gracias! Pero sobre todo GRACIAS por mandarme con tus tres mejores ángeles para que me guiaran, mi familia.

Papi, te agradezco por escucharme incluso en mis silencios, por orientarme, por proteger mis sueños, por cuidarme como el mejor padre y por ser los brazos en los que me sujeto.
Te quiero muchísimo viejito.

Mamá, no sólo eres mi persona favorita eres el impulso en mis pasos y mi fuerza para darlos. Gracias por tu inmenso amor, por tu apoyo, por tus consejos y por ser en mi vida mi sentimiento más sincero. Porque sin ti no lo hubiera podido hacer, gracias.
Te amo mami.

Maithé, gracias por ser mi mejor sonrisa, el hombro en el que me apoyo, la mano amiga en la que más confío, mi mejor ejemplo, la paz de la que me abrigo y porque el sentido de triunfar me lo enseñaste tú. Es un verdadero honor crecer a tu lado.
Te adoro nena.

Por el hermoso equipo que hacemos juntos y porque este logro no es más mío que suyo ...
... ¡gracias!

A usted, estimada Profesora Margarita, mi profunda gratitud y mi sincera admiración. Por su tiempo, dedicación y por sus invaluables enseñanzas
MIL GRACIAS.

UNAM, porque te debo mucho de lo que soy, gracias.

Análisis de Conglomerados para Datos Direccionales

Índice

Introducción

Una de las destrezas más remotas del hombre comprende el agrupamiento de objetos similares que producen una clasificación. La clasificación de objetos ha tenido una relevante importancia en el desarrollo de teorías en muchos campos de la ciencia, como son la medicina, la biología, la psicología, entre otras no menos importantes.

Hoy día existe una considerable cantidad de técnicas numéricas de clasificación, mismas que reciben diversos nombres dependiendo del área de aplicación. No obstante, el término genérico más común es análisis de conglomerados. Es decir, el análisis de conglomerados consiste en asignar o clasificar un conjunto de n individuos u objetos con p características en grupos mutuamente excluyentes y exhaustivos.

Existe una inmensa cantidad de conjuntos de datos que pueden ser agrupados, entre los cuales se hallan aquellos que pueden ser representados en un círculo, ya sea que de manera directa estén medidos en ángulos o que se haga mediante una transformación. A estas variables se les llama variables circulares o direccionales.

El objetivo de este trabajo de tesis es presentar una estadística que permita formar conglomerados en datos circulares, y a través de la maximización de ésta identificar cual es el número óptimo de grupos que configuran los datos.

Para alcanzar este objetivo se han contemplado 3 capítulos. En el capítulo 1 se describen los coeficientes de similaridad y disimilaridad para variables que no son direccionales, estos coeficientes serán la herramienta básica de los métodos jerárquicos de agrupamiento, analizados también en este capítulo. Los métodos jerárquicos de agrupamiento, divididos en métodos aglomerativos y divisivos, permitirán la construcción de diagramas de árbol (dendrogramas), mismos que de manera gráfica constituirán los grupos en los datos. Sin embargo, en este capítulo no sólo se contempla el análisis gráfico de los dendrogramas para la identificación de conglomerados, sino que también se consideran criterios

numéricos, no jerárquicos, de agrupamiento que permiten producir una partición de los individuos u objetos en un número particular de grupos.

En el capítulo 2 se presentan medidas de similaridad y disimilaridad para datos direccionales. Así como también se define la dirección media muestral que depende del vector medio resultante, asimismo se define y analiza la distancia media resultante poblacional y muestral. Todo esto con la intención de definir la estadística que forma los grupos entre los datos direccionales. Al graficar las estadísticas de varios números posibles de grupos se puede fácilmente interpretar la gráfica, la cual se utilizará para determinar el número óptimo de grupos en los datos. Específicamente, el número óptimo de grupos en los datos será el que maximice el valor de la estadística propuesta.

En el capítulo 3 se presenta la aplicación de la estadística sugerida en el capítulo 2, mediante un ejemplo concerniente a la dirección a la que se dirigen las tortugas después de desovar.

Se concluye que la estadística S_k, basada en las diferencias de las distancias medias resultantes muestral y poblacional de los k grupos, permite desarrollar un método jerárquico de agrupamiento divisivo para variables circulares. Además dicha estadística identifica cuál es el número óptimo de grupos en los mismos.

Capítulo 1

Análisis de Conglomerados

(CLUSTER ANALYSIS)

Una de las habilidades más básicas de las criaturas vivientes involucra el agrupamiento de objetos similares que producen una clasificación. La idea de repartir objetos similares en categorías es claramente una idea primitiva de clasificación.

La clasificación ha jugado un papel central en el desarrollo de teorías en muchos campos de la ciencia. Por ejemplo la clasificación de los elementos de la tabla periódica, realizada por Mendeleiev en los años 1860, tuvo un impacto profundo en el entendimiento de la estructura del átomo. Otro ejemplo, esta vez en astronomía, es la clasificación de las estrellas en estrellas enanas y estrellas gigantes, usando la gráfica de temperatura contra luminosidad de Hertsprung-Russell, que afectó ampliamente a las teorías de la evolución estelar.

Un procedimiento de clasificación puede representar simplemente un método conveniente para organizar un gran conjunto de datos de manera que la recuperación de información pueda ser más eficiente. El punto importante es que una clasificación es una división de los objetos o individuos en grupos basada en una serie de reglas.

Durante la segunda mitad del siglo veinte se incrementó dramáticamente el número de técnicas numéricas de clasificación. Hoy en día dichas técnicas son usadas en diferentes campos tales como la arqueología, la psiquiatría, investigación de mercados y astronomía.

Un gran número de nombres se han usado para estos métodos dependiendo del área de aplicación. Por ejemplo, taxonomía numérica es generalmente usada en biología; en psicología es algunas veces utilizado el término análisis Q; en la literatura de la

inteligencia artificial el reconocimiento de patrones no supervisado es como las técnicas de clasificación son comúnmente llamadas. En otras áreas *clumping y grouping* han sido utilizados ocasionalmente. En la actualidad el término genérico más común es análisis de conglomerados (*cluster analysis*).

El problema a las que estas técnicas se orientan puede ser planteado, en general, como sigue:

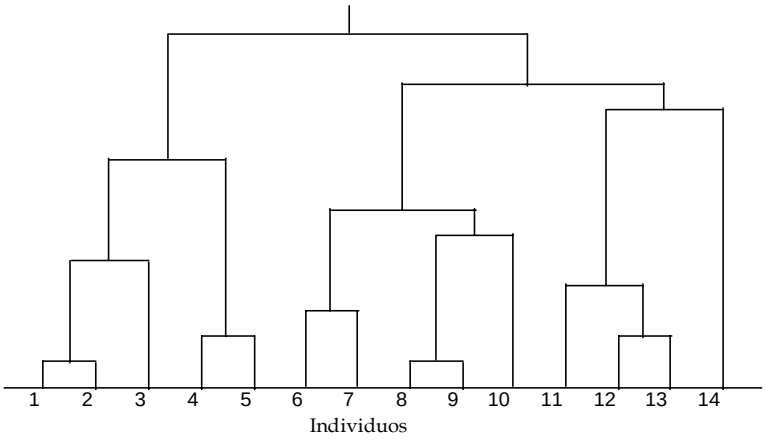
Dada una colección de *n* objetos o individuos, animales, plantas, etc., cada uno de los cuales es descrito por un conjunto de *p* características o variables, se deriva una división en un cierto número de clases. Tanto el número de clases como las propiedades de las clases están por ser determinadas.

El análisis de conglomerados consiste en asignar o clasificar un conjunto de *n* individuos u objetos con *p* características en grupos mutuamente excluyentes y exhaustivos. Los individuos que pertenecen a un mismo grupo son similares unos a otros mientras que son diferentes a los individuos de otros grupos. A este conjunto de grupos usualmente se le llama *partición*.

Los grupos que forman una partición pueden ser subdivididos dentro de conjuntos más pequeños o agrupados en conjuntos más grades, de tal forma que eventualmente se culmina con una estructura jerárquica dada por el conjunto original de individuos, esta estructura es frecuentemente llamada *árbol jerárquico o dendrograma*, cuya definición se verá más adelante.

Ejemplo de un árbol jerárquico[1]:

[1] CHATFIELD, C. Collins "Introduction to Multivariate Analysis" Ed. Chapman and Hall, Londres 1980. pág 213.



Siempre se puede tener una partición desde un árbol jerárquico al graficar una línea horizontal a través del árbol en un punto apropiado; esto algunas veces recibe el nombre de 'corte del árbol'.

De acuerdo con Jain y Dubes (1988) 'el análisis de conglomerados es una herramienta para la exploración de datos y debe ser complementada con técnicas para visualizar datos'.

Para ilustrar el rango de disciplinas en las cuales el análisis de conglomerados ha sido utilizado se describirán brevemente varios de ejemplos:

- Medicina

En 1973, Robert Barclay Fetter desarrolló los Grupos Relacionados de Diagnóstico (GRD), un fascinante uso de los conglomerados para obtener una clasificación de pacientes hospitalizados. Los GRD son un modelo de clasificación que agrupa a los pacientes con base en el consumo de recursos que requiere su atención y en las características clínicas que se presenten. Estos modelos fueron desarrollados inicialmente como una herramienta para administrar los costos y ayudar a las clínicas y hospitales a monitorear la utilización y calidad de los servicios.

- Psiquiatría

Las enfermedades de la mente son más difíciles de encontrar que las enfermedades del cuerpo y ha sido de mucho interés en la psiquiatría el uso de las técnicas del análisis de conglomerados para refinar o redefinir las categorías de los diagnósticos en curso. Mucho de este trabajo ha involucrado pacientes depresivos que son el centro de primordial interés en la existencia de subtipos endógenos y neuróticos.

-Investigación de mercado

Un gran número de ciudades están disponibles para estas pruebas de mercado, pero debido a causas de factores económicos el estudio debe ser restringido a sólo un pequeño número de éstas. La manera de selección de las ciudades para aplicar las pruebas de mercado es primero conglomerar las ciudades en un pequeño número de grupos tal que las ciudades dentro de un grupo sean muy similares la una a la otra, y después elegir una ciudad de cada grupo. Green (1967) adoptó este planteamiento, clasificando 88 ciudades con base en 14 variables que fueron el tamaño de la ciudad, los periódicos de circulación, el ingreso *per capita*, entre otras no menos importantes.

-Educación

Aitkin, Anderson y Hinde (1981) enseñaron conglomerados de distintas maneras con base en varias variables binarias describiendo el comportamiento pedagógico, por ejemplo: ¿Los alumnos tienen una elección de dónde sentarse?, ¿Usan un horario para organizar el trabajo?, ¿Se dan las estrellas a alumnos que producen el trabajo mejor?. Los conglomerados producidos identificaron como 'formal' e 'informal' las maneras de enseñar.

-Arqueología

Hodson (1971) usó la técnica de conglomerados de *k* medias para construir una taxonomía de las herramientas manuales encontradas en las Islas Británicas. Las variables usadas para describir cada una de las herramientas incluía longitud, grosor y la precisión. El análisis dio como resultado dos grupos que contenían herramientas delgadas y pequeñas, y el otro grupo lo formaban la herramientas gruesas y largas.

El análisis de conglomerados cubre una variedad de *objetivos*, estos son:

a) Exploración de datos.
b) Reducción de datos.
c) Generación de hipótesis.
d) Predicción basada en grupos.

Hay tres grandes temas íntimamente relacionados al análisis de grupos que son:

- *Clumping*.
- Disección.
- Variables de agrupamiento.

El término *clumping* es usualmente aplicado a los métodos de agrupamiento donde se detecta que los grupos formados se traslapan. Por ejemplo, al tratar de clasificar palabras de acuerdo a su significado se encontrará que algunas palabras tienen dos o más significados y necesitan ser asignadas en más de un grupo.

El término disección es usado cuando se tiene una población homogénea en la que no hay una manera natural de agrupar a los individuos y aún así se desea dividir a la población en subgrupos. Por ejemplo, cuando se desea dividir a una ciudad en distritos postales y los grupos son claramente arbitrarios.

El objetivo de las variables de agrupamiento es ver si se pueden encontrar subconjuntos de variables que estén altamente correlacionadas entre ellas y que se pueda usar sólo alguna de ellas, o algún promedio de ellas, para poder representar al conjunto total sin tener una pérdida seria de información.

Es por ello que es importante diferenciar entre medidas en las variables y entre los individuos; pues para poder llevar a cabo lo anterior es necesario contar con coeficientes de similaridad (o de disimilaridad) entre cada par de variables. Es lógico pensar que dicha

similaridad es en alguna forma el coeficiente de correlación, entonces dos variables con alta correlación podrían dar los mismos efectos.

Una manera alternativa de unir o agrupar variables es a través del análisis de componentes principales. Al aplicar el análisis de componentes principales, si se encuentra que las primeras dos componentes 'explican' una gran proporción de la varianza total, se puede graficar los datos con respecto a estas dos componentes para cada individuo, de tal forma que se pueda buscar los grupos visualmente. Si más de dos componentes son necesarias para dar una representación satisfactoria de los datos, entonces el análisis de componentes principales no es tan recomendable o seguro y es más fácil tratar con un algoritmo de agrupamiento.

Con respecto a las variables, en la mayoría de los casos hay probabilidad de tener teóricamente límites en el número de variables que pudieran ser utilizadas para producir una clasificación. En la práctica, por supuesto, muchas serán consideradas irrelevantes de acuerdo al propósito del que se trate, y una restricción más fuerte en el número puede incrementarse dependiendo de las consideraciones económicas. Entonces no hay, en general, ninguna base teórica legítima para determinar el número de variables a utilizar y el problema debe aproximarse por consiguiente empíricamente. Es importante considerar que la presencia de variables adicionales que no son importantes puede alterar la estructura del conglomerado.

Un problema más fuerte, común a todas las ramas del análisis multivariado, es la posibilidad de *pérdida de datos*. Esto puede ocurrir debido a una variedad de razones y pude ser tratado de diferentes maneras. La más simple es considerar sólo a los individuos que tengan un conjunto completo de valores de las variables. Sin embargo, en algunos casos esto puede reducir severamente al número de individuos disponibles para el análisis. Una propuesta alternativa es reemplazar los valores perdidos por los valores estimados. Para algunas técnicas multivariadas esto puede ser una alternativa razonable, en análisis de conglomerados no lo es. La media debería ser calculada sólo por aquellos individuos que pertenecen al mismo grupo incluyendo los individuos con datos

incompletos. Pero tal cálculo en un grupo específico no es posible porque los grupos son, por supuesto, desconocidos.

En muchas aplicaciones las variables que describen a los objetos no serán medidas en las mismas unidades. De hecho, frecuentemente dichas variables son de tipos completamente diferentes, algunas categóricas, otras ordinales y otras tienen una escala de intervalo. La solución sugerida con mayor frecuencia es la simple estandarización antes del análisis, usando la desviación estándar calculada del conjunto completo de objetos. Sin embargo, Fleiss y Zubin demuestran que esto puede tener serias desventajas, como diluir las diferencias entre grupos de variables, un punto dado a notar también por Duda y Hart.

Se han hecho muchas sugerencias acerca de como se podrían utilizar simultáneamente, en un análisis, variables de diferentes tipos. El planteamiento más simple es convertir todas las variables en forma binaria. Esto tiene la ventaja de ser directo, pero la desventaja es que se sacrifica potencialmente información útil. Una alternativa más atractiva es usar un coeficiente de similaridad que pueda incorporar información de diferentes tipos de variables de un modo razonable.

1.1 Coeficiente de similaridad y disimilaridad

Hasta ahora se ha dicho que para que los individuos pertenezcan a un mismo grupo se necesita que sean similares entre sí, para ello se requiere una medida de similaridad (o de disimilaridad o diferencia) para cada pareja de individuos. Algunas veces las similaridades son observadas directamente, mientras que en otros casos éstas son derivadas de una matriz con información apropiada. La distancia euclidiana estandarizada es una de las medidas más comunes de disimilaridad.

Un coeficiente de similaridad indica la fuerte relación entre dos objetos, dados los valores del conjunto de las *p* variables en común. La similaridad entre dos objetos *i* y *j*, será alguna función de los valores observados, es decir,

$$s_{ij} = f\left(\mathrm{x}_i, \mathrm{x}_j\right)$$

donde $\mathrm{x}_i = \left\lfloor x_{i1}, x_{i2}, ..., x_{ip} \right\rfloor$ y $\mathrm{x}_j = \left\lfloor x_{j1}, x_{j2}, ..., x_{jp} \right\rfloor$ son los valores observados de las variables de cada objeto o individuo. Muchas funciones han sido propuestas dependiendo, en parte, del tipo de variable concerniente (cuantitativa, categórica, binaria, ordinal, etc.).

Normalmente se considera a la similaridad como una relación simétrica requiriendo que $s_{ij} = s_{ji}$. La mayoría de los coeficientes de similaridad son no negativos y son ajustados para tener a la unidad como límite superior, aunque algunos son correlaciones, de manera que $-1 \leq s_{ij} \leq 1$.

Asociado con cada medida de similaridad, limitada por cero y la unidad, hay una disimilaridad $d_{ij} = 1 - s_{ij}$ que es simétrica y no negativa. El grado de similaridad entre dos objetos se incrementa con s_{ij} y decrece con d_{ij}. Es natural para un objeto tener la máxima similaridad con él mismo, así que $s_{ii} = 1$ y $d_{ii} = 0$.

Un coeficiente de disimilaridad es una función *d* que va de *P* x *P* a los reales no negativos, la cual [2]:

$$d\left(A, B\right) \geq 0, \qquad \text{para todo } A, B \in P$$
$$d\left(A, A\right) = 0, \qquad \text{para todo } A \in P$$
$$d\left(A, B\right) = d\left(B, A\right), \quad \text{para todo } A, B \in P$$

[2] JARDINE & SIBSON, "Mathematical Taxonomy", Ed. Jhon Wiley & Sons Ltd., 1971. pág. 6.

1.1.1 Disimilaridades y medidas de distancia

Una función de valor real $d(A,B)$ que va de P x P es una *función de distancia* si satisface, para todo $A, B, C \in P$, las siguientes propiedades [3]:

i) $d(A,B) = d(B,A)$

ii) $d(A,B) \geq 0$

iii) $d(A,A) = 0$

Para muchas funciones de distancia las siguientes propiedades también se cumplen:

iv) $d(A,A) = 0$ si y sólo si $A = B$

v) $d(A,B) \leq d(A,C) + d(C,B)$

Si d cumple de i) – v) es llamada una *métrica*.

Algunos coeficientes de disimilaridad tienen la propiedad de métrica, $d_{ij} + d_{ik} \geq d_{jk}$ para todo *i, j* y *k,* en cuyo caso son conocidos generalmente como *medidas de distancia*. La medida de distancia más comúnmente utilizada y la más familiar es la Euclidiana, pero puede ser muy poco satisfactoria puesto que su valor depende principalmente de las escalas elegidas para las variables.

Una alternativa es usar la distancia de Mahalanobis, para dos individuos *i* y *j* con vectores de medidas x_i y x_j respectivamente.

$$d_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j) .$$

[3] MARDIA, Kantilal Varichand et. al. "Multivariate Analysis" Ed. Academic Press, Londres 1995. pág. 376.

La matriz S en la fórmula anterior usualmente se toma por ser la matriz de varianzas y covarianzas estimadas.

A pesar de que la distancia Euclidiana es la más usada en el contexto de conglomerados, se han empleado otras medidas de distancia.

1.1.2 Medidas de similaridad para variables dicotómicas

Los valores de las variables dicotómicas en algunos casos indican la presencia, o bien la ausencia, de alguna característica, pero también pueden indicar si el individuo tiene alguna de las dos características alternativas, por ejemplo hombre/mujer o áspero/liso. Tales datos de dos individuos *i* y *j* pueden ser arreglados en una tabla de 2x2. Dicha tabla, como se usa en las aplicaciones de conglomerados, es principalmente una manera conveniente de colocar los datos y no debe confundirse con la usual tabla de contingencia 2x2.

		Individuo i		
		1	2	
Individuo	1	A	B	A+B
j	2	C	D	C+D
		A+C	B+D	P

Por ejemplo:

	Variable									
	1	2	3	4	5	6	7	8	9	10
Individuo1	1	0	0	0	1	1	0	0	1	0
Individuo2	0	0	0	0	1	0	0	1	1	0

La correspondiente tabla de 2x2 es:

		Individuo 1		
		1	0	
Individuo 2	1	2	1	3
	0	2	5	7
		4	6	10

Los coeficientes de similaridad más simples y más comúnmente usados son los siguientes para variables dicotómicas.

i) $\frac{A+D}{P}$

ii) $\frac{A}{A+B+C}$

iii) $\frac{2A}{2A+B+C}$

iv) $\frac{2(A+D)}{2(A+D)+B+C}$

v) $\frac{A}{A+2(B+C)}$

Los dos coeficientes más utilizados en la práctica son los coeficientes (i) y el coeficiente de Jaccard (ii). El primero es simplemente la relación entre el número total de variables correspondientes a los dos individuos, con respecto al número de variables total; el segundo es la correspondiente relación cuando la correspondencia 'negativa' D es ignorada.

Sokal y Sneath (1963) dan una discusión completa de coeficientes de similaridad para el uso de datos binarios y sostienen que cada juego de datos debe ser considerado en sus cualidades por el investigador más familiar con el material involucrado.

1.1.3 Medidas de similaridad para variables cuantitativas

Además de la conocida distancia Euclidiana, una medida de similaridad que se ha usado ampliamente en las variables cuantitativas es el coeficiente de correlación muestral de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cuando es usado como una medida de similaridad para dos individuos, su cálculo involucra el promedio de los valores de las variables cuantitativas diferentes para producir un 'valor medio de variable' para cada individuo.

Se ha sugerido a menudo que el coeficiente de correlación es una medida útil de similaridad en esas situaciones donde 'el tamaño' absoluto sólo se ve como menos importante que 'la forma'. Por ejemplo, en la clasificación de animales y plantas el tamaño absoluto del organismo o de otras partes son frecuentemente menos importantes que las formas.

1.1.4 Medidas de similaridad para variables de tipo mixto

Un coeficiente de similaridad sugerido por Gower (1971) es particularmente útil en este tipo de datos, definido como

$$s_{ij} = \frac{\sum_{k=1}^{P} w_{ijk} s_{ijk}}{\sum_{k=1}^{P} w_{ijk}}$$

En esta fórmula, s_{ijk} es la similaridad entre el *i*-ésimo y *j*-ésimo individuos medidos por la *k*-ésima variable y w_{ijk} es típicamente 1 ó 0 dependiendo de si la comparación es considerada válida o no para la *k*-ésima variable. Se asigna un cero cuando la variable *k* es desconocida para uno o ambos individuos. Para los datos categóricos, s_{ijk} toma el valor uno cuando los dos individuos tienen el mismo valor y toma el valor de cero en otro caso.

1.2 Los dendrogramas y la desigualdad ultramétrica

Un árbol puede ser definido como un anidamiento secuencial de particiones de los individuos en *g* grupos, donde *g* varía de 1 a *n* (que es el total de individuos) con la propiedad de que las particiones en *g* y en (*g+1*) grupos es tal que (*g*-1) de los grupos son idénticos mientras que el resto de los individuos forman un grupo en el primer caso y dos grupos en el segundo caso. Hartigan (1975) definió un árbol como una familia de grupos, en donde dos grupos son disjuntos o están incluidos uno en el otro. La estructura jerárquica es frecuentemente representada por un diagrama bidimensional. Este diagrama es llamado *diagrama de árbol o dendrograma*.

Es decir, un dendrograma es un diagrama de árbol en el cual el eje de las equis representa a los 'objetos', mientras que el eje de las yes representa distancias. Las ramas del árbol dan el orden de las *n*-1 uniones; la primera horquilla representa la primera unión, la segunda horquilla la segunda unión, y así sucesivamente hasta que todos juntos están en el tronco[4].

Dado un conjunto de distancias observadas entre todas las parejas de individuos, existen muchas maneras en las cuales la distancia entre grupos o individuos puede ser definida. Habiendo elegido la definición más conveniente, el diagrama de árbol es graficado de tal forma que dos grupos se unen a través de una distancia derivada apropiada.

[4] MARDIA, Kantilal Varichand et. al. "Multivariate Analysis" Ed. Academic Press, Londres 1995. pág. 372.

Es decir, el diagrama de árbol también implica un nuevo conjunto de distancias entre individuos, las cuales pueden ser encontradas a partir de la distancia en el nivel más bajo del eslabón que une a dos individuos en el diagrama de árbol. Esta distancia derivada satisface las condiciones de un coeficiente de disimilaridad métrico, y además satisface la *desigualdad ultramétrica*:

$$d_{rs}^{*} \leq \max\left(d_{rt}^{*}, d_{ts}^{*}\right)$$

para todos los individuos *r, s, t*.

Una condición necesaria y suficiente para que un coeficiente de disimilaridad sea representado exactamente por un dendrograma es que satisfaga la desigualdad ultramétrica. Pero los coeficientes de disimilaridad más comunes no satisfacen la desigualdad anterior, así que se puede decir que usualmente no hay una estructura jerárquica genuina.

1.3 Métodos jerárquicos de agrupamiento

Una dificultad inmediata en el análisis de conglomerados es que no hay una manera satisfactoria de definir 'grupo'. Se desea que los grupos sean parte de un *p*-espacio donde los puntos estén densamente ubicados, pero que a la vez estén separados por partes con una densidad baja. Por otro lado, se desea que los grupos sean internamente coherentes pero separados de otros grupos.

Un método jerárquico de agrupamiento trata de encontrar un árbol tal que las distancias ultramétricas derivadas sean en algún sentido tan cercanas como sea posible a las distancias observadas. Esto explica porque un procedimiento para encontrar un árbol de un conjunto dado de disimilaridades observadas es algunas veces llamado una *transformación ultramétrica*.

Los métodos de agrupamiento pueden ser aplicados al mismo conjunto de datos y producir estructuras que sean substancialmente diferentes. Esto es debido a que la elección del método de agrupamiento implica imponer una estructura a la población.

La habilidad de los métodos de agrupamiento es que detectan la no existencia de grupos bien establecidos. Si una clasificación no existe, un problema más fuerte es que los datos pueden admitir más de una clasificación y la solución radicará en el propósito de los investigadores.

Se observa que hay muchos problemas prácticos involucrados en el análisis de conglomerados. Los resultados dependerán de una variedad de consideraciones, del método que se elija y cuáles variables fueron contempladas por ser importantes.

Existe una variedad de técnicas convenientes para proporcionar despliegues gráficos informativos de datos multivariados. Dichas técnicas son frecuentemente útiles para detectar la presencia de grupos, y además a menudo son más útiles aún para prevenir una demanda excesiva de la estructura de grupos producida por técnicas más complejas.

Las condiciones matemáticas que debería satisfacer un 'buen' método jerárquico de agrupamiento según Jardine y Sibson (1971) son:

- Los resultados producidos por un método no deben depender en la manera en la que los individuos estén etiquetados.

- Se requiere que un procedimiento de agrupamiento esté 'bien definido', esto es que se obtenga siempre el mismo árbol del mismo conjunto de disimilaridades observadas. La dificultad con esta condición incrementa cuando hay diferencias iguales, las cuales son resueltas en un orden arbitrario durante el proceso secuencial de encontrar el árbol. El método de la liga simple está 'bien definido' pero muchos otros no.

- La condición de continuidad, sugerida por Jardine y Sibson, trata de que un pequeño cambio en los datos debería sólo producir un pequeño cambio en el árbol resultante.

Otro conjunto importante de condiciones, a las que Jardine y Sibson llamaron condiciones 'ajuste conjunto', son:

- Si se añade o substrae sólo un individuo del conjunto original debería verse cambiada en muy poco la estructura del árbol, aunque algunas veces la clasificación puede cambiar en un sentido no tan trivial.

- Si se corta el árbol de tal manera que todos los individuos permanezcan en una sola rama del árbol, entonces la estructura de la rama debería permanecer invariante cuando los métodos de agrupamiento se vuelvan a aplicar al conjunto restante de individuos.

1.3.1 Algoritmos de agrupamiento jerárquicos

Es importante distinguir cuidadosamente entre un método de agrupamiento y un algoritmo para llevarlo a cabo. Técnicamente un método de agrupamiento mapea un conjunto de coeficientes de diferencias observadas a un nuevo conjunto de diferencias las cuales satisfacen la desigualdad ultramétrica y de ahí se describe un árbol jerárquico. Es importante realzar que hay muchos algoritmos diferentes en la actualidad para encontrar este mapeo.

En una clasificación jerárquica los datos no son particionados en un particular número de clases de grupos en un solo paso. En cambio la clasificación consiste en una serie de particiones que puede ir de un solo grupo contando con todos los individuos, a n grupos que cuentan con un solo individuo.

Con estos métodos, divisiones o fusiones, una vez hechos son irrevocables, así que cuando un algoritmo de conglomerado ha unido a dos individuos estos no pueden ser separados subsecuentemente. Como Kaufman y Rousseeuw (1990) comentaron 'un método jerárquico padece el defecto que nunca puede reparar lo que se hizo en pasos anteriores'.

Los grupos son formados por un proceso o algoritmos aglomerativos o divisivos.

- Los *algoritmos aglomerativos* empiezan por grupos de sólo un individuo. Los grupos más cercanos son gradualmente unidos hasta que finalmente todos los individuos están en un solo grupo.

- Los *algoritmos divisivos* operan por la división sucesiva de grupos, empezando con un solo grupo de *n* individuos y terminando con *n* grupos de sólo un individuo.

1.3.2 El método de la liga simple

El método más importante para encontrar un árbol jerárquico es el método llamado el método de la liga simple.

Este método fue descrito primero por Florek (1951) y más tarde por Sneath (1957) y por Johson (1967). El método de la liga simple está cercanamente relacionado a ciertos aspectos de teoría de gráficas. Una *gráfica* es un conjunto de nodos y de aristas entre parejas de nodos. Un conjunto de observaciones y sus disimilaridades pueden ser representados en una gráfica como nodos y aristas respectivamente. Una *gráfica de árbol expandido* es un conjunto de aristas las cuales proporcionan un único camino entre cada par de nodos. Un *árbol de expansión mínima* es el más corto de todos los árboles extendidos[5], como se mencionará más adelante.

[5] EVERITT, Brian S. "Cluster Analysis". Ed. Edward Arnold. ed 3°. Londres. 1993. pág. 57,60.

Este método puede ser definido como sigue: para cualquier distancia d^*, el conjunto de todos los individuos está dividido en g ($\leq n$) grupos, para los cuales se cumple que los individuos *r* y *s* están en el mismo grupo si existe una cadena de individuos r, *a, b, ..., q, s*, tales que las disimilaridades observadas en la cadena, llamadas $d_{ra}, d_{ab}, ..., d_{qs}$, son todas menores o iguales a d^*.

Existen muchos algoritmos numéricos diferentes para encontrar el método de la liga simple**.** El más fácil de realizar es el siguiente:

i) Iniciar con *n* 'grupos'; cada uno contiene justo a un individuo.
ii) Unir a los dos individuos más cercanos, por ejemplo *r* y *s*, en un solo grupo, entonces existen en este paso (*n*-1) grupos.
iii) La diferencia entre este nuevo grupo y cualquier otro individuo *t*, está definida por $\min(d_{rt}, d_{st})$.
iv) Unir a los dos grupos más cercanos, los cuales tendrán cada uno dos individuos o un individuo y el grupo formado en ii).
v) Construir nuevas diferencias entre los (*n*-2) grupos. Entonces se continua hasta combinar los grupos de tal forma que en cada escenario el número de grupos es reducido por uno y la diferencia entre cualesquiera dos grupos está definida por ser la diferencia entre los miembros más cercanos.

Un nombre alternativo que en algunas ocasiones se le da a este método es 'el método del vecino más cercano'.

El tipo de algoritmo descrito es un algoritmo aglomerativo, ya que al contar inicialmente con una matriz de distancias entre los individuos, éste opera por series de uniones, empezando por *n* grupos de sólo un individuo y terminando con un solo grupo de *n* individuos.

El método de la liga simple es el más solicitado matemáticamente. Éste es el único método jerárquico de agrupamiento que satisface todas las condiciones sugeridas por Jardine y Sibson y también tiene ventajas computacionales.

El método de la liga simple da soluciones invariantes bajo una transformación monótona de las medidas de disimilaridad.

La desventaja principal del método de la liga simple es llamado el efecto 'de encadenamiento' el cual incrementa cuando aparentemente los grupos distintos son unidos muy rápidamente por unos pocos puntos intermediarios.

A manera de ejemplo de la operación del algoritmo de la liga simple, el método será aplicado a la siguiente matriz de distancias[6]:

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

La entrada más pequeña en la matriz es la correspondiente a los individuos 1 y 2, consecuentemente éstos se unen en un conglomerado de dos miembros, las distancias entre este conglomerado y los otros tres individuos son obtenidas como:

$$d_{(12)3} = \min\left\lfloor d_{13,} d_{23} \right\rfloor = d_{23} = 5$$
$$d_{(12)4} = \min\left[d_{14,} d_{24} \right] = d_{24} = 9$$
$$d_{(12)5} = \min\left[d_{15,} d_{25} \right] = d_{25} = 8$$

[6] EVERITT, Brian S. "Cluster Analysis". Ed. Edward Arnold. ed 3°. Londres. 1993. pág. 58-61.

Ahora se puede construir una nueva matriz cuyas entradas son distancias entre individuos y distancias entre los individuos y el conglomerado. Es decir:

	(12)	3	4	5
(12)	0			
3	5	0		
4	9	4	0	
5	8	5	3	0

La entrada más pequeña en esta última matriz es la que corresponde a la distancia ente los individuos 4 y 5, así estos forman un segundo conglomerado constituido por dos miembros, y a su vez se tiene un nuevo conjunto de distancias que son calculadas de la siguiente manera:

$$d_{(12)3} = 5$$
$$d_{(12)(45)} = \min\left[d_{14,}d_{15,}d_{24,}d_{25}\right] = d_{25} = 8$$
$$d_{(45)3} = \min\left[d_{34,}d_{35}\right] = d_{34} = 4$$

Nótese que la distancia $d_{(12)3} = 5$ no se ve modificada en este paso, dicha distancia es la misma que con la que se ya se contaba.

Estas distancias se pueden escribir en una nueva matriz:

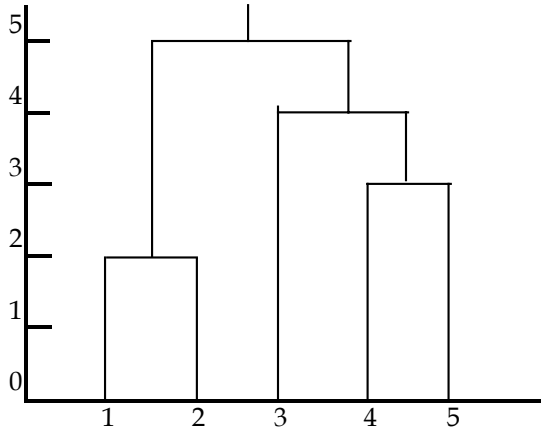
	(12)	3	(45)
(12)	0		
3	5	0	
(45)	8	4	0

La entrada más pequeña ahora es $d_{(45)3}$, por lo que el individuo 3 es añadido al conglomerado formado por los individuos 4 y 5. Finalmente, los grupos contienen a los individuos 1,2 y 3,4,5 y están unidos en un solo conglomerado.

Las particiones producidas en cada escenario, son las siguientes:

Etapa	Grupos
P_5	[1], [2], [3], [4], [5]
P_4	[1,2], [3], [4], [5]
P_3	[1,2], [3], [4,5]
P_2	[1,2], [3,4,5]
P_1	[1,2,3,4,5]

El correspondiente dendrograma se muestra a continuación:



Un punto importante a notar acerca de los resultados es que los conglomerados proceden jerárquicamente, cada uno es obtenido por la fusión de los conglomerados del nivel previo.

Un algoritmo alternativo es descrito por Goweer y Ross (1969) que se deriva del árbol de la liga simple vía un mecanismo llamado el árbol de mínima expansión. El árbol de mínima expansión no es un árbol jerárquico, sino una red que atraviesa todos los puntos (o individuos) por un conjunto de líneas rectas cuyas longitudes son iguales a las disimilaridades correspondientes entre dichos puntos.

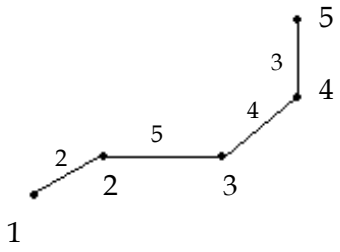
El árbol de mínima expansión es elegido de tal forma que:

i) todos los pares de puntos estén conectados.
ii) la suma de las longitudes de las líneas rectas que unen a los puntos sea la mínima.

Es fácil ver que el árbol de mínima expansión no contendrá ningún 'lazo', y que cada punto es visitado por al menos una línea. Si hay igualdad en las disimilaridades, el árbol de mínima expansión no será único.

A continuación se presenta el árbol de mínima expansión correspondiente a la matriz de disimilaridades de los datos vistos con anterioridad.

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0



1.3.3 Otros métodos jerárquicos de agrupamiento

- En el método de la liga completa o el vecino más lejano, la 'distancia' entre dos grupos está definida como la disimilaridad entre la pareja de individuos más lejana. En un sentido es exactamente lo opuesto a la definición de la liga simple.

- En el método del centroide, la 'distancia' entre dos grupos está definida por ser la 'distancia' entre el grupo de centroides (o grupo de vectores medios). Es decir, con este método, los grupos una vez formados, son representados por los valores medios de cada variable, que es su vector de medias (vector medio), y la distancia entre los grupos es ahora definida en términos de distancia entre dos vectores medios.

- En el método de grupos promedio, la 'distancia' entre dos grupos es definida como el promedio de las disimilaridades entre todos los pares de individuos, tal que hay un individuo en cada grupo.

- El método de conglomerados jerárquico de Ward (1963) está basado en la suma de cuadrados de cada grupo en vez de las ligas entre grupos. En cada fase el número de grupos es reducido en uno, combinando los dos grupos que dan el posible aumento más pequeño en la suma de cuadrados total dentro del grupo. Por supuesto cuando se inicia con *n* grupos de un solo individuo la suma de cuadrados total es cero.

- El método de Wishart, algunas veces llamado análisis de modo, busca los 'puntos densos', donde *k* o más puntos (o individuos) están contenidos dentro de una hiperesfera de radio R. Iniciando con un valor 'pequeño' de R, el método se parece a una hiperesfera de radio R al rededor de cada punto Si el número de puntos es por lo menos *k*, entonces el punto del centro se llama un *punto denso*. El parámetro R se aumenta gradualmente para que cada vez más puntos se vuelvan densos, hasta que todos los puntos permanezcan dentro de una sola hiperesfera. Si el parámetro *k* es igual a 1, es fácil ver que el método es equivalente al de la liga simple.

1.4 Métodos de optimización para el análisis de conglomerados

En los métodos de optimización para el análisis de conglomerados se consideran un conjunto de técnicas de agrupamiento para producir una partición de los individuos en un número particular de grupos, al minimizar o maximizar algún criterio numérico. Tales técnicas de optimización difieren de los métodos descritos anteriormente, ya que estos no forman clasificaciones jerárquicas de los datos. En un inicio, en estos métodos se asume que el número de grupos ha sido fijado por el investigador.

1.4.1 Criterios de agrupamiento

Se han sugerido muchos criterios de agrupamiento, la mayoría surgen comúnmente de consideraciones a las siguientes matrices, las cuales pueden ser calculadas de una partición de los datos.

$$T = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}\right)\left(x_{ij} - \bar{x}\right)'$$

$$W = \frac{1}{n-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_j\right)\left(x_{ij} - \bar{x}_i\right)'$$

$$B = \sum_{i=1}^{g} n_i\left(\bar{x}_i - \bar{x}\right)\left(\bar{x}_i - \bar{x}\right)'$$

Estas matrices de *p*x*p* (*p* es el número de variables) representan respectivamente la dispersión total (T), la dispersión dentro de los grupos (*within-group dispertion*) y la dispersión entre grupos (*between-group dispertion*), y satisface la ecuación:

$$T = W + B$$

Para p = 1 esta ecuación representa una relación entre escalares; simplemente la división de la suma de cuadrados total de una variable en la suma de cuadrados dentro del grupo y la suma de cuadrados entre grupos, familiar al análisis de varianza. En este caso un criterio natural para agrupar podría ser elegir la partición correspondiente al mínimo valor de la suma de cuadrados dentro del grupo, o equivalentemente, el máximo valor del término entre grupos. Para p >1 derivar el criterio de agrupamiento de la ecuación anterior no es tan claro, y se han sugerido muchas alternativas.

1.4.2 Minimización de la traza de la matriz W

Singleton y Kautz (1965) desarrollan una extensión obvia de la minimización del criterio de la suma de cuadrados dentro de los grupos sugerida en el caso p = 1, cuando los datos no son univariados, y es minimizar la suma de la suma de cuadrados dentro de los grupos, sobre todas las variables, y esto es minimizar la traza de la matriz W. Esto puede ser demostrado por ser equivalente a minimizar la suma de las distancias Euclidianas al cuadrado entre individuos y la media del grupo, esto es:

$$E = \sum d_{i,c(i)}^2$$

donde $d_{i,c(i)}$ es la distancia Euclidiana del individuo i a la media del grupo al cual éste es asignado (minimizar la traza de la matriz W es, por supuesto, equivalente también a maximizar la traza de la matriz B).

1.4.3 Minimización del determinante de la matriz W

En el análisis de varianza multivariado, una de las pruebas para las diferencias en los vectores medios de grupo está basada en la razón de los determinantes de las matrices de dispersión T y W. Grandes valores de $\frac{\det(T)}{\det(W)}$ indica que los vectores medios de grupo

difieren. Tales consideraciones llevaron a Friedman y Rubin (1967) a sugerir como un criterio la maximización de esta proporción. Subsecuentemente para todas las particiones de los *n* individuos en *g* grupos, T permanece el mismo y la maximización de $\frac{\det(T)}{\det(W)}$ es equivalente la minimización del det(W). Este criterio ha sido estudiado en detalle por Marriott (1971,1982).

1.4.4 Maximización de la traza de la matriz (BW^{-1})

Otro criterio sugerido por Friedman y Rubin (1967) es la maximización de la traza de la matriz obtenida del producto de la matriz de dispersión entre grupos y la inversa la matriz de dispersión dentro de grupos, es decir, la maximización de la matriz BW^{-1} . Esta función también es usada en el contexto del análisis de varianza multivariado, y es equivalente a lo que Rao (1952) llama la generalización de la distancia de Mahalanobis a más de dos grupos.

1.4.5 Optimización de los criterios de agrupamiento

Una vez que se ha seleccionado un criterio numérico conveniente de agrupamiento, se necesita dar la consideración de cómo elegir la *g* partición de los datos que lleve a su optimización. En teoría por supuesto el problema es simple; desafortunadamente el problema en la práctica no es así. Incluso con las computadoras de hoy, los números involucrados son inmensos, la enumeración completa de cada posible partición de *n* individuos en los *g* grupos simplemente no es posible. Algunos ejemplos tomados del autor Spath (1980) servirán para ilustrar la magnitud del problema[7]:

[7] EVERITT, Brian S. "Cluster Analysis". Ed. Edward Arnold. ed 3°. Londres. 1993. pág. 93.

N(15,3) = 2,375,101

N(20,4) = 45,232,115,901

N(25,8) =690,223,721,118,368,580

N(100,5) =10^{68}

donde N(n,g) es el número de distintas particiones de n individuos en g grupos no vacíos.

Una expresión general es dada por Liu (1968):

$$N(n,g) = \frac{1}{g}\sum_{i=0}^{g}(-1)^{g-i}\binom{g}{i}i^n$$

La poca practicidad de examinar cada posible partición ha llevado al desarrollo de algoritmos diseñados para buscar el valor óptimo de un criterio de agrupamiento que reestructure las particiones existentes y guarde la nueva sólo si proporciona una mejora; éstos son llamados algoritmos 'cuesta arriba' aunque en el caso del criterio que requiere minimización dichos algoritmos deben ser quizás denominados 'cuesta abajo'. Los pasos esenciales en estos algoritmos son:

a) Encontrar alguna partición inicial de los individuos en el número requerido de grupos.
b) Calcular el cambio en el criterio de agrupamiento producido por el movimiento de cada individuo del grupo donde estaba a otro.
c) Hacer el cambio que lleve a la mejora más grande en el valor del criterio del agrupamiento.
d) Repetir los pasos (b) y (c) hasta que el movimiento de un solo individuo no cause la mejora del criterio de agrupamiento.

Una vez eligiendo donde empezar, el proceso es ejecutado en una gran variedad de maneras. Una configuración del grupo inicial podría especificarse con base en el conocimiento anterior; podría ser el resultado de algún otro tipo de método de

agrupamiento, por ejemplo de algún método jerárquico. Una partición inicial podría escogerse al azar, o g puntos en el espacio p-dimensional podrían ser seleccionados de alguna manera para actuar como centros del grupo inicial. Las soluciones iniciales diferentes pueden llevar a un óptimo local diferente del criterio del agrupamiento, aunque con datos bien estructurados es razonable esperar convergencia al mismo, esperanzadamente global, óptimo de la mayoría de las configuraciones iniciales.

Antes de proceder a analizar las propiedades y los inconvenientes de la optimización de los criterios de agrupamiento, es útil considerar un pequeño ejemplo numérico de la aplicación del tipo de algoritmo descrito.

Considérese el siguiente conjunto de datos en siete individuos con dos variables[8]:

Individuo	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Este conjunto de datos está por ser agrupado en 2 grupos usando el método de minimización de la traza de la matriz W. Como primer paso para encontrar una partición inicial sensata se usa la distancia Euclidiana y se definen los grupos medios iniciales:

	Grupo 1	Grupo2
Individuo	1	4
Vector medio	[1.0 , 1.0]	[5.0 , 7.0]

[8] Ibídem. pág. 95-96.

Los individuos del grupo se examinan ahora en sucesión y se asignan al grupo medio al que ellos son cercanos, en términos de distancia Euclidiana. El vector medio se vuelve a calcular cada vez que un nuevo miembro es añadido. Esto permite seguir la siguiente serie de pasos:

	Grupo 1		Grupo 2	
	Individuo	Vector medio	Individuo	Vector medio
Paso 1	1	[1.0, 1.0]	4	[5.0, 7.0]
Paso 2	1, 2	[1.2, 1.5]	4	[5.0, 7.0]
Paso 3	1, 2, 3	[1.8, 2.3]	4	[5.0, 7.0]
Paso 4	1, 2, 3	[1.8, 2.3]	4, 5	[4.2, 6.0]
Paso 5	1, 2, 3	[1.8, 2.3]	4, 5, 6	[4.3, 5.7]
Paso 6	1, 2, 3	[1.8, 2.3]	4, 5, 6, 7	[4.1, 5.4]

Esto da la clasificación inicial; los dos grupos en esta fase tienen las siguientes características:

Grupo 1 Individuos 1, 2 y 3
Vector medio = [1.8 , 2.3]
Traza(W_1) = 6.84

Grupo 2 Individuos 4, 5, 6 y 7
Vector medio = [4.1 , 5.4]
Traza(W_2) = 5.38

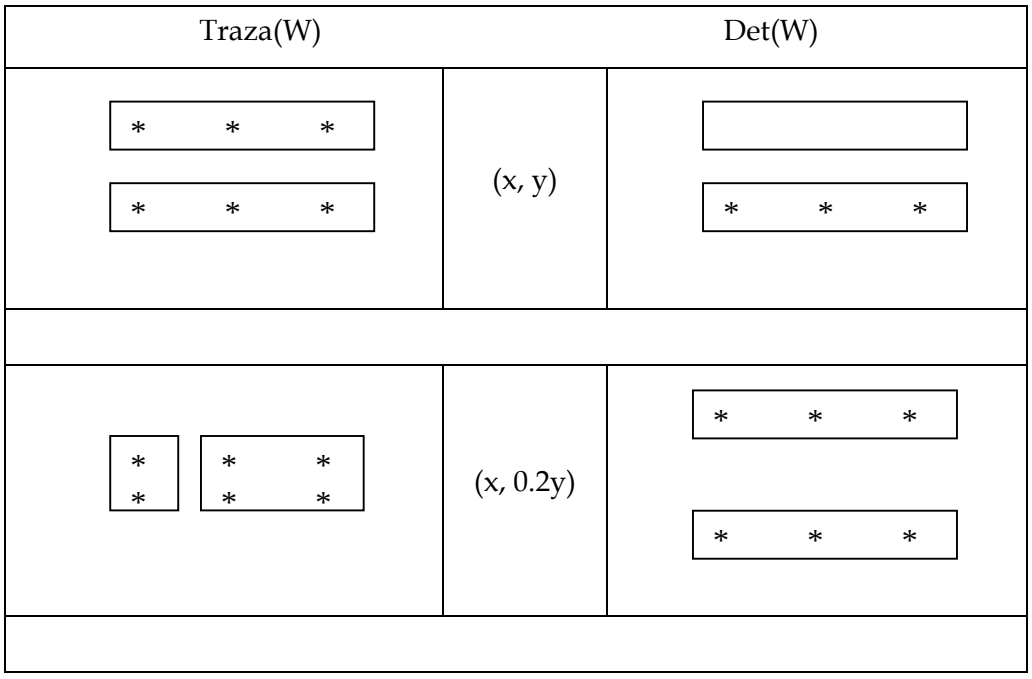
En este punto la traza(W) = 6.84 + 5.38 =12.22.

Considérese ahora que el individuo 3 se mueve al segundo grupo, teniendo así que traza(W_1) = 0.63, traza(W_2) = 7.90 y traza(W) = 8.53. Ya que el este movimiento causa un decremento en el criterio de agrupamiento, el movimiento es hecho, y el proceso interactivo continúa ahora de esta nueva partición.

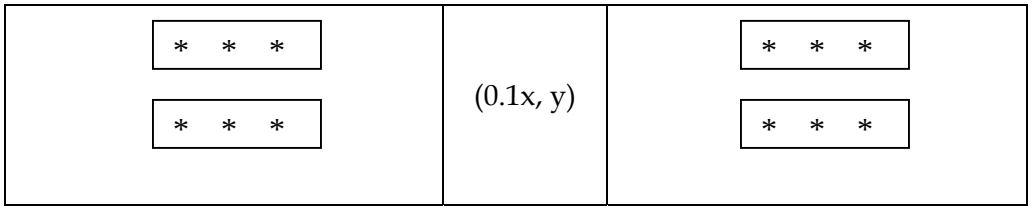
1.4.6 Propiedades e inconvenientes de la optimización de los criterios de agrupamiento

El criterio numérico de agrupamiento más usado comúnmente es la minimización de la traza(W), a pesar de que es bien sabido que sufre serios problemas. Primeramente el método es dependiente de una escala. Pueden obtenerse soluciones diferentes de los datos iniciales y de los datos estandarizados de alguna manera particular.

La dependencia de la escala del método de la traza(W) fue la motivación detrás de la búsqueda de Friedman y Rubin (1967) para un criterio alternativo que no estuviera afectado por la escala. Sus sugerencias, que se basaron en minimizar el det(W), se han usado ampliamente. A continuación se presenta una ilustración de la falta de dependencia de la escala de este último criterio en comparación con el de la traza(W) [9].



[9] Ibídem. pág. 99.



Desafortunadamente el criterio del det(W) supone que todo el grupo en los datos tiene la misma forma, y otra vez esto puede causar problemas cuando los datos no satisfacen dicho requisito.

En un esfuerzo por superar el problema de la 'forma similar' Scott y Symons (1971) sugirieron un método de agrupamiento basado en la minimización de

$$\prod_{i=1}^{g} \det(W_i)^{n_i}$$

La condición de que cada grupo contenga al menos p+1 individuos es necesario evitar una matriz de dispersión singular dentro del grupo, el determinante de la cual sería cero. Una posibilidad alternativa descrita por Maronna y Jacovkis (1974) es la minimización de

$$\sum_{i=1}^{g} (n_i - 1) \det(W_i)^{1/p}$$

La tendencia de que los criterios como la traza(W) y el det(W) den igual tamaño en los grupos ha sido comentada por varios autores. En un esfuerzo por superar este problema, Symons (1981) sugirió otros dos criterios de agrupamiento:

$$n \ln \det(W) - 2\sum n_i \ln n_i$$

$$\sum \left(n_i \ln \det(W_i) - 2 n_i \ln n_i \right)$$

Marriott (1982) concluye que los criterios sugeridos son dignos de un estudio extenso.

1.4.7 Selección del número de grupos

En la mayoría de las aplicaciones de los métodos de optimización del análisis de conglomerados, el investigador tendrá la 'estimación' del número de grupos en los datos, y cuando esto sucede surgen una gran variedad de métodos. La mayoría de éstos son relativamente informales e involucran, esencialmente, gráficas del valor de los criterios de agrupamiento contra el número de grupos. Los cambios grandes de nivel en la gráfica normalmente son tomados como sugerentes de un número particular de grupos. Como sucede en los procedimientos similares, los dendrogramas, donde el juzgar el acercamiento en las gráficas puede ser muy subjetivo.

Un método sugerido por Calinski y Arabas (1974) es tomar el valor de *g* que corresponda al máximo valor de C, donde C está dado por :

$$C = \frac{\dfrac{\text{traza(B)}}{\text{g - 1}}}{\dfrac{\text{traza(W)}}{n - g}}$$

Por otro lado, Marriott (1971) sugiere como un posible procedimiento de evaluar el número de grupos el tomar el valor de g tal que minimice

$$g^2 \det(W) / \det(T)$$

1.4.8 Aplicaciones de los métodos de optimización

Existen muchas aplicaciones de los tipos de optimización de los métodos de agrupamiento y en seguida se darán algunos ejemplos de ello.

- Clasificación de pacientes psiquiátricos.

Las enfermedades de la mente son menos concretas que las del cuerpo, y una clasificación de las enfermedades psiquiátricas ha sido siempre difícil además de ser un tema controversial. Los métodos de agrupamiento se han usado frecuentemente en esfuerzos para refinar o incluso redefinir sistemas de diagnóstico psiquiátricos actuales. Los siguientes autores han hecho estudios al respecto: Zubin (1938), Lorr (1963), Everitt, Gourlay y Kennedy (1971). Los últimos autores buscaron minimizar la traza(W), en dos diferentes conjuntos de pacientes psiquiátricos, uno de Estados Unidos de América y el otro del Reino Unido. Cada conjunto consistía de 250 pacientes medidos en 45 estados mentales**.** Los grupos encontrados correspondían a las categorías de diagnóstico estándares como depresión, esquizofrenia y manía, aunque en cada caso se encontró un gran grupo 'mezclado' de pacientes con diagnósticos muy diferentes.

- Clasificación del dolor 'no específico' de espalda baja .

La ambigüedad de la presencia de diagnósticos disponibles para los dolores de espalda baja, según Henrich (1985), son 'perjudiciales a la moral del paciente e impide la investigación para el tratamiento óptimo y prevención'. Por consiguiente estos autores aplicaron varios métodos del análisis multivariado a un conjunto de 132 signos y síntomas coleccionados en 301 pacientes que padecen un dolor no-específico de espalda baja, en la búsqueda de clasificación útil. Entre estas técnicas estaban la minimización de la traza(W) y la minimización del det(W). Aunque los resultados de los diferentes métodos no fueron completamente consistentes, podrían identificarse cinco tipos de descripción del grupo estable.

1) Un grupo de pacientes que demuestran altos puntajes en los índices de dolor general.
2) Un grupo de pacientes con puntajes altos en los índices de dolor bilateral.
3) Un grupo con pacientes que muy frecuentemente su dolor cambia de lado
4) Un grupo de pacientes etiquetado por la ausencia de señales y síntomas.

5) Un grupo de pacientes que predominantemente muestran la presencia de cambios en el disco anterior, la ausencia de reflejos, la presencia de ciática y dolor ipselateral en corrección con una condición aguda.

- Juicio estético en pintores.

Un crítico del siglo XVII, Roger De Piles, expresó en términos cuantitativos una serie de juicios estéticos en 56 pintores, usando cuatro juicios conceptuales lógicos pero complejos. De Piles propuso para dividir 'las partes principales del arte en cuatro columnas referentes al ingenio: Composición, Diseño, Colorido y Expresión', y en cada dimensión los 56 pintores consiguieron un puntaje en una escala entre 0 y 20; donde la calificación de 20 fue reservada para la 'perfección soberana a la que ningún hombre ha llegado totalmente.'

Por otro lado, las escuelas a las que cada pintor pertenece son: a = Renacentista, b = Manierista, c = Seicento, d = Veneciana, e = Lombard, f = del Siglo XVI, g = del Siglo XVII y h = Francesa.

A continuación se presentan los datos de los 56 artistas:

	Pintor	Composición	Diseño	Color	Expresión	Escuela
1	Albani	14	14	10	6	e
2	Durer	8	10	10	8	f
3	Del Sarto	12	16	9	8	a
4	Barocci	14	15	6	10	c
5	Bassano	6	8	17	0	d
6	Del Piombo	8	13	16	7	a
7	Bellini	4	6	14	0	d
8	B	10	8	8	4	h
9	Le Brun		16	8	16	h
10	Veronese		10	16	3	d
11	The Carracci		17	13	13	e
12	Corregio	13	13	15	12	e
13	Volterra		15	5	8	b
14	Dipenbeck		10	14	6	g

15	Domenichino		17	9	17	e
16	Giogione	8	9	18	4	d
17	Guercino	18	10	10	4	e
18	Guido Reni	14	13	9	12	e
19	Holbein	9	10	16	13	f
20	Da Udine	10	8	16	3	a
21	J. Jordaens	10	8	16	6	g
22	L. Jordaens	13	12	9	6	c
23	Josepin	10	10	6	2	c
24	Romano	15	16	4	14	a
25	Lanfranco	14	13	10	5	e
26	Da Vinci	15	16	4	14	a
27	Van Leyden		6	6	4	f
28	Michelangelo		17	4	8	a
29	Caravaggio	6	6	16	0	e
30	Murillo	6	8	15	4	d
31	Venius	13	14	10	10	g
32	Vecchio	5	6	16	0	d
33	Giovane	12	9	14	6	d
34	Parmigiano	10	15	6	6	b
35	Penni	0	15	8	0	a
36	P a	15	16	7	6	a
37	Cortona	16	14	12	6	c
38	P	4	12	10	4	a
39	Polidore da Cara		17	8	15	a
40	Pordenone		14	17	5	d
41	Pourbus	4	15	6	6	f
42	P	15	17	6	15	h
43	Primaticcio	15	14	7	10	b
44	Raphael	17	18	12	18	a
45	Rembrandt		6	17	12	g
46	Rubens	18	13	17	17	g
47	Salviata	13	15	8	8	b
48	Le Sueur	15	15	4	15	h
49	Teniers	15	12	13	6	g
50	Testa	11	15	0	6	c
51	Tintoretto		14	16	4	d
52	Titian	12	15	18	6	d
53	Van Dyck		10	17	13	g
54	Vanius	15	15	12	13	c
55	T. Zuccaro	13	14	10	9	b
56	F	10	13	8	8	b

Con el propósito de organizar los datos se agrupó a los pintores utilizando el método de la minimización del det(W). Se calculó de dos a cuatro grupos y en cada caso, se consideraron cuatro configuraciones arbitrarias de partida. Los resultados se muestran a continuación; la letra hace referencia a la escuela a la cual pertenece el artista y el número corresponde a lista anterior.

- DOS GRUPOS

Grupo 1: n = 35
1(e), 3(a), 4(c), 9(h), 10(d), 11(e), 12(e), 13(b), 15(e), 17(e), 18(e), 22(c), 23(c), 24(a), 25(e), 26(a), 28(a), 31(g), 34(b), 36(a), 37(c), 39(a), 42(h), 43(b), 44(a), 46(g), 47(b), 48(h), 49(g), 50(c), 51(d), 52(d), 54(c), 55(b), 56(b)

Grupo 2: n = 21
2(f), 5(d), 6(a), 7(d), 8(h), 14(g), 16(d), 19(f), 20(a), 21(g), 27(f), 29(e), 30(d), 32(d), 33(d), 35(a), 38(a), 40(d), 41(f), 45(g), 53(g)

- TRES GRUPOS

Grupo 1: n = 13
5(d), 6(a), 7(d), 16(d), 20(a), 29(e), 30(d), 32(d), 35(a), 38(a), 40(d), 41(f), 52(d)

Grupo 2: n = 27
1(e), 2(f), 3(a), 4(c), 8(h), 10(d), 13(b), 14(g), 17(e), 21(g), 22(c), 23(c), 25(e), 27(f), 28(a), 31(g), 33(d), 34(b), 36(a), 37(c), 43(b), 47(b), 49(g), 50(c), 51(d), 56(b)

Grupo 3: n = 16
9(h), 11(e), 12(e), 15(e), 18(e), 19(f), 24(a), 26(a), 39(a), 42(h), 44(a), 45(g), 46(g), 48(h), 53(g), 54(c)

- CUATRO GRUPOS

Grupo 1: n = 16
2(f), 5(d), 7(d), 8(h), 14(g), 16(d), 19(f), 20(a), 21(g), 27(f), 29(e), 30(d), 32(d), 33(d), 45(g), 53(g)

Grupo 2: n = 15
9(h), 11(e), 12(e), 15(e), 18(e), 24(a), 26(a), 31(g), 39(a), 42(h), 44(a), 46(g), 48(h), 54(c), 56(b)

Grupo 3: n = 18
1(e), 3(a), 4(c), 10(d), 13(b), 17(e), 22(c), 23(c), 25(e), 34(b), 36(a), 37(c), 43(b), 47(b), 49(g), 50(c), 51(d), 55(b)

Grupo 4: n = 7
6(a), 28(a), 35(a), 38(a), 40(d), 41(f), 52(d)

Es difícil de especular sobre estos resultados sin ser un historiador de arte informado. Sin embargo, se puede concluir que la correspondencia entre grupos y la escuela de un artista es relativamente pequeña.

Capítulo 2

Análisis de Conglomerados para Datos Direccionales

El tópico del análisis de conglomerados para datos direccionales, o datos circulares, ha recibido poca mención en la literatura. Hasta el momento se ha recalcado que las estadísticas empleadas para datos lineales son inapropiadas para datos direccionales, puesto que hay que considerar el hecho que 1° y 359° están sólo 2° separados. Cuando se utiliza la estadística convencional para datos direccionales es necesario imponer una linealidad en los datos cortando el círculo para formar una línea que no sea cerrada. Sin embargo, esto es especialmente peligroso en el análisis de grupos, ya que el corte puede alterar grupos que se encuentren cerca o sobre la posición del corte.

Una alternativa es utilizar la medida de distancia circular dada por $\delta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$ como la medida de disimilaridad entre los dos valores muestreados θ_i y θ_j. La medida δ_{ij} toma como valor el más pequeño de los arcos entre θ_i y θ_j. Otra elección natural para medir la distancia en un círculo, como se ha menciona en el Apéndice B es $d_{ij} = 1 - \cos(\theta_i - \theta_j)$, la cual toma valores entre [0,1] donde el valor de uno indica que las observaciones están separadas lo más posible que es 180°.

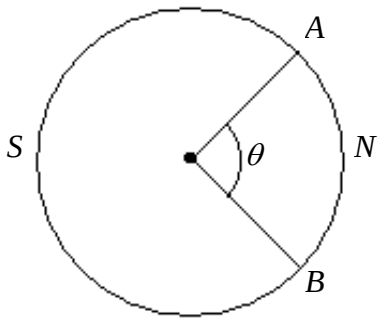
En este capítulo se presenta la estadística alternativa que explota la periodicidad inherente de los datos circulares. Al graficar las estadísticas de varios números posibles de grupos se puede fácilmente interpretar la gráfica, la cual se utilizará para determinar el número óptimo de grupos en los datos. Específicamente, el número óptimo de grupos en los datos será el que maximice el valor de la estadística sugerida.

2.1 Medidas de disimilaridad para datos direccionales

Las variables circulares necesitan, con base en lo anterior, métodos estadísticos y de medición distintos a los utilizados para datos lineales. Una solución fue sugerida por Ackerman (1997) y por Jammalamadaka[1], al definir el uso de la medida de distancia circular adecuada entre dos puntos a la longitud menor de los arcos formados entre los dos puntos en la circunferencia, es decir, que para cualquier pareja de ángulos α y β se tiene que:

$$\delta_{ij} = \delta(\theta_i, \theta_j) = \min(\theta_i - \theta_j, 2\pi - (\theta_i - \theta_j)) = \pi - | \pi - | \theta_i - \theta_j ||$$

Por ejemplo, en la gráfica que se presenta a continuación, la distancia entre A y B puede ser la longitud del arco ANB o la del arco ASB. Según δ_{ij}, la distancia sería la longitud de arco ANB. La distancia circular δ_{ij} toma valores entre $[0, \pi]$.



Otra elección definida por Jammalamadaka[2] para medir la distancia en un círculo, como se menciona en el Apéndice B, es:

$$d_{ij} = d(\theta_i, \theta_j) = 1 - \cos(\theta_i - \theta_j)$$

donde θ_i y θ_j representan los ángulos correspondientes a los puntos A y B.

[1] JAMMALAMADAKA, S. Roo, "Topics in Circular Statistics" Ed. World scientific, pág. 15.
[2] Ibídem. pág. 16.

Esta última distancia circular toma valores entre [0,1], donde el valor de uno indica que las observaciones están separadas lo más posible, que es 180°. Si θ es el ángulo ente los puntos A y B, es claro que la función de distancia d_{ij} es monótona creciente con respecto a θ, tomando el valor de 0 cuando $\theta = 0$ y crece hasta 2 si $\theta = \pi$.

Es importante, para poder utilizar la distancia como instrumento de decisión, determinar si las distancias denotadas como δ_{ij} y d_{ij} cumplen las propiedades de medida de disimilaridad. Se recuerda que una medida ρ entre *a* y *b* se dice de disimilaridad si:

- $\rho(a,b) \geq 0 \qquad \forall a,b$ (Positiva)
- $\rho(a,a) = 0 \qquad \forall a$ (Nulidad)
- $\rho(a,b) = \rho(b,a) \quad \forall a,b$ (Simetría)

Proposición 1: La distancia circular definida como $d_{ij} = d(\theta_i,\theta_j) = 1-\cos(\theta_i - \theta_j)$ es una medida de disimilaridad.

Demostración:

La positividad de la distancia se tiene ya que $\cos(\theta_i - \theta_j)$ está entre [-1,1] por tanto $d(\theta_i,\theta_j) \geq 0$. Además, se sabe que cos(0) = 1, por lo que se tiene que $1-\cos(0) = 0$, y para cualquiera θ_i se tiene que $d(\theta_i,\theta_i) = 1-\cos(\theta_i - \theta_i) = 0$. La simetría de la disimilaridad circular se tiene gracias a la paridad de la función coseno. Es decir, $\cos(\theta) = \cos(-\theta)$ implica directamente $d(\theta_i,\theta_j) = 1-\cos(\theta_i - \theta_j) = 1-\cos(\theta_j - \theta_i) = d(\theta_j,\theta_i)$.

■

***Proposición* 2.** La distancia circular definida como $\delta_{ij} = \delta(\theta_i, \theta_j) = \pi - |\pi - |\theta_i - \theta_j||$ es una medida de disimilaridad.

Demostración:

Dado que el máximo valor que toma la diferencia entre dos ángulos medidos en radianes está entre -2π y 2π se tiene que $|\pi - |\theta_i - \theta_j||$ toma valores entre $-\pi$ y π por tanto $\pi - |\pi - |\theta_i - \theta_j||$ tiene como rango $[0, \pi]$ con lo cual se tiene que $\delta(\theta_i, \theta_j) \geq 0$ para todo θ_i y θ_j. La nulidad es obvia y la simetría se obtiene del valor absoluto, ya que $|\theta_i - \theta_j| = |\theta_j - \theta_i|$

■

2.2 Evaluación de la presencia de grupos

Para una variable aleatoria circular θ, que toma valores en el círculo unitario $[0, 2\pi)$, una medida de localización y dispersión está dada por el primer momento trigonométrico

$$E\left[e^{i\theta}\right] = \rho e^{i\mu}$$

Esta cantidad define un vector desde el origen, cuya dirección está dada por μ y cuya longitud está dada por ρ. Los parámetros μ y ρ son llamados, respectivamente, la dirección media y la distancia media resultante de θ. Se tiene así que μ es una medida de localización de la distribución, mientras que ρ es una medida de dispersión. Se puede notar que la existencia del primer momento trigonométrico está garantizado, ya que éste también es la función característica evaluada en uno.

Para un conjunto de mediciones angulares $\theta_1, \theta_2, ..., \theta_n$, las estimaciones muestrales de μ y ρ son obtenidas mediante el tratamiento de datos como un vector unitario. La dirección media muestral es la dirección del vector resultante, cuya definición se encuentra en el Apéndice A, formado por las *n* observaciones, y la distancia media resultante muestral es obtenida al dividir la longitud del vector resultante entre el tamaño de la muestra. Más formalmente, se tiene

$$S = \sum_{i=1}^{n} \operatorname{sen}\theta_i \quad \text{y} \quad C = \sum_{i=1}^{n} \cos\theta_i$$

y la dirección media de la muestra está dada por

$$\bar{\theta} = \begin{cases} \arctan S/C & \text{si } C > 0 \\ \arctan S/C + \pi & \text{si } C < 0 \\ \pi/2 & \text{si } C = 0, S > 0 \\ -\pi/2 & \text{si } C = 0, S < 0 \end{cases}$$

y la distancia media resultante muestral es $\bar{r} = \dfrac{\sqrt{S^2 + C^2}}{n}$, donde $\bar{r} \in [0,1]$. Los valores de $\bar{r}$ cercanos a 0 indican una gran dispersión, mientras que los valores cercanos a 1 señalan que los datos se encuentran altamente concentrados.

Lo anterior se puede corroborar al observar los casos extremos. Si todas las observaciones son idénticas, entonces la longitud del vector resultante obtenido será de longitud igual a *n*, haciendo que la longitud de la media resultante sea igual a 1. Si los datos están igualmente dispersos por todo el círculo, entonces tanto *S* como *C* serán cero, dando una media resultante de la muestra de cero.

En lo sucesivo se usará la medida de dispersión $\bar{r}$ para identificar a los grupos en el conjunto de datos.

Los posibles grupos serán propuestos de acuerdo a la longitud de arco más grande, o espacios, entre las observaciones. Por ejemplo, los dos espacios más grandes serán considerados para determinar si hay dos grupos significativos de puntos. En general, los *k* grupos de puntos son obtenidos mediante los *k* espacios más grandes.

Para evaluar la significancia de los grupos se puede inspeccionar la dispersión de la observaciones en los grupos propuestos.

Se ha mencionado que ρ, distancia media resultante de θ, es una medida de dispersión y a continuación se presenta una proposición que prueba cómo calcularla.

Proposición 3. Sea θ uniformemente distribuida en un arco formado de *a* a *b*, donde $a,b \in [0, 2\pi)$. Entonces la distancia media resultante de θ es

$$\rho(a,b) = \frac{\text{sen}(\| a,b \| /2)}{\| a,b \| /2}$$

donde $\| a,b \| = b\text{-}a \pmod{2\pi}$.

Demostración:

Por facilidad de la notación se supondrá que $b > a$, de tal forma que $\| a,b \| = b\text{-}a$. Por definición, se tiene que $E\left[e^{i\theta}\right] = \rho e^{i\mu}$, y sustituyendo en este primer momento trigonométrico la relación de Euler, $e^{i\theta} = \cos\theta + i\text{sen}\theta$, se tiene

$$E[\cos\theta] = \rho\cos\mu \text{ y } E[\text{sen}\theta] = \rho\text{sen}\mu$$

Entonces, $\left(E[\cos\theta]\right)^2 = \rho^2(\cos\mu)^2$ y $\left(E[\operatorname{sen}\theta]\right)^2 = \rho^2(\operatorname{sen}\mu)^2$.

Si se suman las expresiones anteriores se obtiene

$$\left(E[\cos\theta]\right)^2 + \left(E[\operatorname{sen}\theta]\right)^2 = \rho^2(\cos\mu)^2 + \rho^2(\operatorname{sen}\mu)^2$$

Por lo que,

$$\rho^2 = \left(E[\cos\theta]\right)^2 + \left(E[\operatorname{sen}\theta]\right)^2$$

$$\rho = \left\{\left(E[\cos\theta]\right)^2 + \left(E[\operatorname{sen}\theta]\right)^2\right\}^{1/2}$$

Las esperanzas al cuadrado se calculan fácilmente si se piensa en ellas de la siguiente manera

$$\left(E[\cos\theta]\right)^2 = \left[\int_a^b \cos\phi \frac{d\theta}{b-a}\right]^2$$

$$= \left(\frac{1}{b-a}\right)^2 \left[\int_a^b \cos\phi d\theta\right]^2$$

$$= \left(\frac{1}{b-a}\right)^2 [\operatorname{sen} b - \operatorname{sen} a]^2$$

por otro lado

$$\left(E[\operatorname{sen}\theta]\right)^2 = \left[\int_a^b \operatorname{sen}\phi \frac{d\theta}{b-a}\right]^2$$

$$= \left(\frac{1}{b-a}\right)^2 \left[\int_a^b \operatorname{sen}\phi d\theta\right]^2$$

$$= \left(\frac{1}{b-a}\right)^2 [\cos a - \cos b]^2$$

sustituyendo el cálculo de las esperanzas al cuadrado en la expresión anterior se tiene

$$\rho = \left\{\left(\frac{1}{b-a}\right)^2 \left[(\operatorname{sen} b - \operatorname{sen} a)^2 + (\cos b - \cos a)^2\right]\right\}^{1/2}$$

desarrollando los cuadrados y recordando las siguientes identidades trigonométricas $\text{sen}^2 + \cos^2 = 1$ y $\cos(b-a) = \cos b \cos a + \text{sen} b \text{sen} a$ se tiene

$$\rho = \left(\frac{1}{b-a}\right)[2 - 2\cos(b-a)]^{1/2}$$

$$\rho = \left(\frac{1}{b-a}\right)\left[4\left(\frac{1}{2} - \frac{1}{2}\cos(b-a)\right)\right]^{1/2}$$

$$\rho = \left(\frac{2}{b-a}\right)\left[\frac{1}{2} - \frac{1}{2}\cos\left(2\frac{b-a}{2}\right)\right]^{1/2}$$

Por lo que finalmente se concluye que

$$\rho = \frac{\text{sen}((b-a)/2)}{(b\text{ - a})/2}$$

■

Como se ha mencionado anteriormente, los *k* grupos de puntos son determinados por los *k* espacios más grandes constituidos entre una observación y su inmediata anterior. Se denotará a los puntos medios de estos espacios por $m_1, m_2, ..., m_k$. Estos puntos medios serán utilizados para dividir al círculo, de tal manera que los *k* espacios estarán cada uno entre dos de estos puntos medios. Si los puntos del *i*-ésimo grupo están localizados en el arco formado de a_i a b_i, donde $a_i, b_i \in \{m_1, m_2, ..., m_k\}$, i = 1,2,...,*k*, y si los puntos están uniformemente distribuidos en este arco, su distancia media resultante es $p_i = \rho(a_i, b_i)$, como se definió en la proposición anterior.

Se denotará a la distancia media resultante de los *k* grupos por $\overline{r_1}, \overline{r_2}, ..., \overline{r_k}$. Al restarle p_i a $\overline{r_i}$ se representa la concentración de los puntos en el *i*-ésimo grupo y da una medida de qué tan significativo es el grupo. Para los *k* grupos se suma, sumando estos valores sobre todos los grupos se tiene la siguiente estadística:

$$S_k = \sum_{i=1}^{k} \left(\overline{r_i} - p_i \right)$$

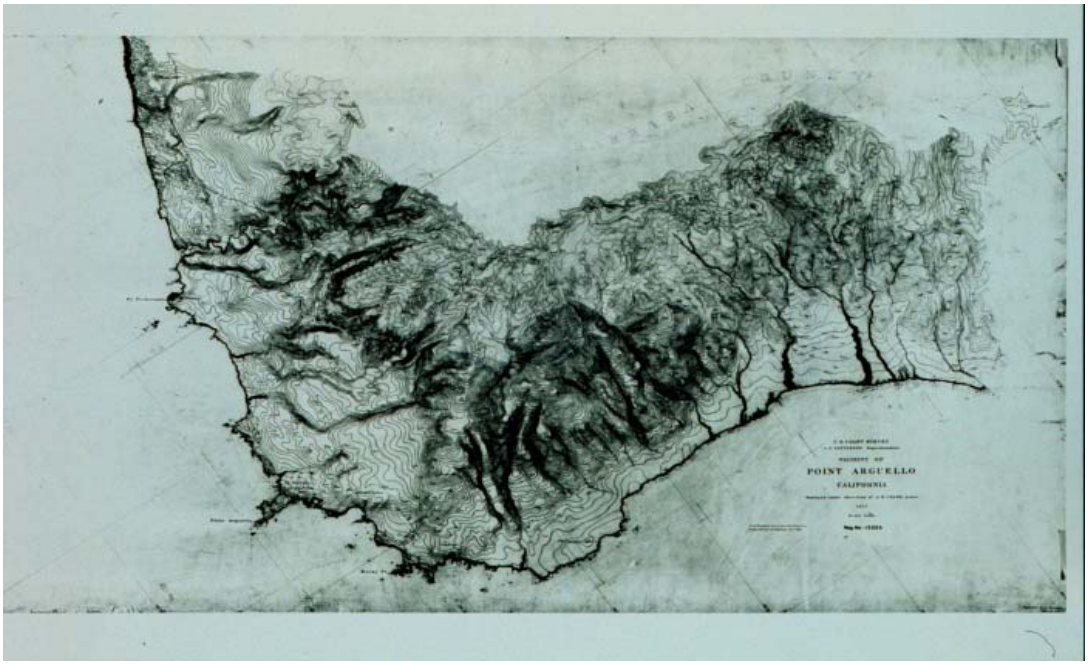
Para un número propuesto de grupos, *k*, S_k mide la concentración de los *k* grupos. Un número relativamente grande de S_k indica un alto grado de agrupamiento entre los grupos propuestos.

Es posible que S_k tome valores negativos, no obstante, esto sólo ocurrirá cuando los datos estén completamente distribuidos en forma equitativa en el círculo, y el último de los grupos tenga una distancia media resultante menor que la distancia media resultante de una distribución uniforme en un arco que contiene a ese grupo. Sin embargo, esto está garantizado porque el máximo de S_k sobre *k* es no negativo.

Graficando S_k contra k, es posible identificar la significancia de los incrementos sucesivos del número de grupos. El número óptimo de grupos es el valor k_0 que maximiza S_k, ya que k_0 grupos producen los grupos de puntos con mayor concentración, relativo a los datos uniformemente distribuidos.

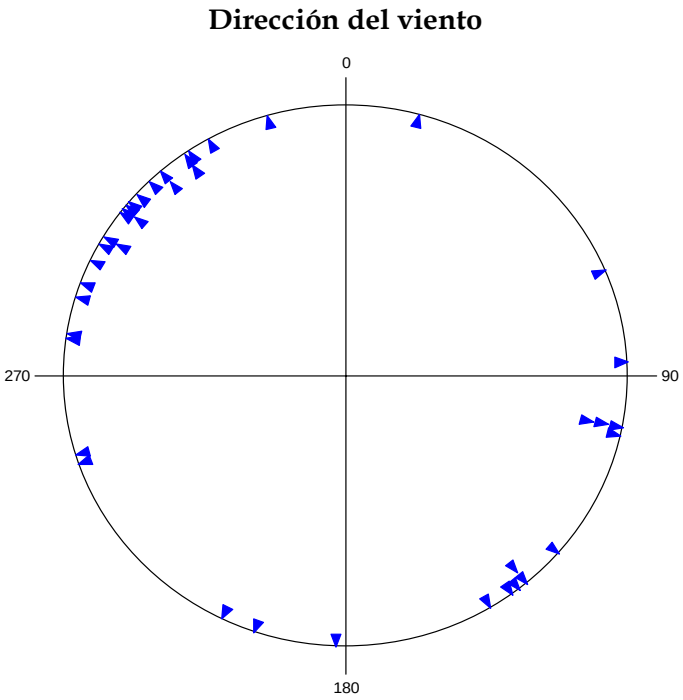
2.2.1 Ejemplo

El siguiente conjunto de datos consiste en 40 observaciones de la dirección del viento, obtenidas de la estación climatológica National Oceanic and Atmospheric Administration (NOAA), en Point Arguello en la costa de California.

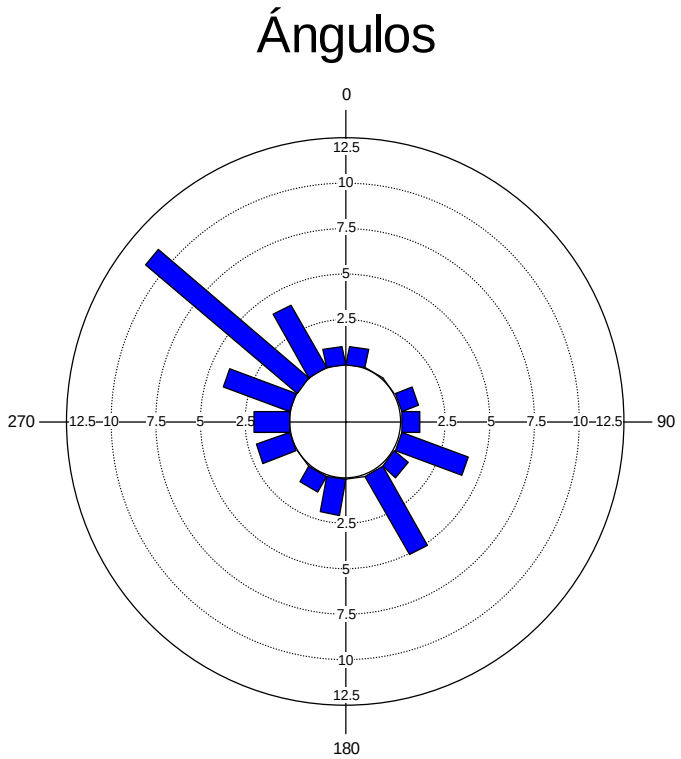


Direcciones del viento en grados			
15	142	287	310
67	144	290	312
87	149	295	316
101	182	299	319
101	199	301	319
101	206	301	325
103	251	307	325
131	253	308	326
140	278	308	331
140	279	309	344

Gráficamente se ven de la siguiente manera,



Y su histograma es el siguiente,

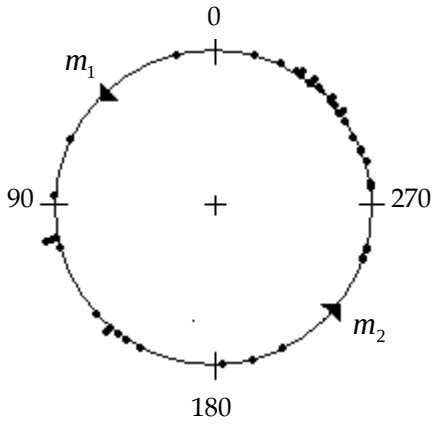


La distancia media resultante muestral, $\bar{r}$, de las direcciones del viento es 0.27349.

Nótese que cuando se calcula S_1, es decir, cuando se supone que sólo hay un grupo, no hay particiones en el círculo. Sin embargo, no se debe omitir su cálculo, y para ello se calculará $\bar{r}_1$ y p_1.

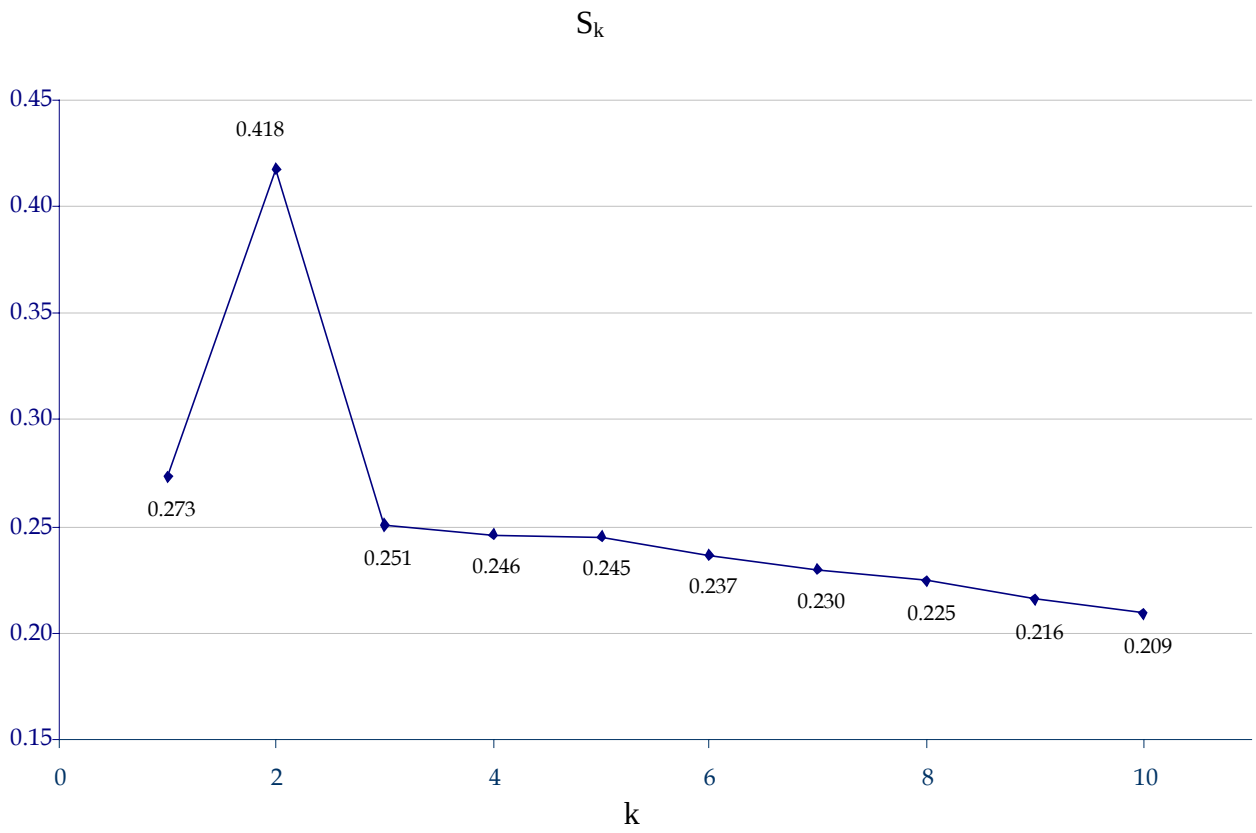
Con base en la ecuación $\rho(a,b) = \dfrac{\text{sen}(\| a,b \| /2)}{\| a,b \| /2}$ se tiene que $p_1 = \rho(0,2\pi) = 0$ y por lo tanto $S_1 = \bar{r}_1 = \bar{r}$, la distancia media resultante del total de la muestra.

En la siguiente figura se muestra los puntos medios de los espacios más grandes, etiquetados como m_1 y m_2 respectivamente



En esta figura podría parecer que existen 2 grupos significativos en el conjunto de datos: los puntos en el arco formado de m_1 a m_2, y los otros permanecen en el arco complementario. Para evaluar el agrupamiento se iniciará por calcular la distancia media resultante conjunta de la muestra, $\bar{r}$ = 0.27349.

Para determinar el número óptimo de grupos en el conjunto de datos, se procede a calcular S_k para $k = \overline{1,10}$, y graficar S_k contra k.



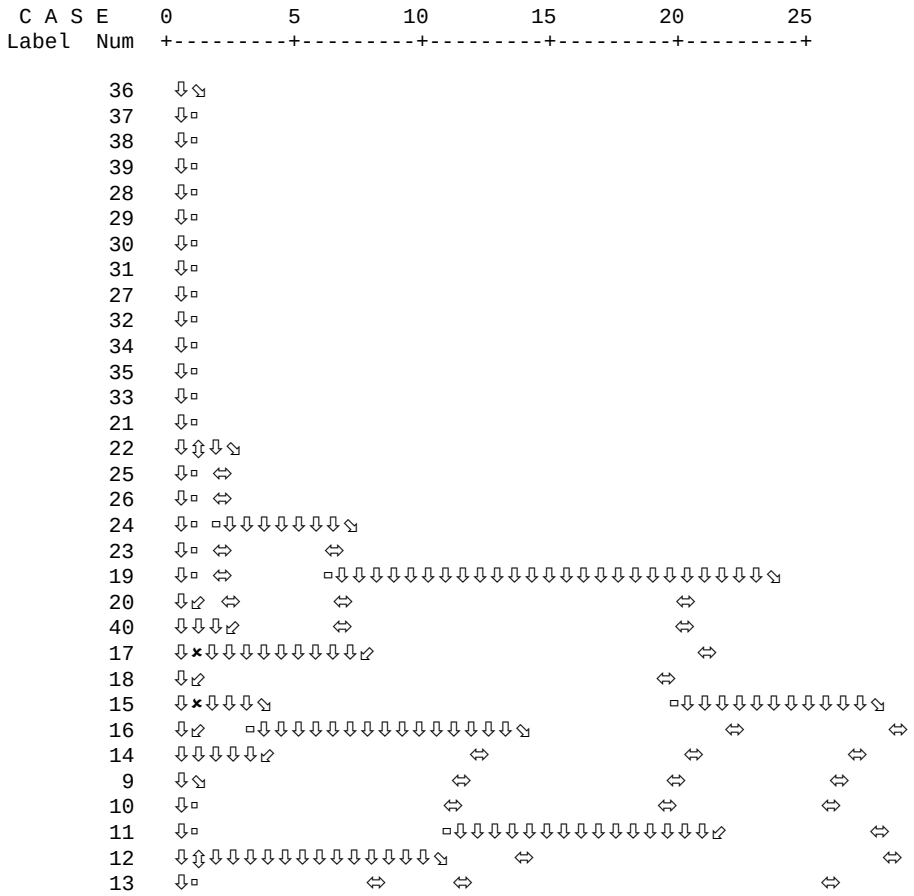
En la gráfica anterior se puede apreciar que S_k es maximizada por k =2, indicando que dos grupos son los grupos más significativos. Es decir, la distribución de estos dos grupos, en sus respectivos arcos, son los más concentrados significativamente en relación con las distribuciones uniformes en los mismos arcos.

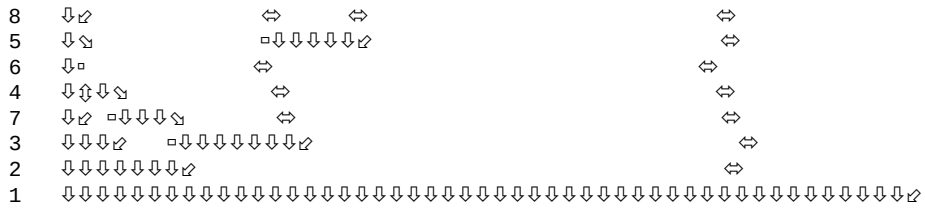
En la misma gráfica se puede observar una disminución en el valor de S_3 que indica que el tercer grupo formado por las tres observaciones justo después de los 180° no son exactamente tan ajenas de los otros dos grupos para garantizar la partición. La rápida disminución en los sucesivos valores de S_k muestra que los grupos subsecuentes son aún menos justificados.

Con base en que lo óptimo son dos grupos, dichos grupos serían los siguientes:

Grupos	Observaciones				
1	67	87	101	101	101
	103	131	140	140	142
	144	149	182	199	206
2	251	253	278	279	287
	290	295	299	301	301
	307	308	308	309	310
	312	316	319	319	325
	325	326	331	344	15

A continuación se muestra el respectivo dendrograma utilizando el algoritmo de la liga simple, cuya matriz de disimilaridad es construida usando la distancia del coseno.





El dendrograma indica la presencia de los mismos dos grupos elegidos anteriormente con base en la estadística $S_k = \sum_{i=1}^{k} \left(\bar{r}_i - p_i \right)$. Es decir, en las direcciones del viento en el Punto Arguello se pueden hacer dos grupos, uno formado por aquellas con orientación noreste y otro con vientos en dirección suroeste.

Sin embargo, el hecho de que otro agrupamiento sea significativo es un poco subjetivo; este caso sucede a menudo. Con la evidencia en la gráfica de S_k contra k y la estadística S_k, se puede tener más confianza en la elección de dos grupos para el conjunto de datos.

Capítulo 3

Aplicación

ORIENTACIÓN DE LAS TORTUGAS

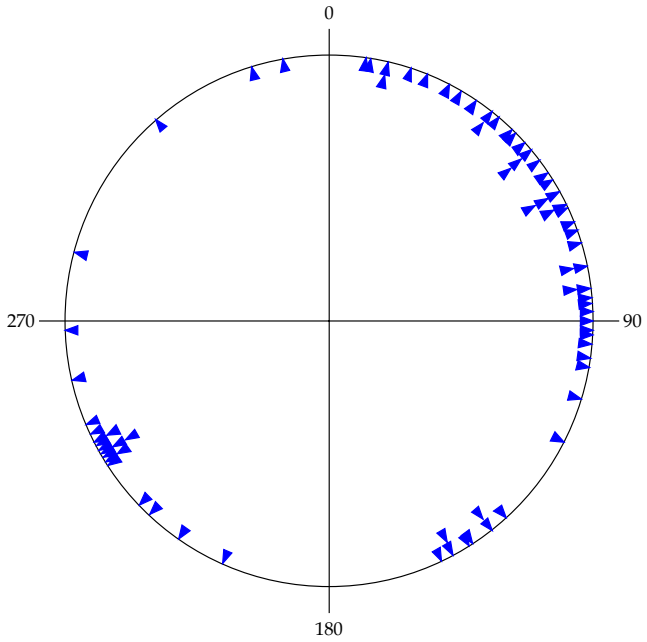
Como se ha mencionado, los datos direccionales son utilizados con frecuencia en diversas ciencias, entre las que destaca la biología, ya que el uso de estos datos puede verse reflejado en diversas ramas de ésta. Algunos ejemplos donde se manejan los datos circulares en la biología son la orientación de los animales, la migración y los ritmos biológicos, pues las variables de interés en este último caso se miden en tiempo.

Por lo anterior, y a manera de ejemplificar la aplicación de la estadística sugerida en este trabajo de tesis, se presenta el siguiente conjunto de datos referentes a las direcciones que toman 76 tortugas después de desovar; los ángulos se consideran en dirección dextrógira.

Dirección (en grados) de las tortugas después de desovar.							
8	9	13	13	18	22	27	30
34	38	38	40	44	45	48	50
50	50	53	56	58	61	61	61
64	64	65	68	70	73	78	78
83	83	85	86	88	90	92	93
95	98	100	107	117	138	142	142
147	148	152	152	155	204	215	223
226	237	238	238	239	240	240	240
241	242	243	243	245	247	257	268
285	319	343	350				

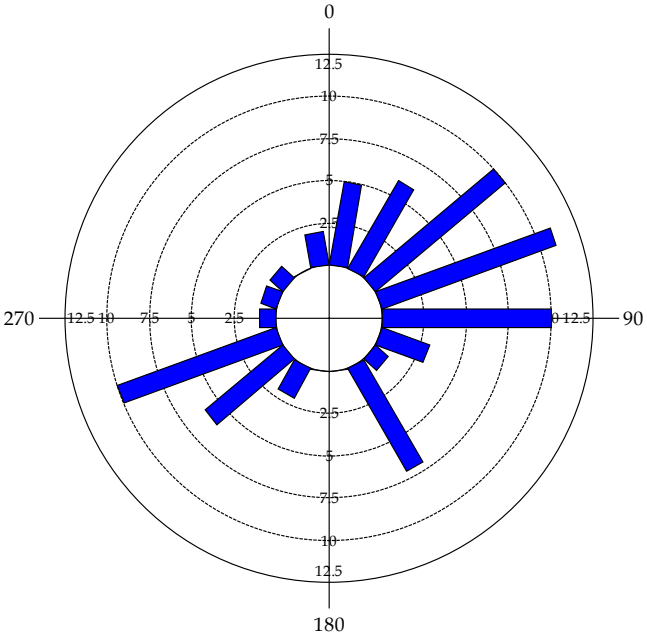
Gráficamente se ven de la siguiente manera,

Orientación de las tortugas



Y su histograma circular es el siguiente,

Histograma Circular



Utilizando la distancia $\delta_{ij} = \delta(\theta_i, \theta_j) = \pi - |\pi - |\theta_i - \theta_j||$, se construye la matriz de distancias de los datos,

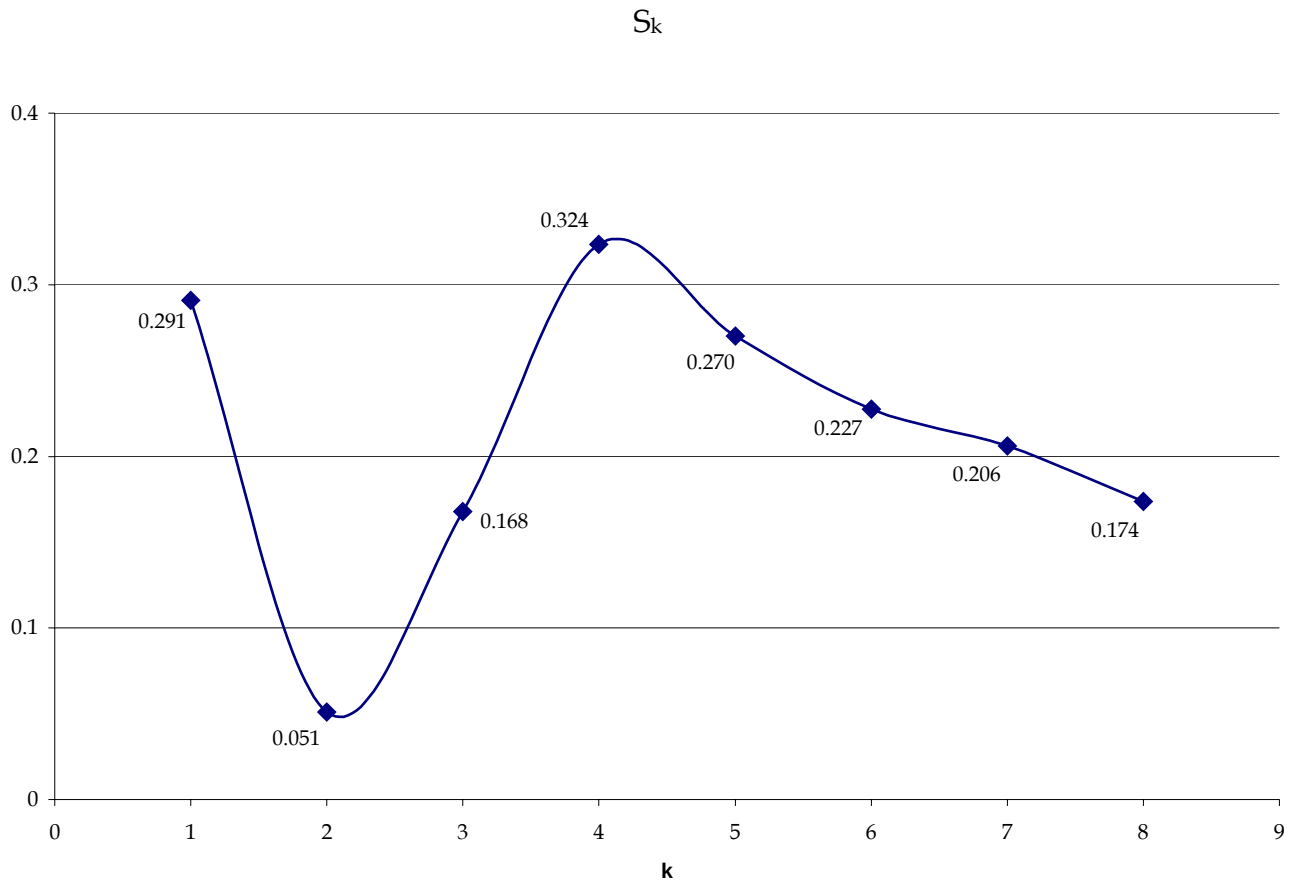
·	8	9	13	18	22	27	30	34	38	40	44	45	48	50	53	56	58	61	64	65	68	70	73	78	83	85	86	88	90	92	93	95	98	100	107	117	138	142	147	148	152	155	204	215	223	226	237	238	239	240	241	242	243	245	247	257	268	285	319	343	350
8	0																																																															
9	1	0																																																																
13	5	4	0																																																														
18	10	9	5	0																																																													
22	14	13	9	4	0																																																													
27	19	18	14	9	5	0																																																												
30	22	21	17	12	8	3	0																																																											
34	26	25	21	16	12	7	4	0																																																										
38	30	29	25	20	16	11	8	4	0																																																								
40	32	31	27	22	18	13	10	6	2	0																																																							
44	36	35	31	26	22	17	14	10	6	4	0																																																					
45	37	36	32	27	23	18	15	11	7	5	1	0																																																				
48	40	39	35	30	26	21	18	14	10	8	4	3	0																																																		
50	42	41	37	32	28	23	20	16	12	10	6	5	2	0																																																
53	45	44	40	35	31	26	23	19	15	13	9	8	5	3	0																																														
56	48	47	43	38	34	29	26	22	18	16	12	11	8	6	3	0																																													
58	50	49	45	40	36	31	28	24	20	18	14	13	10	8	5	2	0																																												
61	53	52	48	43	39	34	31	27	23	21	17	16	13	11	8	5	3	0																																											
64	56	55	51	46	42	37	34	30	26	24	20	19	16	14	11	8	6	3	0																																										
65	57	56	52	47	43	38	35	31	27	25	21	20	17	15	12	9	7	4	1	0																																									
68	60	59	55	50	46	41	38	34	30	28	24	23	20	18	15	12	10	7	4	3	0																																									
70	62	61	57	52	48	43	40	36	32	30	26	25	22	20	17	14	12	9	6	5	2	0																																								
73	65	64	60	55	51	46	43	39	35	33	29	28	25	23	20	17	15	12	9	8	5	3	0																																								
78	70	69	65	60	56	51	48	44	40	38	34	33	30	28	25	22	20	17	14	13	10	8	5	0																																							
83	75	74	70	65	61	56	53	49	45	43	39	38	35	33	30	27	25	22	19	18	15	13	10	5	0																																						
85	77	76	72	67	63	58	55	51	47	45	41	40	37	35	32	29	27	24	21	20	17	15	12	7	2	0																																					
86	78	77	73	68	64	59	56	52	48	46	42	41	38	36	33	30	28	25	22	21	18	16	13	8	3	1	0																																					
88	80	79	75	70	66	61	58	54	50	48	44	43	40	38	35	32	30	27	24	23	20	18	15	10	5	3	2	0																																					
90	82	81	77	72	68	63	60	56	52	50	46	45	42	40	37	34	32	29	26	25	22	20	17	12	7	5	4	2	0																																					
92	84	83	79	74	70	65	62	58	54	52	48	47	44	42	39	36	34	31	28	27	24	22	19	14	9	7	6	4	2	0																																					
93	85	84	80	75	71	66	63	59	55	53	49	48	45	43	40	37	35	32	29	28	25	23	20	15	10	8	7	5	3	1	0																																					
95	87	86	82	77	73	68	65	61	57	55	51	50	47	45	42	39	37	34	31	30	27	25	22	17	12	10	9	7	5	3	2	0																																					
98	90	89	85	80	76	71	68	64	60	58	54	53	50	48	45	42	40	37	34	33	30	28	25	20	15	13	12	10	8	6	5	3	0																																					
100	92	91	87	82	78	73	70	66	62	60	56	55	52	50	47	44	42	39	36	35	32	30	27	22	17	15	14	12	10	8	7	5	2	0																																					
107	99	98	94	89	85	80	77	73	69	67	63	62	59	57	54	51	49	46	43	42	39	37	34	29	24	22	21	19	17	15	14	12	9	7	0																																					
117	109	108	104	99	95	90	87	83	79	77	73	72	69	67	64	61	59	56	53	52	49	47	44	39	34	32	31	29	27	25	24	22	19	17	10	0																																					
138	130	129	125	120	116	111	108	104	100	98	94	93	90	88	85	82	80	77	74	73	70	68	65	60	55	53	52	50	48	46	45	43	40	38	31	21	0																																					
142	134	133	129	124	120	115	112	108	104	102	98	97	94	92	89	86	84	81	78	77	74	72	69	64	59	57	56	54	52	50	49	47	44	42	35	25	4	0																																					
147	139	138	134	129	125	120	117	113	109	107	103	102	99	97	94	91	89	86	83	82	79	77	74	69	64	62	61	59	57	55	54	52	49	47	40	30	9	5	0																																					
148	140	139	135	130	126	121	118	114	110	108	104	103	100	98	95	92	90	87	84	83	80	78	75	70	65	63	62	60	58	56	55	53	50	48	41	31	10	6	1	0																																					
152	144	143	139	134	130	125	122	118	114	112	108	107	104	102	99	96	94	91	88	87	84	82	79	74	69	67	66	64	62	60	59	57	54	52	45	35	14	10	5	4	0																																					
155	147	146	142	137	133	128	125	121	117	115	111	110	107	105	102	99	97	94	91	90	87	85	82	77	72	70	69	67	65	63	62	60	57	55	48	38	17	13	8	7	3	0																																					
204	164	165	169	174	178	177	174	170	166	164	160	159	156	154	151	148	146	143	140	139	136	134	131	126	121	119	118	116	114	112	111	109	106	104	97	87	66	62	57	56	52	49	0																																					
215	153	154	158	163	167	172	175	179	177	175	171	170	167	165	162	159	157	154	151	150	147	145	142	137	132	130	129	127	125	123	122	120	117	115	108	98	77	73	68	67	63	60	11	0																																					
223	145	146	150	155	159	164	167	171	175	177	179	178	175	173	170	167	165	162	159	158	155	153	150	145	140	138	137	135	133	131	130	128	125	123	116	106	85	81	76	75	71	68	19	8	0																																					
226	142	143	147	152	156	161	164	168	172	174	178	179	178	176	173	170	168	165	162	161	158	156	153	148	143	141	140	138	136	134	133	131	128	126	119	109	88	84	79	78	74	71	22	11	3	0																																					
237	131	132	136	141	145	150	153	157	161	163	167	168	171	173	176	179	179	176	173	172	169	167	164	159	154	152	151	149	147	145	144	142	139	137	130	120	99	95	90	89	85	82	33	22	14	11	0																																					
238	130	131	135	140	144	149	152	156	160	162	166	167	170	172	175	178	180	177	174	173	170	168	165	160	155	153	152	150	148	146	145	143	140	138	131	121	100	96	91	90	86	83	34	23	15	12	1	0																																					
239	129	130	134	139	143	148	151	155	159	161	165	166	169	171	174	177	179	178	175	174	171	169	166	161	156	154	153	151	149	147	146	144	141	139	132	122	101	97	92	91	87	84	35	24	16	13	2	1	0																																					
240	128	129	133	138	142	147	150	154	158	160	164	165	168	170	173	176	178	179	176	175	172	170	167	162	157	155	154	152	150	148	147	145	142	140	133	123	102	98	93	92	88	85	36	25	17	14	3	2	1	0																																					
241	127	128	132	137	141	146	149	153	157	159	163	164	167	169	172	175	177	180	177	176	173	171	168	163	158	156	155	153	151	149	148	146	143	141	134	124	103	99	94	93	89	86	37	26	18	15	4	3	2	1	0																																					
242	126	127	131	136	140	145	148	152	156	158	162	163	166	168	171	174	176	179	178	177	174	172	169	164	159	157	156	154	152	150	149	147	144	142	135	125	104	100	95	94	90	87	38	27	19	16	5	4	3	2	1	0																																				
243	125	126	130	135	139	144	147	151	155	157	161	162	165	167	170	173	175	178	179	178	175	173	170	165	160	158	157	155	153	151	150	148	145	143	136	126	105	101	96	95	91	88	39	28	20	17	6	5	4	3	2	1	0																																		
245	123	124	128	133	137	142	145	149	153	155	159	160	163	165	168	171	173	176	179	180	177	175	172	167	162	160	159	157	155	153	152	150	147	145	138	128	107	103	98	97	93	90	41	30	22	19	8	7	6	5	4	3	2	0																																
247	121	122	126	131	135	140	143	147	151	153	157	158	161	163	166	169	171	174	177	178	179	177	174	169	164	162	161	159	157	155	154	152	149	147	140	130	109	105	100	99	95	92	43	32	24	21	10	9	8	7	6	5	4	2	0																																
257	111	112	116	121	125	130	133	137	141	143	147	148	151	153	156	159	161	164	167	168	171	173	176	179	174	172	171	169	167	165	164	162	159	157	150	140	119	115	110	109	105	102	53	42	34	31	20	19	18	17	16	15	14	12	10	0																																
268	100	101	105	110	114	119	122	126	130	132	136	137	140	142	145	148	150	153	156	157	160	162	165	170	175	177	178	180	178	176	175	173	170	168	161	151	130	126	121	120	116	113	64	53	45	42	31	30	29	28	27	26	25	23	21	11	0																																
285	83	84	88	93	97	102	105	109	113	115	119	120	123	125	128	131	133	136	139	140	143	145	148	153	158	160	161	163	165	167	168	170	173	175	178	168	147	143	138	137	133	130	81	70	62	59	48	47	46	45	44	43	42	40	38	28	17	0																																
319	49	50	54	59	63	68	71	75	79	81	85	86	89	91	94	97	99	102	105	106	109	111	114	119	124	126	127	129	131	133	134	136	139	141	148	158	179	177	172	171	167	164	115	104	96	93	82	81	80	79	78	77	76	74	72	62	51	34	0																																
343	25	26	30	35	39	44	47	51	55	57	61	62	65	67	70	73	75	78	81	82	85	87	90	95	100	102	103	105	107	109	110	112	115	117	124	134	155	159	164	165	169	172	139	128	120	117	106	105	104	103	102	101	100	98	96	86	75	58	24	0																																
350	18	19	23	28	32	37	40	44	48	50	54	55	58	60	63	66	68	71	74	75	78	80	83	88	93	95	96	98	100	102	103	105	108	110	117	127	148	152	157	158	162	165	146	135	127	124	113	112	111	110	109	108	107	105	103	93	82	65	31	7	0																																

Una vez calculadas las distancias entre los datos, se identifican los espacios más grandes que se forman entre parejas de datos subsecuentes, es decir, se determinan los espacios más grandes constituidos entre una observación y su inmediata anterior. Estos espacios se pueden precisar a partir de un análisis visual de la gráfica de los datos o bien, observando la segunda y última diagonales de la matriz de distancias. Los k espacios más grandes se presentan a continuación, para $k = \overline{2,8}$, ya que cuando $k = 1$ no hay particiones en el círculo.

k	°		Distancia	Puntos medios
2	155	204	49	179.5
	285	319	34	302
3	155	204	49	179.5
	285	319	34	302
	319	343	24	331
4	155	204	49	179.5
	285	319	34	302
	319	343	24	331
	117	138	21	127.5
5	155	204	49	179.5
	285	319	34	302
	319	343	24	331
	117	138	21	127.5
	350	8	18	359
6	155	204	49	179.5
	285	319	34	302
	319	343	24	331
	117	138	21	127.5
	350	8	18	359
	268	285	17	276.5
7	155	204	49	179.5
	285	319	34	302
	319	343	24	331
	117	138	21	127.5
	350	8	18	359
	268	285	17	276.5
	257	268	11	262.5
8	155	204	49	179.5
	285	319	34	302
	319	343	24	331
	117	138	21	127.5
	350	8	18	359
	268	285	17	276.5
	257	268	11	262.5
	226	237	11	231.5

Donde la primera columna, k, precisa la cantidad de arcos o espacios; la segunda y tercera comprenden los ángulos que forman dichos arcos; en la cuarta columna se indica la distancia entre los ángulos que forman los k espacios; y en la última columna se calculó los puntos medios de los mencionados espacios.

Con base en los datos anteriores, y con el fin de determinar el número óptimo de grupos en el conjunto de datos, se calcula S_k para k = $\overline{1,8}$, y grafica S_k contra k.

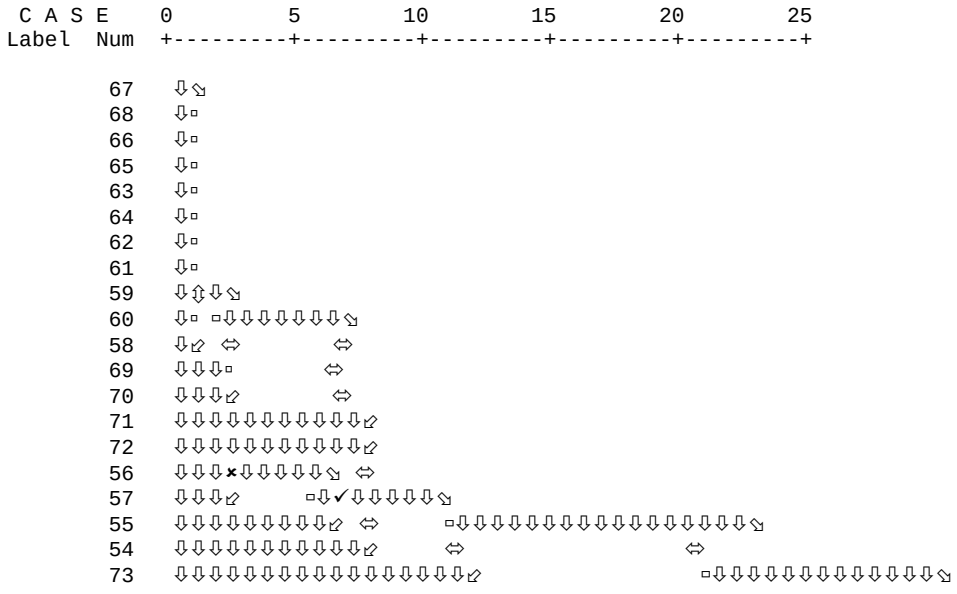


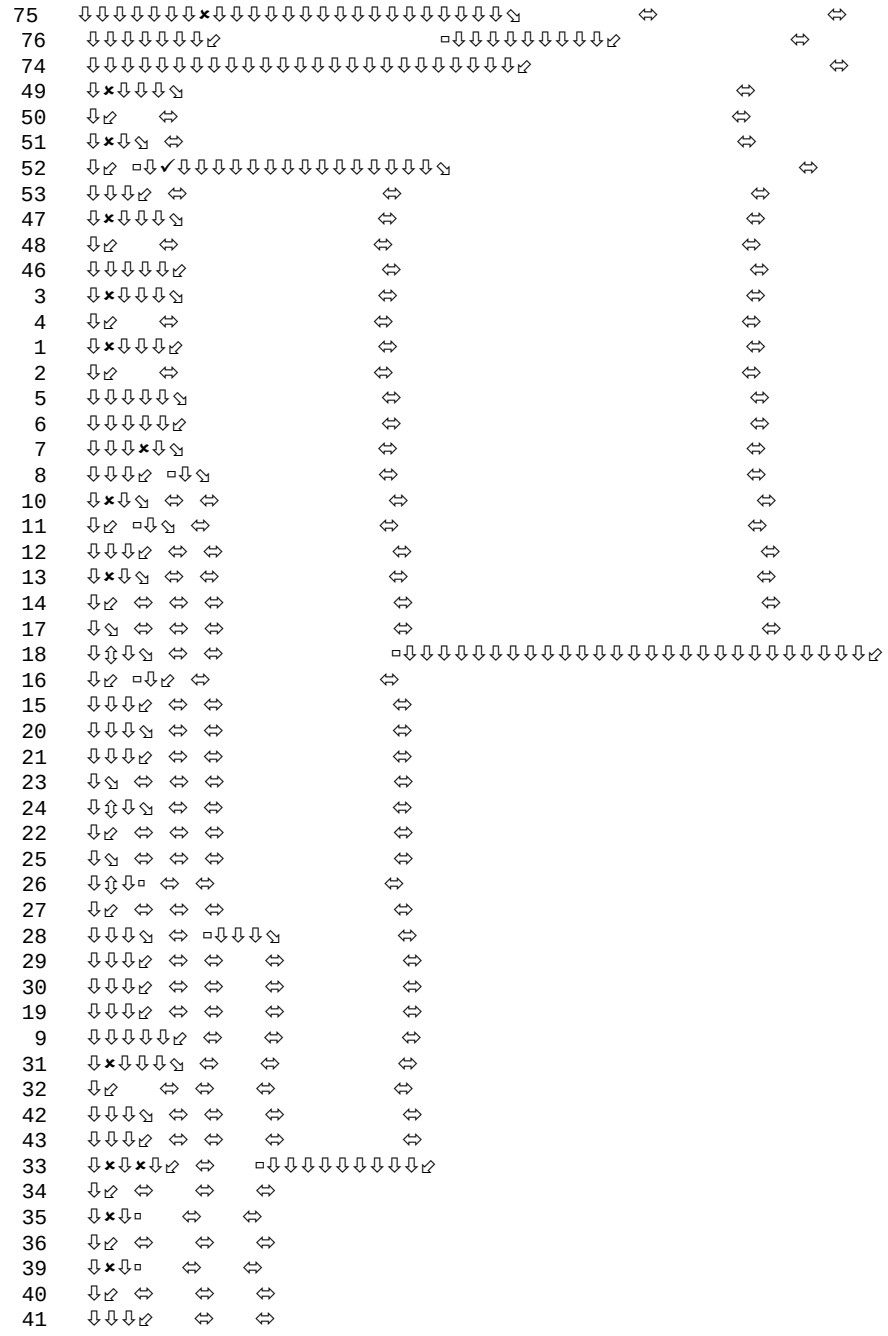
De la gráfica se interpreta que S_k alcanza su máximo en k = 4, indicando que cuatro grupos son los grupos más significativos. Es decir, la distribución de estos cuatro grupos en sus respectivos arcos son los más concentrados significativamente en relación con las distribuciones uniformes en los mismos arcos.

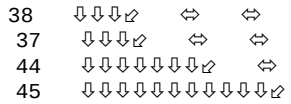
De acuerdo a la estadística S_k, los cuatro grupos óptimos en los datos son:

Grupos	Observaciones								$\bar{r}$	ρ
1	343	350	8	9	13	13	18	22	0.862054387	0.716871227
	27	30	34	38	38	40	44	45		
	48	50	50	50	53	56	58	61		
	61	61	64	64	65	68	70	73		
	78	78	83	83	85	86	88	90		
	92	93	95	98	100	107	117			
2	138	142	142	147	148	152	152	155	0.995322862	0.966031407
3	204	215	223	226	237	238	238	239	0.958568707	0.820126415
	240	240	240	241	242	243	243	245		
	247	257	268	285						
4	343	350							1	0.989359828

A continuación se muestra el respectivo dendrograma, usando la distancia del coseno y utilizando el algoritmo de la liga simple.







En el dendrograma también se halla la presencia de cuatro grupos. Es decir, en cuanto a la orientación que eligen las tortugas después de desovar se pueden detectar cuatro grupos, uno formado por aquellas que optan por la orientación norte y noreste, el segundo formado por aquellas que prefieren la orientación sureste, el tercero constituido por las que se dirigen con orientación suroeste y un último formado por las tortugas que se orientan al noroeste.

Conclusiones

- La habilidad de los métodos de agrupamiento es que detectan la no existencia de grupos bien establecidos. Si una clasificación no existe, un problema más fuerte es que los datos pueden aceptar más de una clasificación y ya dependerá del propósito de los investigadores.

- Los métodos de agrupamiento pueden ser aplicados al mismo conjunto de datos y producir estructuras substancialmente diferentes. Esto es debido a que la elección del método de agrupamiento implica imponer una estructura a la población.

- El método jerárquico de agrupamiento de la liga simple es aquél que mejor cumple con todas las condiciones matemáticas establecidas por Jardine y Sibson.

- Los dendrogramas son la herramienta gráfica más valiosa para la formación de grupos. Sin embargo, los dendrogramas no detectan el número óptimo de grupos en los datos y esto pudiese ser un inconveniente en el objetivo de algunas investigaciones.

- Cuando se desea hallar el número óptimo de grupos en los datos, lo útil es la maximización o minimización de algún método numérico que produzca una partición de los individuos u objetos. No obstante, y a diferencia de los dendrogramas, estos no necesariamente forman clasificaciones jerárquicas de los datos.

- Resulta especialmente peligroso darle a las variables circulares el mismo trato que a las lineales, en particular en la formación de grupos. Es por ello que se deben considerar medidas de similaridad (y disimilaridad) y métodos estadísticos diferentes y específicos para variables direccionales.

- La estadística S_k, basada en las diferencias de las distancias medias resultantes muestral y poblacional de los k grupos, permite desarrollar un método jerárquico de agrupamiento divisivo para variables circulares. Además dicha estadística identifica cuál es el número óptimo de grupos en los mismos.

- Se debe desarrollar más la investigación en el análisis de conglomerados para datos direccionales, pues la estadística presentada en este trabajo de tesis sólo considera la formación de grupos en aquellos individuos u objetos en los que se ha medido sólo una característica de interés, debido a que se carece de alguna otra estadística que haga lo propio en casos multivariados.

APÉNDICE A

Conceptos Fundamentales de Estadística Circular

A.1 Medidas de localización

Las direcciones son medidas en ángulos en un rango de 0° a 360° o, equivalentemente, de 0 a $2p$ radianes. La dirección cero (norte en algunos casos, en otros el eje de las equis) es completamente arbitrario. La medición de la dirección de los ángulos es cíclica, y por tanto, a este tipo de variable se le llama *variable circular o direccional.* Dichas variables son totalmente diferentes de las otras cantidades como la longitud, peso, temperatura, voltaje, a las cuales se les llamará *variables lineales*.

Las variables circulares también se encuentran en experimentos que miden tiempo, por ejemplo, un periodo de 24 horas corresponde a una vuelta completa de 360 grados, es decir, se puede comparar una hora con un ángulo de 15 grados y medio día con 180 grados. Situación similar ocurre con un mes, un año o cualquier otro periodo de un evento cíclico pues puede ser representado en un círculo.

El análisis de variables que son medidas en ángulos tiene muchas aplicaciones en áreas como la biología, geología, geografía, meteorología, astronomía, física, economía y medicina.

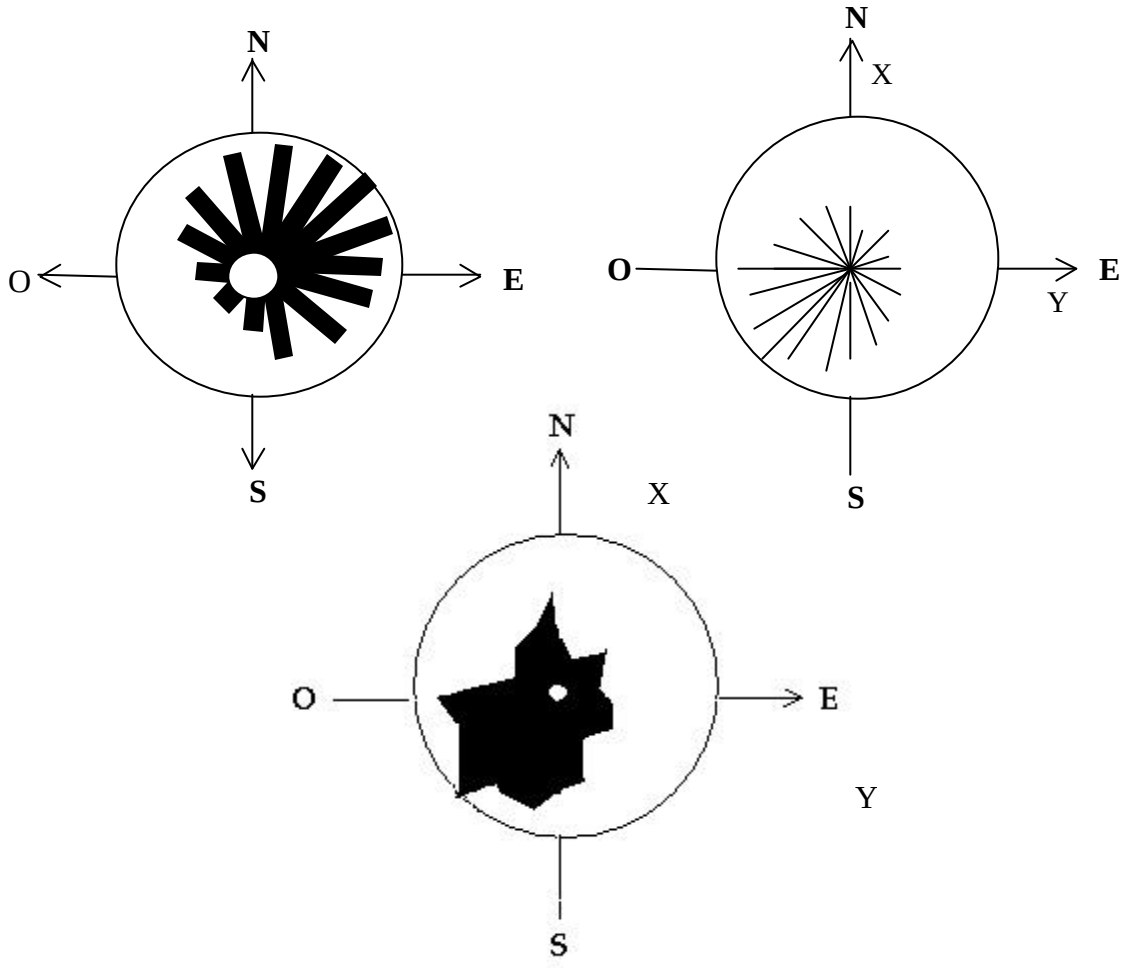
Por otro lado, la suma o diferencia de los ángulos podría exceder el intervalo de 0° a 360° por lo que se tiene que reducir a módulo 360°. Cabe notar que si α y β son variables circulares, también lo es α-β; así como también que en el caso de no importar si la desviación es en el sentido de las agujas del reloj (sentido dextrógiro) o en sentido contrario (sentido levógiro) se elegirá la *distancia angular* $|\alpha,\beta|$. Nótese que $|\alpha,\beta|$ no es periódica y es, por consiguiente, una variable lineal, no una circular. Por consecuencia, sólo los métodos lineales debieran ser aplicados a distancias angulares.

A.1.1 Presentación gráfica

Las direcciones pueden ser representadas gráficamente en una circunferencia por semilíneas que comienzan en el origen, *O*, a estas presentaciones se les llama *diagramas de dispersión.*

Para una gran cantidad de datos puede ser necesario ordenar las direcciones observadas en grupos. En cuyo caso es conveniente graficar un *histograma circular*, como se muestra a continuación. Como en un histograma lineal, las barras deben ser rectangulares y representar adecuadamente las frecuencias.

También se pueden unir las líneas y rellenar los espacios entre ellas para poder dar otra representación gráfica de las direcciones observadas.



A.1.2 Vector medio

Supóngase que se tiene una muestra de tres direcciones dada por los siguientes ángulos[1]:

$$f_1 = 80° \qquad f_2 = 350° \qquad f_3 = 50°$$

Se quiere definir un promedio de las direcciones o un ángulo medio. Resulta obvio pensar que en este caso un ángulo medio apropiado estaría entre 0° y 50°. Al calcular la media aritmética, se tendría:

$$\frac{1}{3}(f_1 + f_2 + f_3) = 160°$$

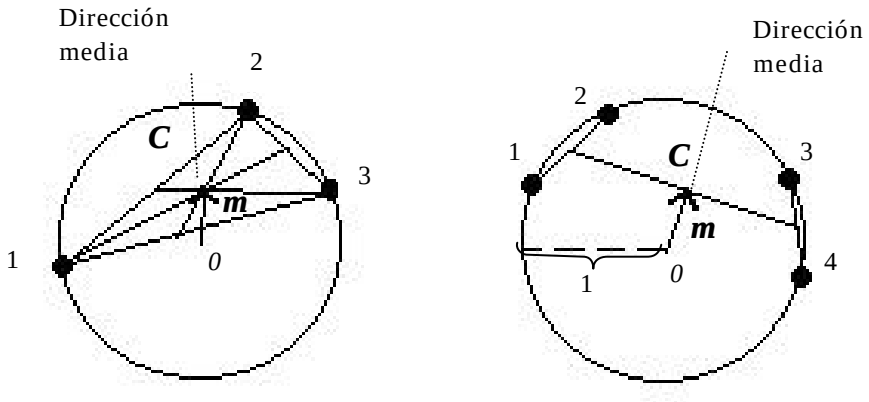
Claramente este resultado no es aceptable. Cuando se reemplaza 350° por su ángulo equivalente ángulo –10°, se obtiene un mejor valor, esto es:

$$\frac{1}{3}(80° - 10° + 50°) = 40°$$

Sin embargo, en general, con más de tres direcciones, no sabemos manejar los ángulos y, por consiguiente, la media aritmética de los ángulos falla al ser aplicada.

Una estadística adecuada para la dirección media está basada en un procedimiento realmente diferente. Se considera un diagrama como el que se muestra a continuación:

[1] BATSCHELET, Edward. “Circular Statistics in Biology”. Ed. Academic Press. Londres. 1981. pág. 7.



Se pide que el círculo sea unitario, es decir, cuyo radio sea de longitud uno. A cada punto se le asigna una masa de igual valor, *M*, y se encuentra *el centro de masa, C*, también llamado *centro de gravedad*. Si este centro es diferente del origen *0*, la línea *0C* define una dirección llamada *la dirección media de la muestra.*

Se mostrarán dos maneras de determinar el centro de masa, una con el álgebra de vectores, y la otra con las funciones trigonométricas.

A.1.2.1 Aplicando el álgebra de vectores

Cada punto en el círculo unitario puede ser representado mediante un vector unitario. Sean los vectores unitarios, $e_1, e_2, ..., e_n$, que constituyen la muestra de las direcciones. Por definición , $| e_i |$ =1 para todo *i* =1,...,n.

Sea el vector :

$$m = \frac{1}{\sum_{i=1}^{n} M_i} \left(M_1 e_1 + M_2 e_2 + ... + M_n e_n \right)$$

el que indica el centro de masa. Si se supone que $M_1 = M_2 = ... = M_n = M$, entonces se tiene que $\sum_{i=1}^{n} M_i = nM$ y se puede simplificar la expresión anterior a :

$$m = \frac{1}{n}(e_1 + e_2 + ... + e_n)$$

Por consiguiente, se tiene que formar el *vector resultante* $\sum_{i=1}^{n} e_i$ y dividir su longitud entre *n*. Se denotará por *m* al *vector medio* de la muestra.

Sea *R* la longitud del vector resultante y *r* la longitud del vector medio, es decir,

$$\left| \sum_{i=1}^{n} e_i \right| = R \quad |m| = r$$

Entonces, $r = R / n$

El centro de masa, *C*, puede caer en la circunferencia del círculo unitario, pero sólo en el caso excepcional cuando todas las masas están juntas en un solo punto. En otro caso, el centro de masa permanece dentro del círculo unitario. Teniendo así que

$$0 \leq R \leq n$$

$$0 \leq r \leq 1$$

A.1.2.2 Aplicando funciones trigonométricas

Se usa un sistema de coordenadas rectangulares con ejes X y Y y un origen *0*. Sea f_i una de los *n* ángulos observados y e_i el correspondiente vector unitario.

Sean x_i y y_i las componentes rectangulares de e_i. Entonces, por definición del seno y del coseno, se tiene

$$x_i = \cos f_i, \quad y_i = \operatorname{sen} f_i$$

Sean $\bar{x}$ y $\bar{y}$ las coordenadas rectangulares del centro de masa. Entonces

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \quad \text{y} \quad \bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n)$$

Partiendo de la definición de x_i y y_i, se tiene que

$$\bar{x} = \frac{1}{n}(\cos f_1 + \cos f_2 + \ldots + \cos f_n)$$

$$\bar{y} = \frac{1}{n}(\operatorname{sen} f_1 + \operatorname{sen} f_2 + \ldots + \operatorname{sen} f_n)$$

Sea otra vez R la longitud del vector resultante con componentes $\sum x_i$ y $\sum y_i$ y sea r la longitud del vector medio con componentes $\bar{x}$ y $\bar{y}$. Entonces

$$r = \left(\bar{x}^2 + \bar{y}^2\right)^{1/2}$$

$$R = \left[\left(\sum x_i\right)^2 + \left(\sum y_i\right)^2\right]^{1/2}, \qquad R = nr$$

También se puede escribir,

$$r = \frac{1}{n}\left[\left(\sum \cos f_i\right)^2 + \left(\sum \operatorname{sen} f_i\right)^2\right]^{1/2}$$

Un caso especial ocurre cuando $\bar{x}$ =0 , $\bar{y}$ =0 y así r = 0, en este caso el vector medio es igual al vector cero. Si se descarta este caso, el vector medio tiene un ángulo bien definido

contra el eje positivo de las equis. Se le llamará a este *el ángulo medio de la muestra* y se denotará como $\overline{f}$. Para calcular $\overline{f}$ aplicamos las ecuaciones anteriores. Obteniendo que

$$\overline{f} = \begin{cases} \arctan(\overline{y}/\overline{x}) & \text{si } \overline{x} > 0 \\ 180° + \arctan(\overline{y}/\overline{x}) & \text{si } \overline{x} < 0 \end{cases}$$

Los casos excepcionales son

$$\overline{f} = \begin{cases} 90° & \text{si } \overline{x} = 0 \text{ y } \overline{y} > 0 \\ 270° & \text{si } \overline{x} = 0 \text{ y } \overline{y} < 0 \\ \text{indeterminado} & \text{si } \overline{x} = 0 \text{ y } \overline{y} = 0 \end{cases}$$

Como una comprobación del cálculo se pueden usar las fórmulas

$$\cos\overline{f} = \overline{x}/r \quad \text{y} \quad \text{sen}\overline{f} = \overline{y}/r$$

A.1.3 Propiedades del vector medio

Ya que el centro de masa está definido independientemente del sistema de coordenadas, el vector medio no depende de la dirección cero.

Supóngase una rotación de la dirección cero por un ángulo, y ; entonces la muestra de valores, $f_1, f_2, ..., f_n$, se expresa como:

$$f_i' = f_i - y \quad \text{para } i = 1,...n$$

De manera similar, para el nuevo ángulo medio se tiene

$$\overline{f}' = \overline{f} - y$$

No obstante, la longitud del vector medio, *r*, permanece invariante.

De las relaciones trigonométricas se obtiene que

$$\operatorname{sen}(f_i - \overline{f}) = \operatorname{sen} f_i \cos \overline{f} - \cos f_i \operatorname{sen} \overline{f}$$

sumando sobre *i* de 1 a *n* y usando las ecuaciones anteriores se tiene

$$\begin{aligned} \sum \operatorname{sen}(f_i - \overline{f}) &= \cos \overline{f} \sum \operatorname{sen} f_i - \operatorname{sen} \overline{f} \sum \cos f_i \\ &= (\overline{x}/r) n \overline{y} - (\overline{y}/r) n \overline{x} \end{aligned}$$

es decir, $\sum \operatorname{sen}(f_i - \overline{f}) = 0$

Los términos negativos y positivos se cancelan. Para una pequeña desviación, se sabe que

$$\operatorname{sen}(f_i - \overline{f}) \approx f_i - \overline{f}$$

Entonces, la ecuación $\sum \operatorname{sen}(f_i - \overline{f}) = 0$ es análoga a la igualdad:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

en un análisis estadístico lineal.

Por otro lado de las relaciones trigonométricas también se obtiene que

$$\cos(f_i - \overline{f}) = \cos f_i \cos \overline{f} + \operatorname{sen} f_i \operatorname{sen} \overline{f}$$

sumando sobre *i* de 1 a *n* y usando las ecuaciones anteriores se tiene

$$\begin{aligned}\sum \cos(f_i - \bar{f}) &= \cos\bar{f}\sum \cos f_i + \operatorname{sen}\bar{f}\sum \operatorname{sen} f_i \\ &= (\bar{x}/r)n\bar{x} + (\bar{y}/r)n\bar{y} \\ &= \frac{n}{r}(\bar{x}^2 + \bar{y}^2) = \frac{n}{r}r^2\end{aligned}$$

es decir, $$\sum_{i=1}^{n} \cos(f_i - \bar{f}) = nr$$

Esta última ecuación se puede reescribir de la siguiente forma

$$\frac{1}{n}\sum_{i=1}^{n} 2\left[1 - \cos(f_i - \bar{f})\right] = 2(1-r)$$

Haciendo uso nuevamente de resultados trigonométricos, se sabe que para una desviación pequeña se tiene que

$$2\left[1 - \cos(f_i - \bar{f})\right] = (f_i - \bar{f})^2$$

Por consiguiente, $$\frac{1}{n}\sum (f_i - \bar{f})^2 \approx 2(1-r)$$

Entonces, esta última ecuación es análoga a

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$$

en un análisis estadístico lineal.

Existe una tercera analogía entre la estadística circular y lineal: la fórmula descubierta por Jacob Steiner (1796-1863) que establece:

$$\sum_{i=1}^{n}(x_i-u)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-u)^2$$

donde u es un número arbitrario. De la fórmula de Steiner se concluye que $\sum_{i=1}^{n}(x_i-u)^2$ alcanza su mínimo para $u=\bar{x}$. Como se demostrará más adelante, la fórmula análoga en estadística circular establece que:

$$\sum\cos(f_i-y)=\sum\cos(f_i-\bar{f})\cos(\bar{f}-y)$$

para un ángulo arbitrario y. Aquí, el lado izquierdo de la igualad alcanza su máximo si $\cos(\bar{f}-y)=1$ lo cual implica que $\bar{f}=y$ (mod 360°). Este máximo es nr de acuerdo a la ecuación antes mencionada: $\sum_{i=1}^{n}\cos(f_i-\bar{f})=nr$.

Para demostrar que $\sum\cos(f_i-y)=\sum\cos(f_i-\bar{f})\cos(\bar{f}-y)$ se divide a f_i-y en dos partes:

$$f_i-y=(f_i-\bar{f})+(\bar{f}-y)$$

y se aplica la ecuación trigonométrica:

$$\begin{aligned}\cos(f_i-y)&=\cos[(f_i-\bar{f})+(\bar{f}-y)]\\&=\cos(f_i-\bar{f})\cos(\bar{f}-y)-\operatorname{sen}(f_i-\bar{f})\operatorname{sen}(\bar{f}-y)\end{aligned}$$

Entonces, $$\sum_{i=1}^{n}\cos(f_i-y)=\sum_{i=1}^{n}\cos(f_i-\bar{f})\cos(\bar{f}-y)-\sum_{i=1}^{n}\operatorname{sen}(f_i-\bar{f})\operatorname{sen}(\bar{f}-y)$$

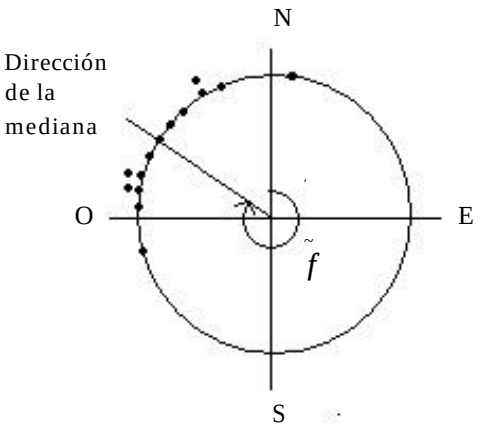
el último término del lado derecho de la desigualdad desaparece como consecuencia de lo ya visto, es decir, $\sum_{i=1}^{n}\operatorname{sen}(f_i-\bar{f})=0$. Y así se completa la demostración.

Haciendo un sumario de las principales analogías, se tiene:

Estadística lineal	Estadística circular
$(x_i - \bar{x})$	$\operatorname{sen}(f_i - \bar{f})$
$\sum (x_i - \bar{x}) = 0$	$\sum \operatorname{sen}(f_i - \bar{f}) = 0$
$(x_i - \bar{x})^2$	$2[1 - \cos(f_i - \bar{f})]$
$\frac{1}{n}\sum (x_i - \bar{x})^2 = s^2$	$\frac{1}{n}\sum 2[1 - \cos(f_i - \bar{f})] = 2(1 - r)$

A.1.4 Dirección de la mediana

Algunas veces es más fácil usar otras medidas de localización en lugar de las vistas anteriormente. Para este propósito se divide la muestra circular por un diámetro de tal manera que la mitad de los puntos de la muestra permanezca en un lado y la otra mitad en el otro lado del diámetro. Si la muestra es unimodal y si el tamaño de la muestra, *n*, es un número impar, entonces el diámetro es definido de una sola forma. Si *n* es par, el diámetro pasa en medio de dos puntos de la muestra. El ángulo del diámetro, medido sobre el lado donde los puntos de la muestra están concentrados es llamado el ángulo mediano, denotado por $\tilde{f}$.



Se debe señalar que una medida de localización (por ejemplo, una dirección preferida) es un valor práctico sólo si los datos están concentrados en un conglomerado alrededor de la media.

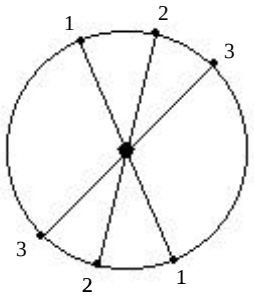
Por ejemplo, si se sabe que un grupo de ratones estaba activo entre las 21:00 y 22:30 horas y después de un descanso vuelven a estar activos entre la 1:20 y 2:40 de la madrugada pero después nuevamente inactivos, una ”hora media” de actividad caería en un periodo de descanso. De tal forma que una media no tiene un significado intuitivo.

A.1.5 Muestras multimodales

Los ángulos medios y los ángulos de la mediana son estadísticas que son intuitivamente significativas sólo para muestras unimodales. Un caso donde se presenta la bimodalidad circular es aquel donde existe la elección entre dos direcciones; no obstante, también se presentan con mucha frecuencia las muestras cuadrimodales.

Si el ángulo entre dos modas es completamente arbitrario ningún método estándar es válido para separar la muestra en dos muestras unimodales. Existe, sin embargo, un tipo especial importante de muestra bimodal que permite la aplicación de la mayoría de técnicas estadísticas, *el caso con dos modas iguales y opuestas*.

Una situación similar ocurre si se observa la posición de las líneas rectas no dirigidas o los ejes no dirigidos. Entonces se puede no hacer distinción entre dos puntos diametralmente opuestos. Si se grafica un diagrama, como el que se presenta a continuación, se obtienen dos puntos para cada línea recta, y el diagrama toma la forma de una figura simétrica con respecto al centro, es decir, la figura coincidiría con ella misma si se rota 180°. En ambos casos, tanto en las líneas rectas no dirigidas como en los ejes no dirigidos, se habla de *datos axiales*.



Para analizar los datos axiales se supone que la posición de una rotación de una línea recta contra una dirección del cero puede ser fijado por el ángulo en el intervalo de 0° a 180° solamente, mientras que para los vectores se tiene que usar el rango completo de 0° a 360°. En otras palabras: para los datos axiales se pueden reducir todos los ángulos módulo 180°. Para aplicar la estadística circular se tiene que relacionar el periodo de 180° con una vuelta completa, de la misma manera como se relaciona 24 horas con 360°. Todo lo que se tiene que hacer es duplicar cada ángulo y reducir los múltiplos módulo 360°. El resultado es una muestra circular unimodal. Entonces se pueden aplicar los métodos estadísticos vistos anteriormente.

El método de duplicar los ángulos ha sido usado en geología desde que éste fue introducido por Krumbein (1939). En años recientes dicho método ha sido aceptado por los biólogos.

Ocasionalmente, incluso las muestras cuadrimodales han sido analizadas. Aquí, las cuatro modas están separadas 90° una de la otra. Un ejemplo donde se presenta este caso es la orientación de los animales puede seguir un patrón. Para transformar una muestra

cuadrimodal en una muestra unimodal y aplicar las técnicas estadísticas, sólo se tiene que cuadruplicar los ángulos observados y reducir los múltiplos módulo 360°.

En general, si hay ? modas igualmente espaciadas, se multiplica cada uno de los n ángulos observados, f_i, por ? y obtener de esta manera una muestra modificada: $?f_1, ?f_2, ..., ?f_n$. Los ángulos pueden ser reducidos módulo 360. Con esta nueva muestra se calcula un vector medio denotado por m_u. Las coordenadas polares de m_u son la longitud del vector medio, r_u, y el ángulo medio, $\overline{f_u}$.

A.2 Medidas de dispersión, sesgo y kurtosis

El ángulo medio de una muestra tomada de una distribución unimodal indica una dirección preferida, pero esto no indica ninguna información de que tan dispersos están los valores de la muestra alrededor de la media. Una medida de dispersión es tan importante como una medida de localización.

A.2.1 Medidas de concentración

El caso extremo de máxima concentración es cuando toda la muestra de puntos cae en un solo punto en el círculo, la longitud del vector medio, *r*, es 1. Cuando la muestra de puntos está cercana entre si, es decir, concentrada en un arco de no más de 20°, el centro de masa está todavía muy cercana a la circunferencia del círculo unitario, y *r* es a lo más 1. Menos concentración lleva a disminuir los valores de *r*. El valor más bajo, es decir, cuando *r* =0 es cuando no hay concentración alrededor de una sola dirección. Entonces, en muestras unimodales, la longitud del vector medio, *r*, sirve como medida de concentración.

Se puede tomar como ejemplo el caso que más se ocupa en experimentos, la trayectoria de un animal al moverse de un lugar a otro, pues difícilmente es una línea recta. Una manera de resolver el problema es dividir la trayectoria en pequeñas secciones las cuales son registradas en intervalos de tiempos iguales. Así la trayectoria es reducida a sucesiones de vectores $?_1, ?_2, ..., .$ Las direcciones de estos vectores pueden ser graficadas como puntos en un círculo unitario y la longitud del vector medio, *r*, es determinado. Si *r* es grande, cercano a 1, indica que la trayectoria es una lo más cercano a una línea recta. Sin embargo, si *r* es pequeña, cercana a 0, las desviaciones de una línea recta son pronunciadas. Así *r* puede servir como un índice de que tanto se la trayectoria sigue una línea recta.

En el caso señalado la medida de *r* no siempre es práctica de calcular. Se puede obtener un índice con valores numéricos cercanos al de *r*, observando que el vector suma $?_1 + ?_2 + ...$ es igual al vector $\vec{PQ}$, donde *P* es el punto inicial y *Q* el punto final de la trayectoria. Sea *D* la distancia entre *P* y *Q*, y sea *W* la actual longitud de la trayectoria.

Entonces, $d = D/W$ es aproximadamente igual a *r*. Este índice modificado fue usado por Ferlin (1973), Duelli (1975), y por Hamilton (1977).

Un valor de r >0 o de *d* >0 no prueba que la orientación es significativa en un sentido estadístico. Los índices *r* y *d* son puramente descriptivos.

A.2.2 Varianza angular y desviación angular

Mientras *r* decrece de 1 a 0 la dispersión se incrementa, entonces parece natural considerar a 1-*r* como una medida de dispersión. Sin embargo, la comparación, hecha con anterioridad, entre la ecuación $\frac{1}{n}\sum(f_i - \overline{f})^2 \approx 2(1-r)$ y $\frac{1}{n}\sum(x_i - \overline{x})^2 = s^2$ sugiere considerar a 2(1-*r*) en lugar de 1-*r* como una estadística idónea.

Por lo tanto, se define

$$s^2 = 2(1-r)$$

como *varianza angular*. Esta cantidad es equivalente a la varianza $\frac{1}{n}\sum(x_i - \bar{x})^2 = s^2$ en estadística lineal.

Tomando la raíz cuadrada, se obtiene una medida de dispersión que es equivalente a la desviación estándar en estadística lineal. Es decir,

$$s = [2(1-r)]^{1/2}$$

es llamada la *desviación media angular* o, si no existe peligro de confusión, simplemente *desviación angular*. Esta es una medida en radianes, para obtener la desviación angular en grados, se tiene lo siguiente

$$s(\text{grados}) = \frac{180^\circ}{p}[2(1-r)]^{1/2}$$

Esta medida fue introducida por Batschelet (1965) y desde entonces se utiliza con frecuencia.

Para una muestra bimodal con modas separadas por 180°, se aplica el método de duplicar los ángulos. Para la muestra modificada se calcula la longitud del vector medio, r_2, y usando la ecuación $s(\text{grados}) = \frac{180^\circ}{p}[2(1-r)]^{1/2}$ se calcula la desviación media estándar; por estar ésta basada en r_2, se denota por s_2. Finalmente, para regresar a la muestra bimodal original, se cancela el efecto de haber duplicado los ángulos por la división de s_2 entre dos. Así, el valor de la desviación angular es

$$s_1 = s_2 / 2$$

A.2.3 Corrección por agrupamiento

El agrupamiento de los datos angulares ocurre cuando el círculo es subdividido en arcos de igual longitud y la muestra de puntos queda contenida en cada arco. Entonces cada arco es de longitud:

$$l = 2p/k \quad \text{si se desea en radianes o,}$$
$$l = 360°/k \quad \text{en grados}$$

l es llamada la *longitud de clase*. Sean $f_1, f_2, ..., f_k$ los puntos medios de los *k* arcos medidos en grados y $n_1, n_2, ..., n_k$ las frecuencias de los puntos muestreados en los correspondientes arcos. Entonces el tamaño de la muestra es

$$n = n_1 + n_2 + ... + n_k$$

De las ecuaciones, $\bar{x} = \frac{1}{n}(\cos f_i + \cos f_2 + ... + \cos f_n)$ y $\bar{y} = \frac{1}{n}(\text{sen} f_i + \text{sen} f_2 + ... + \text{sen} f_n)$ se sigue que las componentes del vector medio, *m*, son

$$\bar{x} = \frac{1}{n}(n_1 \cos f_1 + n_2 \cos f_2 + ... + n_k \cos f_k)$$

$$\bar{y} = \frac{1}{n}(\text{n}_1 \text{sen} f_1 + n_2 \text{sen} f_2 + ... + n_k \text{sen} f_k)$$

Para calcular la longitud del vector medio, *r*, y el ángulo medio, $\bar{f}$, se procede como si los datos no estuviesen agrupados.

El ángulo medio no requiere de corrección debido al agrupamiento, no obstante la longitud del vector medio estará influenciada por éste. Sin una corrección, *r* tiende a ser un poco más pequeña. Por tanto *r* se tiene que multiplicar por un factor c > 1. El valor corregido es

$$r_c = cr$$

Si la longitud de clase está medida en radianes, el factor de corrección es

$$c = \frac{l/2}{\operatorname{sen} l/2}$$

Si la longitud de clase está medida en grados, entonces se cuenta con una tabla para algunos valores de *c*:

k	*l*	*c*
4	90	1.1107
5	72	1.0690
6	60	1.0472
8	45	1.0262
9	40	1.0206
10	36	1.0166
12	30	1.0115
15	24	1.0073
18	20	1.0051
20	18	1.0041
24	15	1.0029
30	12	1.0018
36	10	1.0013
40	9	1.0010
45	8	1.0008
60	6	1.0005

La corrección por agrupamiento afecta indirectamente a la desviación media angular, *s*. Si s_c denota el valor corregido de *s*, se concluye de $s = [2(1-r)]^{1/2}$ que

$$s_c = [2(1-r_c)]^{1/2}$$

donde s_c es una medida en radianes.

Se debe hacer hincapié que la corrección por grupo da buenos resultados sólo en muestras unimodales y distribuciones claramente simétricas. Si el número de grupos excede de 12, la corrección tiene un efecto mínimo y puede ser omitido.

A.2.4 Otras medidas de dispersión

Una medida de dispersión que puede resultar práctica es el *rango*. Ésta es la longitud del arco más pequeño que contiene a todos los puntos de la muestra. El rango es una medida muy cruda de dispersión. Es importante tener presente que esta medida es significativa sólo si la muestra es tomada de una distribución unimodal.

En experimentos donde se parte de un punto de origen y se busca llegar a un punto de destino específico, como por ejemplo las migraciones, no sólo importa cuánto estén concentradas las direcciones alrededor de la dirección media, sino también importa qué tan cercana esté la dirección media de la dirección que indica el punto de destino específico. Para obtener una medida conveniente para este tipo de comportamiento en estos experimentos, se combina la medida de concentración, *r*, con el ángulo de entre la dirección media y la dirección de destino. Para ello se supone que la dirección que señala el destino en particular forma un ángulo q_0 con el eje positivo de las equis. Además, como siempre, sea *r* la longitud del vector medio y sea $\overline{f}$ el ángulo medio. Entonces

$$u = r\cos(\overline{f} - q_0)$$

es la componente del vector medio con respecto a la dirección de destino. De ahí u es conocida como *la componente de destino*. Ésta toma su valor más alto, u =1, si todos los animales se mueven exactamente al destino deseado. Correspondientemente, la dispersión más grande se da cuando la dirección media se desvía mucho de la dirección de destino**.** Así, la componente de destino u puede servir bien como una medida del comportamiento de migración.

La longitud del vector medio, *r*, depende del tamaño de la muestra, así como también la componente de destino. Una muestra de tamaño pequeño favorece a las componentes de destino más grandes. Por lo tanto, las componentes de destino calculadas de diferentes muestras de diferentes tamaños no pueden ser comparadas entre sí.

A.2.5 Medidas de sesgo y kurtosis

Por razones que van más allá del alcance de esta tesis, el término que básicamente determina el *sesgo* de una muestra circular es

$$r_2 \text{sen}(\overline{f}_2 - 2\overline{f})$$

donde r_2 y $\overline{f}_2$ son las estadísticas ya manejadas con anterioridad al tratar el tema de duplicar los ángulos, no obstante aquí $\overline{f}$ es el ángulo medio de la muestra original. En una muestra simétrica esta expresión desaparece.

Similarmente el término con el que se indica la kurtosis, es

$$r_2 \cos(\overline{f}_2 - 2\overline{f})$$

Las medidas de sesgo y kurtosis son significativas sólo para distribuciones unimodales.

Mardia (1972) define la medida de sesgo como $g_1 = \frac{r_2 \text{sen}(\overline{f}_2 - \overline{f})}{\text{s}^3}$ donde *s* es la desviación media angular. El denominador sirve para eliminar posibles efectos de dispersión.

En una muestra cuyo tamaño no exceda de 20 elementos, el sesgo y la kurtosis pueden ocurrir como un efecto espurio causado por la fluctuación aleatoria. De ahí, las medidas de tales desviaciones sólo deberían ser calculadas para muestras grandes.

A.3 Estimación puntual de los parámetros

Se retomarán las estadísticas m, r y $\bar{f}$ de una muestra dada de ángulos, y se considerará, para una población hipotética, la siguiente notación:

	Estadística (muestra)	Parámetro (población)
Vector medio	M	m
Longitud del vector medio	R	r
Ángulo medio	$\bar{f}$	q

La estimación de m, r y q se denotará por $\hat{m}$, $\hat{r}$ y $\hat{q}$ respectivamente. Entonces es posible basar las estimaciones de los parámetros en las estadísticas correspondientes. Así provisionalmente se sugiere:

$$\hat{m} = m$$

$$\hat{r} = r$$

$$\hat{q} = \bar{f}$$

Las ventajas de estas estimaciones sólo pueden ser discutidas en conexión con una distribución particular.

APÉNDICE B

Herramientas Matemáticas

En el análisis de direcciones y de eventos periódicos, el sistema de coordenadas tiene que ser cambiado frecuentemente. Algunas veces las coordenadas rectangulares son una herramienta apropiada, pero en otras ocasiones las coordenadas polares son más útiles. El cambio de un sistema a otro necesita aplicaciones cuidadosas de funciones trigonométricas.

B.1 Ángulos

La posición de un punto, *P*, en el plano cartesiano puede ser determinado únicamente por dos coordenadas, *x* y *y*,. Pero *P* también puede ser caracterizado por un ángulo, f. Si *P* coincide con el origen, *O*, ningún ángulo queda definido. Por lo que se supone, en lo consiguiente, que *P* es distinto de *O*. Se introduce una semilínea, *l*, y se supone que originalmente coincide con el eje positivo de las equis. Se dice que el eje positivo de las equis tiene dirección cero y es también llamado el *eje polar*. Ahora se rota la semilínea en sentido levógiro hasta que ésta pasa por primera vez por el punto *P*. Entonces a la magnitud de rotación se le llama un *ángulo*. Para obtener un ángulo negativo se gira la semilínea en sentido dextrógiro.

La definición anterior de un ángulo no es siempre conveniente. En la medición de direcciones no se está interesado en la suma o resta de rotaciones completas; lo que se desea sólo es asociar un ángulo con una dirección dada. Por lo que se tomarán los ángulos de la siguiente manera:

$$q = f \pmod{360°}$$

lo que significa que q y f difieren uno del otro por un múltiplo de 360, es decir,

$$q = f \pm k * 360°$$

donde *k* es un entero positivo.

Hasta este momento se ha tratado la definición de ángulo de una línea recta direccionada, no obstante se puede dar un trato semejante a los ángulos en caso de que la línea recta no lleve dirección, pues la única diferencia radica en los ángulos q y $f \pm 180°$ indican la dirección de la misma línea recta. Es decir, dos ángulos q y f son equivalentes, o en terminología tradicional, congruentes, si estos difieren por un múltiplo de 180°, que es

$$q = f \pmod{180°}$$

Un ángulo utilizado para medir una dirección en un plano horizontal es llamado un *acimut* en astronomía y geografía si la dirección del cero señala al norte y la rotación es realizada en el sentido de las manecillas del reloj (como el movimiento de las estrellas). El acimut es ligeramente diferente del ángulo introducido antes. Sin embargo, esto es irrelevante para el tratamiento matemático y para las aplicaciones.

Para determinar completamente la posición de un punto, *P*, en el plano, se tiene que combinar el ángulo, f, con la distancia, *r*, del origen. Ambas medidas, *r* y f, son llamadas *las coordenadas polares de P*. Se recuerda que hay un punto excepcional, el origen, *O*, éste está dado por *r* =0 solamente, f no está definido.

Se debe tomar la siguiente precaución: *No todos los ángulos son variables circulares*. Si los ángulos son medidos en un sector que fue restringido por adelantado, estos ángulos se comportan como variables lineales y pueden ser tratados por el análisis de la estadística

lineal. Esto es, por ejemplo, el caso cuando los animales entran a una área limitada por un semicírculo con el ángulo en los límites de 0 ° a 180°[2].

De la misma manera, en un estudio referente a la actividad del plankton, realizado por R. Margalef (1957), la actividad se concentraba en la primavera y el verano, y no había prácticamente ninguna actividad en el invierno. Por lo tanto, la distribución anual de actividad tenía un hueco natural y era razonable de tratar la distribución como lineal y no como circular[3]. La latitud de una posición en la tierra no es una variable circular incluso cuando es medida por un ángulo. Sin embargo la longitud sí es una variable circular. En ocasiones, los ángulos negativos y positivos no son distinguidos unos de otros. Estos ángulos son llamados *distancia angular* (cuya definición se dará más adelante). Nuevamente una distancia angular no es una variable circular.

Algunas veces se tiene que cambiar la unidad con la cual con la cual un ángulo es medido, ya sea en radianes o bien en ángulos. Incluso las unidades de los ángulos son usados en conexión con los periodos de tiempo. Si el periodo de tiempo es un día, es decir, 24 horas, un ángulo de 15° corresponden a una hora y 1° corresponde a 4 minutos. Las dificultades se incrementan cuando el periodo de tiempo consiste en un año, pues hay años que constan de 365 días y otros de 366 días. El mediodía del 1 de marzo corresponde a 58.7° en el primer caso y a 59.5° en el segundo.

B.2 Vectores

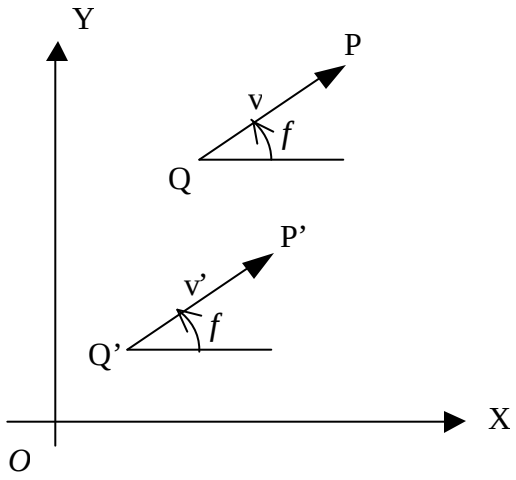
Los vectores fueron inventados por físicos para estudiar conceptos como la fuerza y la velocidad, en los cuales no sólo la magnitud sino la dirección son de importancia. Hoy día los vectores son utilizados frecuentemente para propósitos algebraicos y geométricos. Los vectores son especialmente útiles en el análisis de direcciones.

[2] BATSCHELET Edwuard, “Circular Statistics in Biology”, Ed. Academic Press, Londres., 1981, pág. 231.
[3] Ibídem. pág. 231.

En la siguiente gráfica, fueron dados dos puntos, *P* y *Q*. El segmento de línea direccionado de *Q* a *P* es llamado un *vector* y denotado por $\vec{QP}$ o por una letra. *Q* es llamado la base o cola y *P* la punta del vector.

Dos vectores son considerados iguales si tienen la misma dirección y la misma longitud. Con respecto a la gráfica:

$$\vec{QP} = \vec{Q'P'} \text{ o } v = v'$$



B.3 Funciones trigonométricas

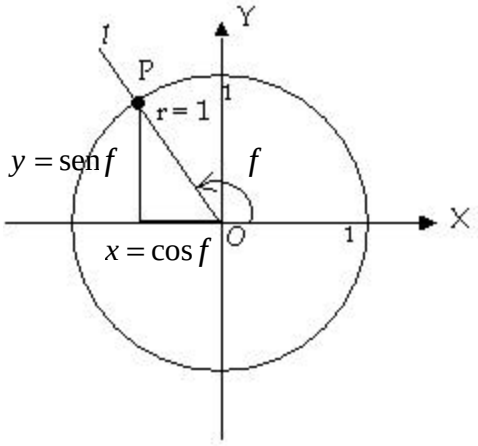
Se supondrá un sistema de coordenadas rectangulares, con la misma longitud, de uno, en ambos ejes. Se introduce una semilínea, *l*, como la empleada para la definición de ángulo. El eje positivo de las equis y la semilínea son las componentes para formar f. También se considera que la intersección del círculo unitario con *l* origina el punto *P*, que es determinado únicamente por el ángulo f.

Definición[4]: Sea *P* un punto con coordenadas polares 1 y f. Denotadas en sus coordenadas rectangulares por *x* y *x*. Entonces el coseno y el seno de f son:

$$\cos f = x$$

$$\operatorname{sen} f = y$$

x y *y* son determinadas únicamente por f.



Si el dominio de f consiste en todos los números reales, y *x* y *y* son *funciones periódicas* de f, una nueva rotación alrededor del círculo unitario genera los mismos valores de *x* y *y*. El periodo es 360°.

Las siguientes relaciones trigonométricas se pueden deducir fácilmente del círculo unitario anterior. Éstas tienen una útil conexión con las distribuciones circulares:

$$\operatorname{sen}^2 + \cos^2 = 1$$

[4] Ibídem. pág. 236.

$$\operatorname{sen}(-f) = \operatorname{sen}(360^\circ - f) = -\operatorname{sen} f$$
$$\cos(-f) = \cos(360^\circ - f) = -\operatorname{sen} f$$

$$\operatorname{sen}(180^\circ - f) = \operatorname{sen} f$$
$$\cos(180^\circ - f) = -\cos f$$
$$\operatorname{sen}(180^\circ + f) = -\operatorname{sen} f$$
$$\cos(180^\circ + f) = -\cos f$$

También se tienen las siguientes fórmulas:

$$\operatorname{sen}(f + j) = \operatorname{sen} f \cos j + \cos f \operatorname{sen} j$$
$$\operatorname{sen}(f - j) = \operatorname{sen} f \cos j - \cos f \operatorname{sen} j$$

$$\cos(f + j) = \cos f \cos j - \operatorname{sen} f \operatorname{sen} j$$
$$\cos(f - j) = \cos f \cos j + \operatorname{sen} f \operatorname{sen} j$$

Frecuentemente es necesaria una tercera función trigonométrica, la función tangente, que se define de la siguiente manera:

$$\tan f = \frac{\operatorname{sen} f}{\cos f}$$

Mientras que el periodo de la función seno y coseno es de 360°, la función tangente tiene un periodo de sólo 180°.

Como una consecuencia de las relaciones anteriores se tiene

$$\tan(f + 180°) = \frac{\operatorname{sen}(f + 180°)}{\cos(f + 180°)} = \frac{-\operatorname{sen} f}{-\cos f} = \tan f$$

En una ecuación como $\cos f = x$, el ángulo no es determinado por un solo valor de *x*. De hecho, hay un infinito de soluciones. Por lo tanto, cuando se definen funciones inversas se tiene que saber el intervalo en el cual el ángulo toma valores. Entonces, mientras $\cos f$ decrece de 1 a -1, si f incrementa de 0° a 180°, f es determinado únicamente en este intervalo. Así un ángulo limitado por $0° \le f \le 180°$ es una función de *x*, llamada *la función inversa de* $x = \cos f$ y se escribe

$$f = \arccos x \quad (-1 \le x \le 1, \quad 0° \le f \le 180°)$$

ó $$f = \cos^{-1} x$$

Similarmente $\text{sen} f$ incrementa de –1 a 1, si f incrementa de –90° a 90°. Entonces *la función inversa de* $y = \text{sen} f$ es

$$f = \text{arcsen } y \quad (-1 \le y \le 1, \quad -90° \le f \le 90°)$$

ó $$f = \text{sen}^{-1} y$$

Además, $u = \tan f$ puede ser resuelto con únicamente el valor de f, si f es limitado por el intervalo que va de -90° a 90°. Entonces, *la función inversa de* $u = \tan f$ es

$$f = \arctan u \quad (-90° < f < 90°)$$

ó $$f = \tan^{-1} u$$

Se hace notar que *u* puede tomar cualquier valor real arbitrario.

Ahora se aplicarán las funciones trigonométricas para obtener la conversión de coordenadas polares a coordenadas rectangulares. Si *r* es la coordenada polar de la distancia, se tiene

$$x = r \cos f \qquad y = r \text{sen} f$$

Sin embargo, la conversión de coordenadas rectangulares en coordenadas polares es menos simple. De las ecuaciones anteriores se sigue que

$$x^2 + y^2 = r^2\left(\cos^2 f + \text{sen}^2 f\right) = r^2$$

y $$r = \left(x^2 + y^2\right)^{1/2}$$

También se sabe que $$\tan f = \frac{\text{sen} f}{\cos f} = \frac{y}{x}$$

Suponiendo que $x \neq 0$. Como se ha mencionado $\arctan\left(y/x\right)$ toma valores entre -90° y 90°, y esto reproduce ángulos polares sólo en el primero y cuarto cuadrantes donde $x > 0$. Para $x < 0$ el punto (x, y) cae en el segundo y tercer cuadrante. De ahí, f toma valores entre 90° y 270°. A partir de que $\tan f$ tiene un periodo de 180°, se tiene que sumar 180° a $\arctan\left(y/x\right)$. Por lo tanto:

$$f = \begin{cases} \arctan\left(y/x\right) & \text{si } x > 0 \\ 180° + \arctan\left(y/x\right) & \text{si } x < 0 \end{cases}$$

Se tiene que completar este resultado por algunos casos excepcionales

$$f = \begin{cases} 90° & \text{si } x = 0 \text{ y } y > 0 \\ 270° & \text{si } x = 0 \text{ y } y < 0 \\ \text{indeterminado} & \text{si } x = 0 \text{ y } y = 0 \end{cases}$$

Para continuar con el estudio de la periodicidad se requiere analizar funciones como $\text{sen}2f$, $\text{sen}3f$, ..., $\cos 2f$, $\cos 3f$, etc. Si f incrementa de 0° a 180°, $2f$ incrementa de 0° a 360°. Entonces, para todos los valores de f de 0° a 180°, $\text{sen}2f$ y $\cos 2f$ toman todos los posibles valores de –1 a 1, y el periodo es 180°.

De la misma manera, si f incrementa de 0° a 120°, $3f$ incrementa de 0° a 360°. De ahí $\text{sen}3f$ y $\cos 3f$ tienen periodo 120°. Se puede seguir así con el estudio de $\text{sen}\, nf$ y $\cos nf$ para cualquier *n* número natural.

Por otro lado, resulta práctico contar con algunas aproximaciones de f cuyos valores sean cercanos a 0°. Éstas son:

$$\text{sen}f \approx f$$

$$\cos f \approx 1-\frac{1}{2}f^2$$

$$2(1-\cos f) \approx f^2$$

donde f es medido en radianes.

Con frecuencia se requiere calcular el ángulo entre dos direcciones dadas. Las direcciones pueden ser representadas por las semilíneas l_1 y l_2 con un vértice común. Las semilíneas dividen al círculo unitario en dos arcos. Uno de ellos es de longitud menor o igual a 180° y el otro tiene una longitud mayor o igual de 180°. Se selecciona el más pequeño de los dos arcos y a este se le llama *distancia angular* de dos direcciones. Sean f y j las coordenadas polares correspondientes a los ángulos de l_1 y l_2 con respecto a una dirección arbitraria del cero. Entonces se denotará a la distancia angular como

$$|f, j|$$

De acuerdo con la definición se tiene la siguiente desigualdad

$$0° \leq |f, j| \leq 180°$$

El cálculo de la distancia angular no resulta trivial. La distancia no es igual a $f-j$, pues la diferencia podría tomar valores entre –360° y 360°. Y el valor absoluto solamente no es

la solución idónea para este caso, pues podría exceder de 180°. Entonces a 360° se le debe restar el valor absoluto. Entonces la solución correcta del problema es

$$|f,j| = \text{más pequeño de los dos ángulos } |f-j| \text{ y } 360° - |f-j|$$

Una alternativa sería recurrir a las ecuaciones vistas con anterioridad, ya que $\cos(f-j)=\cos(j-f)$, lo que elimina la asimetría entre f y j. Además, la función inversa del coseno toma valores entre 0° y 180°. Esto conduce al siguiente resultado práctico

$$|f,j| = \arccos[\cos(f-j)]$$

Otra solución para medir la distancia entre dos direcciones, es la siguiente función

$$d(f,j)=1-\cos(f-j)$$

Si $f=j$, entonces $d(f,j)=0$. Cuando la diferencia $f-j$ incremente en valor absoluto, $d(f,j)$ decrece monótonamente. El máximo valor es 2, cuando f difiere de j por 180°.

Entonces

$$d(f,j)=\begin{cases} 0 & \text{si } |f,j|=0° \\ 1 & \text{si } |f,j|=90° \\ 2 & \text{si } |f,j|=180° \end{cases}$$

La distancia angular $|f,j|$ así como la medida $d(f,j)$ son utilizadas en la estadística circular.

B.4 Rotación del plano

Las coordenadas polares son útiles para demostrar como los puntos en el plano pueden ser rotados alrededor del origen. Sea *P* un punto (*x*, *y*) con coordenadas polares, *r* y f. Si se rota el plano por un ángulo q, *P* se mueve a un punto *P*' con coordenadas rectangulares *x*', *y*' y coordenadas polares *r* y $f' = f + q$. Es decir, con base en las ecuaciones $x = r\cos f$ y $y = r\text{sen} f$, se tiene

$$x' = r\cos(f + q), \qquad y' = r\text{sen}(f + q)$$

O bien, utilizando las ecuaciones vistas con anterioridad

$$\cos(f + q) = \cos f \cos q - \text{sen}\, f \text{sen} q$$
$$\text{sen}(f + q) = \text{sen} f \cos q + \cos f \text{sen} q$$

se tiene

$$x' = r\cos f \cos q - r\text{sen} f \text{sen} q$$
$$y' = r\cos f \text{sen} q + r\text{sen} f \cos q$$

Gracias a que $x = r\cos f$ y $y = r\text{sen} f$, se concluye que

$$x' = x\cos q - y\text{sen} q$$
$$y' = x\text{sen} q + y\cos q$$

Estas ecuaciones muestran como la rotación del plano alrededor del origen afecta a las coordenadas rectangulares.

Para obtener la transformación inversa, que es la rotación por el ángulo $-q$, no se necesita despejar las ecuaciones anteriores con respecto a *x* y *y*; simplemente se puede reemplazar q por $-q$, *x* y *y* por *x*' y *y*' respectivamente. Así se tiene

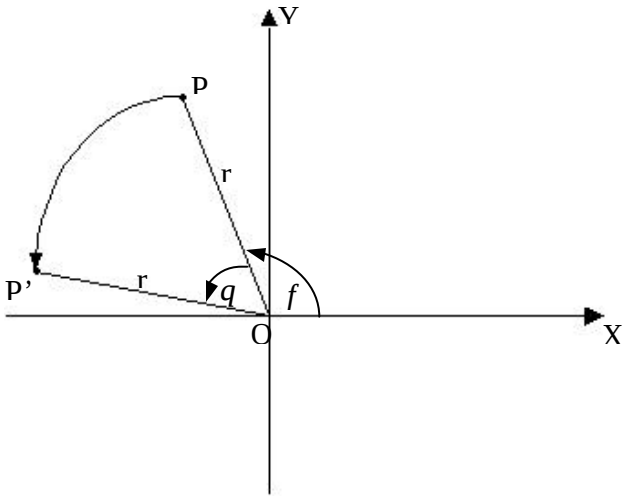
$$x = x' \cos q + y' \operatorname{sen} q$$

$$y = -x' \operatorname{sen} q + y' \cos q$$

Es importante examinar cuáles son los términos que permanecen constantes o invariantes bajo rotaciones del plano alrededor del origen. Naturalmente una función invariante es

$$r^2 = x^2 + y^2$$

Además, la diferencia entre dos ángulos $f - j$ es una función invariante. De lo anterior se deduce que las medidas $|f, j|$ y $d(f, j)$ son medidas invariantes bajo rotaciones.



Bibliografía

- Ackermann, H. (1997). 'A note on circular nonparametrical classification', Biometrical Journal, 5, 557-587.
- Anderberg, M.r. (1973). Cluster Analysis for Applications, New York: Academic Press.
- Batschelet, E. (1981). Circular Statistics in Biology, Lodon: Academic Press.
- Bondy, J. A. (1976). Graph Theory with Applications. Elsevier Science Ltd.
- Chatfield, C. and Collins, A. J. (1980). Introduction to Multivariate Analysis. London: Chapman and Hall.
- Everitt, B. (1993). Cluster Analysis, London: Edward Arnold.
- Fisher, N. I. (1993). Statistical Analysis of Circular Data, Cambridge: Cambridge University Press.
- Hartigan, J. A. (1975). Clustering Algorithms. New York: Wiley.
- Jammalamadaka, S. R. (2001). Topics in Circular Statistics. World Scientific.
- Jardine, N. and Sibson, R. (1971). Mathematical Taxonomy. New York: John Wiley & Sons Ltd.
- Kaufman, L. and Webwer, R.O. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, New York: John Wiley & Sons Ltd.

- Ling, R. F. (1972). 'On the theory and construction of k-clusters'. Comp. J. 15, 326-332.
- Love, M. (1963). Probability Theory. New York: D. Van Nostrand Company.
- Lund, U. (1999). Cluster Analysis for Directional Data, Commun. Statist, 4, 1001-1009.
- Manly Bryan F.J., Multivariate Statistical Methods, Chapman & Hall 2° ed. (1994), 1-145.
- Mardia, K. V. (1972). Statistics of Directional Data, London: Academic Press.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1993). Multivariate Analysis. London: Academic Press.
- Myers, W. and Patil, G. P. (1997). 'Cluster Coordinated Composites of Diverse Datasets on Several Spatial Scales for Designing Extensive Environmental Sample Surveys'. Technical Report Number 97-1103. Center for Statistical Ecology and Environmental Statistics, PA.
- Swokowski, E. W and Cole, J. A. (1996). Álgebra y Trigonometría con Geometría Analítica. Grupo Editorial Iberoamericana.
- Tinsley, H. and Brown, S. D. (2000). Handbook of Applied Multivariate Statistics and Mathematical Modeling. London: Academic Press, 641-663.