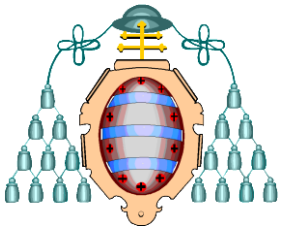


UNIVERSIDAD DE OVIEDO

ESCUELA POLITÉCNICA DE MIERES

MÁSTER EN TELEDETECCIÓN Y SISTEMAS DE INFORMACIÓN GEOGRÁFICA

Descubrimiento de conocimiento en bases de datos espaciales.

TRABAJO FIN DE MÁSTER

AUTOR: Iñaki Díaz Covián

DIRECTOR: Gil González Rodríguez

JULIO, 2014

"La inspiración existe,

pero tiene que encontrarte trabajando".

Pablo Picasso.

Contenido

1. Resumen 4
Abstract 4
2. Introducción 5
2.1. El análisis Cluster 7
2.2. Cluster por individuos y por variables 9
2.3. Clasificación de las técnicas clusters 9
2.4. Medidas de asociación 14
2.5. Etapas en el Análisis Cluster 18
3. Objetivos 21
4. Procedimiento 22
4.1. Área de estudio 22
4.2. Software R 23
4.3. Cluster Jerárquico: “WARD” 24
4.4. Cluster No Jerárquico: K-means o K-medias 32
4.5. Fuzzy K-means o Fuzzy clustering 37
4.6. Diferencias entre ambos métodos 41
4.7. Ocean Data View 42
5. Resultados 43
5.1. Cluster Jerárquico 44
5.2. Cluster No Jerárquico: K-means 52
5.3. Fuzzy K-means 58
6. Conclusiones 64
7. Bibliografía 65

1. Resumen

El análisis cluster es un modo de descubrir conocimientos en bases de datos de forma no supervisada, es decir, es el propio análisis el que agrupa y encuentra los patrones comunes dependiendo de una serie de variables.

A lo largo del trabajo vamos a ver qué es un análisis cluster y cómo funciona, exploraremos diferentes variantes de sus técnicas más comunes, viendo las diferencias entre ellas a la hora de trabajar y sus resultados.

Desarrollaremos estas variantes con un conjunto básico de funciones en el entorno R. Documentaremos estos procedimientos con un ejemplo práctico sobre tipos de cáncer, para a continuación desarrollar un estudio experimental empleando una base de datos de la extensa base de datos oceanográfica (http://odv.awi.de/en/).

Abstract

Cluster analysis is a way of discovering knowledge in databases unsupervised way, namely is the analysis itself that groups find common patterns depending on a number of variables.

Throughout the paper we will see that it is cluster analysis, how it works, we will explore different variants of the most common techniques, seeing the differences between them when working and their results.

These variants will develop a basic set of functions in R. We will document the following setting with a practical example of cancers, to then develop a pilot study using a database of extensive oceanographic database (http://odv.awi .de /en /).

2. Introducción

Desde siempre el hombre ha querido diferenciar entre las clases de objetos, y para ello lo ideal es clasificarlos en categorías. Pero la cantidad de objetos, sucesos y datos que nos encontramos en el día a día es de un volumen mayúsculo para almacenar y procesarlos por un único individuo.

La clasificación o identificación es el proceso o acto de asignar un nuevo objeto u observación en su lugar correspondiente dentro de un conjunto de categorías establecido.

Actualmente y cada vez más trabajamos con datos y estadísticas, es una parte fundamental en determinadas áreas o estudios de investigación, pero también para el ciudadano corriente, nos gusta clasificar todo lo que se encuentra a nuestro alrededor, como los programas de la tv (deportes, series,…) o la bolsa de la compra (frutas, congelados,…).

Al igual que es una actividad humana conceptual básica, la clasificación también es fundamental en la mayoría de las ramas de la ciencia. En Biología, por ejemplo, la clasificación de organismos ha sido una preocupación desde las primeras investigaciones biológicas.

Ya en la Grecia clásica, Aristóteles construyó un sistema de clasificación de especies del reino animal, empezó dividiendo los animales en dos grupos principales:

- Los que tenían sangre roja (correspondiente a los vertebrados).
- Y los que no la tienen (invertebrados).

Además subdividió estos grupos dependiendo de la forma del nacimiento.

También, Teofrasto redactó el primer informe fundamental sobre la estructura y clasificación de las plantas. El resultado fueron unos libros documentados y profundos, abarcando tantos conceptos en sus temas que nos han provisto de la base de las investigaciones biológicas durante mucho tiempo. Hasta los siglos XVII y XVIII cuando los exploradores Europeos profundizaron es este trabajo.

Con estos ejemplos queremos decir que la clasificación ha jugado un papel central en el desarrollo de teorías en diferentes campos de la ciencia. Como pueden ser las citadas biología o zoología, astronomía con la clasificación de las estrellas o en química con la tabla periódica.

La taxonomía (del griego ταξις, taxis, ‘ordenamiento’, y νομος, nomos, ‘norma’ o ‘regla’) es, en su sentido más general, la ciencia de la clasificación. En Biología, la teoría y práctica de la clasificación de los organismos es conocida como taxonomía biológica, la ciencia de ordenar la diversidad biológica en taxones anidados unos dentro de otros, ordenados de forma jerárquica, formando un sistema de clasificación.

Gracias a las técnicas numéricas la taxonomía deja de ser un arte, para tener una base y método científico. Estas técnicas crecen con el desarrollo de los ordenadores y sus altas capacidades para desarrollar operaciones matemáticas, así como con sus áreas de aplicación desde la psiquiatría, arqueología a la investigación de mercados,…

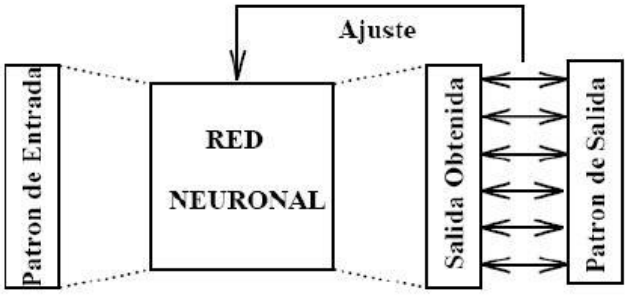
Es en este punto es donde entran en juego los sistemas de clasificación automática. Las técnicas de clasificación automática se pueden agrupar como supervisadas o no supervisadas, las diferencias entre ellas son:

Aprendizaje supervisado

Necesita un experto/investigador que mida el funcionamiento del sistema.

Maneja información de error o de control.

Esta información se emplea para guiar al sistema. Hay varios algoritmos que establecen cómo se realiza esta retroalimentación, el más conocido o empleado es el *backtracking*

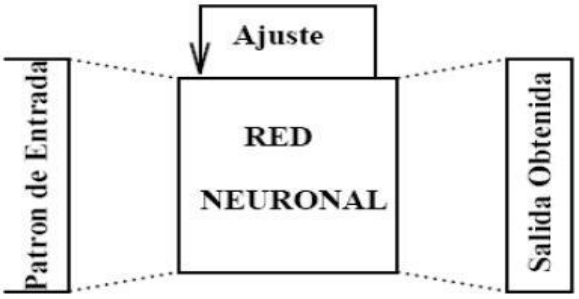


Aprendizaje no supervisado

No utiliza información externa.

Reajuste automático de los parámetros.

Auto organización de la información.



La clasificación no supervisada, en la que nos centraremos, se llama Análisis Cluster. La clasificación supervisada mantiene el mismo nombre en ingeniería y se denomina clasificación de análisis discriminante en estadística.

Como vemos, estos métodos pueden tener diferentes nombres dependiendo de su área, en biología es taxonomía numérica, en psicología es el Q-análisis,… aunque actualmente el más conocido o genérico es Análisis Cluster.

Y todas quieren lo mismo, a un conjunto de *n* objetos, cada uno viene descrito por un conjunto de características o variables, hay que encontrar una división en un número de clases determinadas.

2.1. El análisis Cluster

Análisis Cluster es el nombre genérico de una amplia variedad de procedimientos que pueden ser usados para crear una clasificación. Más concretamente, un método cluster es un procedimiento estadístico multivariante que comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que llamaremos clusters.

En Análisis Cluster poca o ninguna información es conocida sobre la estructura de las categorías, lo cual lo diferencia de los métodos multivariante de asignación y discriminación. Aunque con frecuencia se tiene algunas nociones sobre características deseables e inaceptables a la hora de establecer un determinado esquema de clasificación.

El objetivo operacional es ordenar las observaciones en grupos tales que el grado de asociación natural sea alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes.

En términos operacionales:

El número de formas en las que se pueden clasificar ***m*** observaciones en ***k*** grupos es un número de Stirling de segunda especie (Abramowitz y Stegun, 1968).

$$\mathbb{S}_m^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^m$$

El problema se complica por el hecho de que usualmente el número de grupos es desconocido, por lo que el número de posibilidades es la suma de números de Stirling; así, por ejemplo, en el caso de ***m*** observaciones tendríamos que el número total de posibles clasificaciones sería:

$$\sum_{j=1}^{m} \mathbb{S}_m^{(j)}$$

Que es un número excesivamente grande, por lo que el número de posibles clasificaciones puede ser enorme (por ejemplo, en el caso de 25 observaciones, se tiene que:

$$\sum_{j=1}^{25} \mathbb{S}_{25}^{(j)} > 4 \times 10^{18}$$

Así es necesario encontrar una solución aceptable considerando sólo un pequeño número de alternativas, para ello utilizamos las Técnicas Clusters.

Los usos del Análisis Cluster pueden ser resumidos bajo cuatro objetivos principales:

1. Desarrollar una tipología o clasificación.

2. Investigar esquemas conceptuales útiles para agrupar entidades.

3. Generar hipótesis a través de la exploración de los datos.

4. Intentar determinar si tipos definidos por otros procedimientos están de hecho presentes en un conjunto de datos.

De ellos, la creación de clasificaciones es el más frecuente pero en la mayoría de los casos estos objetivos se combinan.

Precauciones sobre los métodos cluster:

1. Son métodos heurísticos basados en procedimientos descriptivos por lo que no son representativos de la población.

2. La mayor parte han nacido al amparo de ciertas ramas de la ciencia, por lo que, están impregnados de un cierto sesgo procedente de esas disciplinas. Esta cuestión es importante puesto que puede haber, por ejemplo, métodos que sean útiles en psicología pero no en biología o viceversa.

3. Distintos procedimientos clusters pueden generar soluciones diferentes sobre el mismo conjunto de datos. Una razón para ello radica en el hecho ya comentado de que los métodos clusters se han desarrollado a partir de fuentes dispares que han dado origen a reglas diferentes de formación de grupos. De esta manera, es necesaria la existencia de técnicas que puedan ser usadas para determinar qué método produce los grupos naturalmente más homogéneos en los datos.

2.2. Cluster por individuos y por variables

El punto de partida para el Análisis Cluster es, en general, una matriz X que proporciona los valores de las variables para cada uno de los individuos objeto de estudio:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

La **i**-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la **j**-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

Estos procedimientos pueden aplicarse a **X´** obteniéndose así una clasificación de las variables que describen cada individuo. De hecho, muchas de las técnicas cluster existentes (no todas) pueden ser aplicadas para clasificar variables; incluso algunos paquetes estadísticos, como es el caso de BMDP, incluyen implementaciones por separado que permiten realizar análisis cluster por variables (1M) y análisis cluster por individuos (2M).

2.3. Clasificación de las técnicas clusters

La clasificación que vamos a dar está referida a algunas de las distintas técnicas clusters existentes. Es extensa, ya que múltiples son los métodos existentes.

A grandes rasgos se distinguen dos grandes categorías de métodos clusters: métodos jerárquicos y métodos no jerárquicos.

2.3.1 Métodos jerárquicos

Estos métodos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos:

Los métodos **aglomerativos**, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya en el estudio. Se selecciona una medida de similitud, agrupándose los dos grupos o clusters con mayor similitud. A partir de ahí se van formando grupos de forma ascendente Así se continúa hasta que:

1. Se forma un solo grupo.

2. Se alcanza el número de grupos prefijado.

3. Se detecta, a través de un contraste de significación, que hay razones estadísticas para no continuar agrupando clusters, ya que los más similares no son lo suficientemente homogéneos como para determinar una misma agrupación

Los métodos **disociativos** o divisivos, también llamados descendentes, realizan el proceso inverso al anterior. Empiezan con un conglomerado que engloba a todos los individuos. A partir de este grupo inicial se van formando, a través de sucesivas divisiones, grupos cada vez más pequeños. Al final del proceso se tienen tantos grupos como individuos en la muestra estudiada.

Independientemente del proceso de agrupamiento, hay diversos criterios para ir formando los conglomerados; todos estos criterios se basan en una matriz de distancias o similitudes.

Por ejemplo, dentro de los métodos aglomerativos destacan:

1. Método del amalgamamiento simple.

2. Método del amalgamamiento completo.

3. Método del promedio entre grupos.

4. Método del centroide.

5. Método de la mediana.

6. Método de Ward.

Dentro de los métodos disociativos, destacan, además de los anteriores, que siguen siendo válidos:

1. El análisis de asociación.

2. El detector automático de interacción.

Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de ***dendrograma***, en el cual se puede seguir de forma gráfica el procedimiento de unión seguido, mostrando que grupos se van uniendo, en qué nivel concreto lo hacen, así como el valor de la medida de asociación entre los grupos cuando éstos se agrupan (valor que llamaremos nivel de fusión).

2.3.2 Métodos no jerárquicos

Están diseñados para clasificar individuos en una clasificación de *K* clusters, donde *K* se especifica a priori o bien se determina como una parte del proceso. Siendo esta, posiblemente, la principal diferencia respecto de los métodos jerárquicos.

Otra diferencia reside en qué estos métodos se trabaja con la matriz de datos original y no precisan su conversión en una matriz de distancias o similitudes.

La idea central de la mayoría de estos procedimientos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos clusters para obtener una partición *mejor*.

Los diversos algoritmos existentes se diferencian sobre todo en lo que se entiende por *una partición mejor* y en los métodos que deben usarse para conseguir mejoras. La idea general de estos métodos es muy similar a la señalada en los algoritmos descendentes en más de un paso empleados en la optimización sin restricciones en programación no lineal. Tales algoritmos empiezan con un punto inicial y generan una secuencia de movimientos de un punto a otro hasta que se encuentra un óptimo local de la función objetivo.

Los métodos estudiados ahora comienzan con una partición inicial de los individuos en grupos o bien con un conjunto de puntos iniciales sobre los cuales pueden formarse los clusters. En muchos casos, la técnica para establecer una partición inicial es parte del

algoritmo cluster, aunque estas técnicas usualmente son proporcionadas por sí solas más que como una parte del algoritmo cluster.

Pedret en 1986 agrupa los métodos no jerárquicos en cuatro familias:

1. Métodos de Reasignación.

Permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Dentro de estos métodos están:

a) El método K-Medias.

b) El Quick-Cluster análisis.

c) El método de Forgy.

d) El método de las nubes dinámicas.

2. Métodos de búsqueda de la densidad.

Dentro de estos métodos están los que proporcionan una aproximación tipológica y una aproximación probabilística.

En el primer tipo, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre ellos destacan:

a) El análisis modal de Wishart.

b) El método Taxmap.

c) El método de Fortin.

En el segundo tipo se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Se introduce el cluster en la inferencia estadística. El porqué de los clusters está explicado por la existencia de distintas poblaciones y se trata de descubrirlas.

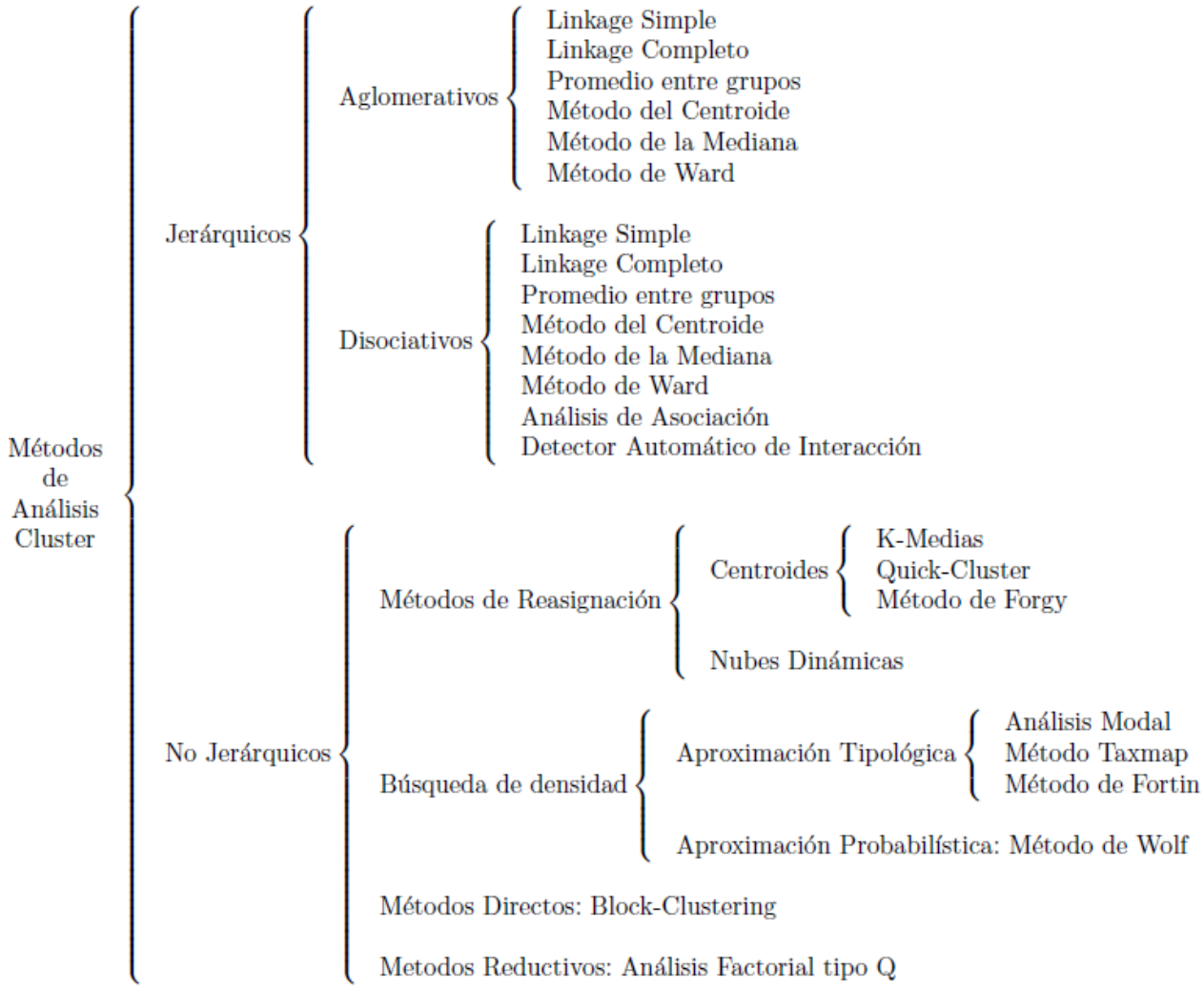
En los anteriores, por el contrario, no importaba si eran diferentes poblaciones y de dónde venían los datos o no, sólo se pretendía obtener conocimiento sobre los datos en sí mismo.
Entre los métodos de este tipo destaca el método de las combinaciones de Wolf.

3. Métodos directos.

Permiten clasificar simultáneamente a los individuos y a las variables. El algoritmo más conocido dentro de este grupo es el Block-Clustering.

4. Métodos de reducción de dimensiones.

Estos métodos consisten en la búsqueda de unos factores en el espacio de los individuos; cada factor corresponde a un grupo. Se les conoce como Análisis Factorial tipo Q.



2.4. Medidas de asociación

El objetivo del análisis cluster consiste en encontrar agrupaciones naturales del conjunto de individuos de la muestra, es necesario definir esas agrupaciones y cómo podemos decir que dos grupos son más o menos similares:

- Como se puede medir la similitud entre dos individuos de la muestra.
- Como se puede evaluar cuando dos clusters pueden ser o no agrupados.

En este punto nos centramos en las posibles funciones que pueden elegirse para medir la similitud entre los grupos que sucesivamente se van formando, distinguiendo primeramente entre:

1. Distancias.
2. Similaridades.

2.4.1. Medidas de asociación entre variables.

Para poder unir variables es necesario tener algunas medidas numéricas que caractericen las relaciones entre variables. La base de trabajo de todas las técnicas cluster es que las medidas numéricas de asociación sean comparables, esto es, si la medida de asociación de una par de variables es 0,72 y el de otro par es 0,59, entonces el primer par está más fuertemente asociado que el segundo. Por supuesto, cada medida refleja asociación en un sentido particular y es necesario elegir una medida apropiada para el problema concreto que se esté tratando.

- Coseno del ángulo de vectores.
- Coeficiente de correlación.
- Medidas para datos binarios o dicotómicos:
    - Medida de Ochiai.
    - Medida Φ.
    - Medidas basadas en coincidencias:
        1. Medida de Russell y Rao.
        2. Medida de parejas simples.
        3. Medida de Jaccard.
        4. Medida de Dice.
        5. Medida de Rogers-Tanimoto.
        6. Medida de Kulcynski
- Medidas basadas en probabilidades condicionadas.

2.4.2. Medidas de asociación entre individuos

Distancia euclídea, de Minkowski y de Mahalanobis.

La distancia euclídea:

$$d_2(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{(x_i - x_j)' (x_i - x_j)} = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}$$

En cuanto a las distancias de Minkowski, éstas proceden de las normas Lp

$$||x_i||_p = \left(\sum_{l=1}^{n} |x_{il}|^p \right)^{\frac{1}{p}} \quad p \geq 1$$

dando origen a:

$$d_p(x_i, x_j) = ||x_i - x_j||_p = \left(\sum_{l=1}^{n} |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}}$$

Algunos casos particulares para valores de p concretos son:

1. Distancia d1 o distancia ciudad (City Block) (p = 1)

$$d_1(x_i, x_j) = \sum_{l=1}^{n} |x_{il} - x_{jl}|$$

2. Distancia de Chebychev o distancia del máximo (p = ∞)

$$d_\infty(x_i, x_j) = \underset{l=1,\ldots,n}{\text{Max}} |x_{il} - x_{jl}|$$

Por otra parte, se puede generalizar la distancia euclídea, a partir de la norma:

$$||x_i||_B = \sqrt{x_i' B x_i}$$

donde B es una matriz definida positiva. La métrica correspondiente a dicha norma es:

$$D_B(x_i, x_j) = \sqrt{(x_i - x_j)' B (x_i - x_j)} = \sqrt{\sum_{l=1}^{n} \sum_{h=1}^{n} b_{lh} x_{il} x_{jh}}$$

definir la distancia de Mahalanobis, para individuos, como:

$$D_S(x_i, x_j) = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

Correlación entre individuos.

Formalmente hablando, el coeficiente de correlación entre vectores de individuos puede ser usado como una medida de asociación entre individuos.

$$\text{Individuo i} \qquad x_i = (x_{i1}, x_{i2}, \ldots, x_{in})'$$

$$\text{Individuo j} \qquad x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$$

$$r_{ij} = \frac{\sum_{l=1}^{n} (x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{s_i s_j}$$

El principal problema de este coeficiente radica en el hecho de que en un vector de datos correspondiente a un individuo hay muchas unidades de medida diferentes, lo cual hace muy difícil comparar las *medias* y las *varianzas*.

Distancias derivadas de la distancia X^2

Se define el estadístico X^2 como:

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

donde p y q son el número de modalidades de las variables estudiadas.

Ahora bien, esta cantidad, que es muy útil para contrastes en tablas de contingencia, no lo es tanto como medida de asociación, puesto que aumenta cuando n crece. Por ello se considera la medida Φ^2, llamada contingencia cuadrática media, definida como

$$\Phi^2 = \frac{\chi^2}{n_{..}}$$

Sin embargo, este coeficiente depende del tamaño de la tabla.

Se han hecho algunos intentos para normalizar la medida al rango [0, 1]. Por ejemplo:

Medida de Tschuprow: $$T = \left(\frac{\Phi^2}{[(p-1)(q-1)]^{\frac{1}{2}}} \right)^{\frac{1}{2}}$$

Medida de Cramer: $$C = \left(\frac{\Phi^2}{\text{Min}(p-1, q-1)} \right)^{\frac{1}{2}}$$

Coeficiente de contingencia de Pearson: $$P = \left(\frac{\Phi^2}{1+\Phi^2} \right)^{\frac{1}{2}} = \left(\frac{\chi^2}{n_{..}+\chi^2} \right)^{\frac{1}{2}}$$

Medidas no métricas: Coeficiente de Bray-Curtis.

El coeficiente de Bray-Curtis viene definido por la expresión:

$$D_{i,j} = \frac{\sum_{l=1}^{n} |x_{il} - x_{jl}|}{\sum_{l=1}^{n} (x_{il} + x_{jl})}$$

MEDIDAS PARA DATOS BINARIOS

Con alguna excepción, las medidas de asociación que se mencionaron para variables de tipo binario pueden ser aplicadas para medir la asociación entre individuos.

2.5. Etapas en el Análisis Cluster

Las etapas a seguir en el empleo de una técnica cluster pueden ser resumidas en los siguientes puntos:

1. Elección de las variables.

La elección inicial del conjunto concreto de características usadas para describir a cada individuo constituye un marco de referencia para establecer las agrupaciones o clusters; dicha elección, posiblemente, refleje la opinión del investigador acerca de su propósito de clasificación. Consecuentemente, la primera cuestión a responder sobre la elección de variables es si son relevantes para el tipo de clasificación que se va buscando. Es importante tener en cuenta que la elección inicial de variables es, en sí misma, una categorización de los datos, para lo cual sólo hay limitadas directrices matemáticas y estadísticas.

La siguiente cuestión que debe considerarse es el número de variables a emplear. En muchas aplicaciones es probable que el investigador se equivoque tomando demasiadas medidas, lo cual puede dar origen a diversos problemas, bien sea a nivel computacional o bien porque dichas variables adicionales oscurezcan la estructura de los grupos.

En muchas aplicaciones las variables que describen los objetos a clasificar no están medidas en las mismas unidades. En efecto, puede haber variables de tipos completamente diferentes, algunas categóricas, otras ordinales e incluso otras que tengan una escala de tipo intervalo.

Es claro que no sería correcto tratar como equivalentes, por ejemplo, el peso medido en kilos, la altura en milímetros y valorar la ansiedad en una escala de cuatro puntos.

Para variables de tipo intervalo, la solución general consiste en tipificar las variables antes del análisis, calculando las desviaciones típicas a partir de todos los individuos. Algunos autores, por ejemplo Fleiss y Zubin (1969), consideran que esta técnica puede tener serias desventajas al diluir las diferencias entre grupos sobre las variables que más discriminen; como alternativa sugieren emplear la desviación estándar entre grupos para tipificar.

Cuando las variables son de tipos diferentes se suele convertir todas las variables en binarias antes de calcular las similaridades. Esta técnica tiene la ventaja de ser muy

clarificadora, pero la desventaja de sacrificar información. Una alternativa más atractiva es usar un coeficiente de similaridad que pueda incorporar información de diferentes tipos de variables de una forma sensible, como el propuesto por Gower en 1971. Asimismo, para variables mixtas existe la posibilidad de hacer un análisis por separado e intentar sintetizar los resultados a partir de los diferentes estudios.

2. Elección de la medida de asociación.

La mayor parte de los métodos cluster requieren establecer una medida de asociación que permita medir la proximidad de los objetos en estudio. Cuando se realiza un Análisis Cluster de individuos, la proximidad suele venir expresada en términos de distancias, mientras que el Análisis Cluster por variables involucra generalmente medidas del tipo coeficiente de correlación, algunas de las cuales tienen interpretaciones en distintos sentidos mientras que otras son difíciles de describir, dado el carácter subjetivo de las mismas.

Destacamos el hecho de estar clasificadas en medidas para variables y para individuos, si bien algunas de ellas pueden considerarse de uso común.

Hay que tener en cuenta, asimismo, la importancia que tienen los tipos de datos a emplear, bien sean estos categóricos o no.

3. Elección de la técnica cluster a emplear en el estudio.

En los métodos jerárquicos las asignaciones de los individuos permanecen estables durante todo el proceso, no permitiendo reasignaciones posteriores a clusters distintos si hubiera lugar a ello, cuestión que sí es factible en los métodos no jerárquicos. Además, en los métodos jerárquicos, el investigador debería sacar sus propias conclusiones mientras que en los procedimientos no jerárquicos el número final de clusters está, por lo general, impuesto de antemano, si bien se han desarrollado, dentro de este tipo de métodos, técnicas que permiten una cierta flexibilidad en el número final de clusters, con el fin de evitar posibles perturbaciones en los resultados definitivos.

Así pues, en algunos problemas prácticos, la elección del método a emplear será relativamente natural, dependiendo, sobre todo, de la naturaleza de los datos usados y de los objetivos finales perseguidos, si bien en otros la elección no será tan clara. Lo que sí es conveniente siempre, a la hora de las aplicaciones prácticas, es no elegir un sólo procedimiento, sino abarcar un amplio abanico de posibilidades y contrastar los resultados obtenidos con cada una de ellas. De este modo, si los resultados finales son

parecidos, podremos obtener unas conclusiones mucho más válidas sobre la estructura natural de los datos.

4. Validación de los resultados e interpretación de los mismos.

Ésta es la última etapa, la más importante, ya que es en ella donde se van a obtener las conclusiones definitivas del estudio.

Son diversos los métodos propuestos para validar un procedimiento cluster. Por ejemplo, cuando se está trabajando con métodos jerárquicos se plantea un problema:

¿En qué medida representa la estructura final obtenida las similitudes o diferencias entre los objetos de estudio?

El argumento más empleado para responder es el empleo del coeficiente de correlación cofenético, propuesto por Sokal y Rohlf en 1962. Dicho coeficiente mide la correlación entre las distancias iniciales, tomadas a partir de los datos originales, y las distancias finales con las cuales los individuos se han unido durante el desarrollo del método. Altos valores de tal coeficiente mostrarán que durante el proceso no ha ocurrido una gran perturbación en lo que concierne a la estructura original de los datos.

En cuanto a los métodos no jerárquicos, la cuestión anterior va perdiendo sentido, mientras que los procedimientos empleados para validar los resultados van encaminados al estudio de la homogeneidad de los grupos encontrados durante el desarrollo del método. Algunos autores han propuesto el empleo de técnicas multivariantes como el análisis multivariante de la varianza (MANOVA), o bien (como BMDP incluye) desarrollar múltiples análisis de la varianza (ANOVA) sobre cada variable en cada cluster.

Estos procedimientos, evidentemente, plantean serios problemas y no deben ser considerados como definitivos. Una técnica usualmente empleada, de tipo remuestreo (Bootstrap), es la de tomar varias submuestras de la muestra original y repetir el análisis sobre cada una. Si tras repetir el análisis sobre ellas se consiguen soluciones aproximadamente iguales, y parecidas a la obtenida con la muestra principal, se puede intuir que la solución obtenida puede ser válida, si bien esto no sería argumento suficiente para adoptar tal decisión. No obstante, este método es más útil empleado de forma inversa, en el sentido de que si las soluciones obtenidas en las diversas submuestras no guardan una cierta similitud, entonces parece evidente que se debiera dudar de la estructura obtenida con la totalidad de la muestra.

3. Objetivos

El objetivo de este trabajo es realizar una introducción al complejo mundo del análisis cluster. A través del trabajo iremos viendo la evolución desde la idea clasificar y agrupar objetos, individuos… Desde la antigüedad con más o menos rigor científico, hasta los métodos que ya son una parte fundamental de los estudios y evolución de muchas ciencias de hoy en día.

Con la ayuda de la teoría exploraremos diversas variantes de las técnicas más comunes de clasificación no supervisada. Finalmente con unos ejemplos prácticos veremos su utilidad en estudios reales, para por último aplicar a una base de datos oceánica y buscar patrones comunes.

4. Procedimiento

4.1. Área de estudio

El área de estudio serán las bases de datos, en concreto un ejemplo en el que veremos 3 métodos de clasificación no supervisada:

Vamos a realizar una prueba con una base de datos del hospital universitario de Wisconsin, esta base de datos "BreastCancer" se encuentra accesible en el UCI Repository Of Machine Learning Databases en:

- ftp://ftp.ics.uci.edu/pub/machine-learning-databases

También contamos con una página de ayuda:

- http://127.0.0.1:30574/library/mlbench/html/BreastCancer.html

En esta página podemos encontrar la descripción, uso, formato, fuente,…

Formato: 699 observaciones en 11 variables, una es carácter variable, 9 son ordinales o nominales y 1 es clase de destino.

El objetivo es relacionar la clase benigna o maligna (que no usaremos en la clasificación), con el cluster realizado con el resto de variables. Cada individuo tiene 9 variables con valores desde 0 a 10.

Trataremos la base de datos para quedarnos con los datos que nos interesen, y a continuación procedemos a realizar los análisis de cluster:

Primero el cluster jerárquico, utilizaremos el método más empleado en la actualidad que es el método de Ward.

Seguidamente el cluster no jerárquico, que será el método *k-means*.

Por último el cluster Fuzzy K-means, una variante difusa del anterior.

Acabaremos viendo las similitudes o no de sus resultados obteniendo al fin 2 grupos para cada tipo de análisis de cluster.

4.2. Software R

R es un lenguaje de programación y un entorno para el análisis estadístico y la realización de gráficos. Debido a su naturaleza es fácilmente adaptable a una gran variedad de tareas. Fue inicialmente escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda. R actualmente es el resultado de un esfuerzo de colaboración de personas del todo el mundo. Desde mediados de 1997 se formó lo que se conoce como núcleo de desarrollo de R, que actualmente es el que tiene la posibilidad de modificación directa del código fuente. Por otra parte, R es un proyecto GNU similar a S, desarrollado este por los Laboratorios Bell. Las diferencias entre R y S son importantes, pero la mayoría del código escrito para S corre bajo R sin modificaciones.

R abarca una amplia gama de técnicas estadísticas que van desde los modelos lineales a las más modernas técnicas de clasificación pasando por los test clásicos y el análisis de series temporales. Proporciona una amplia gama de gráficos que además son fácilmente adaptables y extensibles. La calidad de los gráficos producidos y la posibilidad de incluir en ellos símbolos y fórmulas matemáticas, posibilitan su inclusión en publicaciones que suelen requerir gráficos de alta calidad.

El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL. Además está disponible pre compilado para una multitud de plataformas. La página principal del proyecto es:

http://www.r-project.org.

4.2.2. Instalación en Windows

La descarga de R en el equipo se efectúa desde:

http://cran.es.r-project.org/bin/windows/base/release.htm

Luego se procede con la ejecución, siguiendo las instrucciones. Para la instalación de Rcmdr, se arranca R desde Inicio→Todos los programas→ R. A continuación, Paquetes→Instalar Paquete(s) y elegido el mirror desde el cual se quiere instalar el paquete, por ejemplo Spain (Madrid), se selecciona Rcmdr.

Hemos tenido que instalar los paquetes: *Mlbench, MASS y Sp*.

4.3. Cluster Jerárquico: "WARD"

El método de Ward es un procedimiento jerárquico en que cada paso, se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster.

Lo que se persigue es hacer grupos de mínima variabilidad, es decir, homogéneos.

- x_{ij}^k al valor de la j−ésima variable sobre el i−ésimo individuo del k−ésimo cluster, suponiendo que dicho cluster posee n_k individuos.

- m^k al centroide del cluster *k*, con componentes m_j^k.

- E_k a la suma de cuadrados de los errores del cluster k, o sea, la distancia euclídea al cuadrado entre cada individuo del cluster k a su centroide.

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k)^2 - n_k\sum_{j=1}^{n}(m_j^k)^2$$

- E a la suma de cuadrados de los errores para todos los clusters, o sea, si suponemos que hay h clusters.

$$E = \sum_{k=1}^{h} E_k$$

El proceso comienza con *m* clusters, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del cluster y por lo tanto en este primer paso se tendrá $E_k = 0$ para cada cluster y con ello, $E = 0$. El objetivo del método de Ward es encontrar en cada etapa aquellos dos clusters cuya unión proporcione el menor incremento en la suma total de errores, E.

Supongamos ahora que los clusters C_p y C_q se unen resultando un nuevo clusterC_t. Entonces el incremento de E será

$$\Delta E_{pq} = E_t - E_p - E_q =$$

$$= \left[\sum_{i=1}^{n_t}\sum_{j=1}^{n}(x_{ij}^t)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2\right] - \left[\sum_{i=1}^{n_p}\sum_{j=1}^{n}(x_{ij}^p)^2 - n_p\sum_{j=1}^{n}(m_j^p)^2\right] - \left[\sum_{i=1}^{n_q}\sum_{j=1}^{n}(x_{ij}^q)^2 - n_q\sum_{j=1}^{n}(m_j^q)^2\right] =$$

$$= n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2$$

Ahora bien $n_t m_j^t = n_p m_j^p + n_q m_j^q$

De donde $$n_t^2(m_j^t)^2 = n_p^2(m_j^p)^2 + n_q^2(m_j^q)^2 + 2n_p n_q m_j^p m_j^q$$

Y como $$2m_j^p m_j^q = (m_j^p)^2 + (m_j^q)^2 - (m_j^p - m_j^q)^2$$

Se tiene $$n_t^2(m_j^t)^2 = n_p(n_p + n_q)(m_j^p)^2 + n_q(n_p + n_q)(m_j^q)^2 - n_p n_q(m_j^p - m_j^q)^2$$

Dado que $n_t = n_p + n_q$, dividiendo por n_t^2 se obtiene

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

Con lo cual se obtiene la siguiente expresión de ΔE_{pq}:

$$n_p \sum_{j=1}^{n}(m_j^p)^2 + n_q \sum_{j=1}^{n}(m_j^q)^2 - n_t \sum_{j=1}^{n}\left[\frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2\right]$$

$$\Delta E_{pq} = n_p \sum_{j=1}^{n}(m_j^p)^2 + n_q \sum_{j=1}^{n}(m_j^q)^2 - m_p \sum_{j=1}^{n}(m_j^p)^2 - n_q \sum_{j=1}^{n}(m_j^q)^2 + \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

$$= \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

Así el menor incremento de los errores cuadráticos es proporcional a la distancia euclídea al cuadrado de los centroides de los clusters unidos. La suma *E* es no decreciente y el método, por lo tanto, no presenta los problemas de los métodos del centroide anteriores.

Veamos, para finalizar, cómo se pueden calcular los distintos incrementos a partir de otros calculados con anterioridad.

Sea C_t el cluster resultado de unirC_p y C_q y sea C_r otro cluster distinto a los otros dos. El incremento potencial en E que se produciría con la unión de C_r y C_t es

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^t)^2$$

$$m_j^t = \frac{n_p m_j^p + n_q m_j^q}{n_t}$$
$$n_t = n_p + n_q$$

Ya la expresión

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

Se deduce

$$(m_j^r - m_j^t)^2 = (m_j^r)^2 + (m_j^t)^2 - 2m_j^r m_j^t =$$

$$= (m_j^r)^2 + \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 - 2m_j^r \frac{n_p m_j^p + n_q m_j^q}{n_t} =$$

$$= \frac{n_p(m_j^r)^2 + n_q(m_j^r)^2}{n_t} + \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 -$$

$$- \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 - 2m_j^r \frac{n_p m_j^p + n_q m_j^q}{n_t} =$$

$$= \frac{n_p}{n_t}(m_j^r - m_j^p)^2 + \frac{n_q}{n_t}(m_j^r - m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

Con lo cual

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^t)^2 =$$

$$= \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}\left[\frac{n_p}{n_t}(m_j^r - m_j^p)^2 + \frac{n_q}{n_t}(m_j^r - m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2\right] =$$

$$= \frac{n_r n_p}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^p)^2 + \frac{n_q n_r}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^q)^2 - \frac{n_r n_p n_q}{n_t(n_r + n_t)}\sum_{j=1}^{n}(m_j^p - m_j^q)^2 =$$

$$= \frac{1}{n_r + n_t}\sum_{j=1}^{n}\left[n_r n_p(m_j^r - m_j^p)^2 + n_r n_q(m_j^r - m_j^q)^2 - \frac{n_r n_p n_q}{n_p + n_q}(m_j^p - m_j^q)^2\right] =$$

$$= \frac{1}{n_r + n_t}\left[(n_r + n_p)\Delta E_{rp} + (n_r + n_q)\Delta E_{rq} - n_r \Delta E_{pq}\right]$$

Al igual que en los anteriores métodos del centroide se puede demostrar que la relación anterior se sigue verificando para una distancia que venga definida a partir de una norma que proceda de un producto escalar o que verifique la ley del paralelogramo.

4.3.1. Aplicación con software R

Comenzamos con los siguientes códigos para cargar las diferentes librerías, la base de datos, una ayuda que nos lleva a una página en internet y la cabecera de los datos.

library(mlbench)

library(MASS)

data(BreastCancer)

help(BreastCancer)

head(BreastCancer)

	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	benign
2	1002945	5	4	4	5	7	10	3	2	1	benign
3	1015425	3	1	1	1	2	2	3	1	1	benign
4	1016277	6	8	8	1	3	4	3	7	1	benign
5	1017023	4	1	1	3	2	1	3	1	1	benign
6	1017122	8	10	10	8	7	10	9	7	1	malignant

Pasamos la variable class a numérica: benigno = 0 y Maligno = 1,

BreastCancer$Class<-1*(BreastCancer$Class=="malignant")

	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	0
2	1002945	5	4	4	5	7	10	3	2	1	0
3	1015425	3	1	1	1	2	2	3	1	1	0
4	1016277	6	8	8	1	3	4	3	7	1	0
5	1017023	4	1	1	3	2	1	3	1	1	0
6	1017122	8	10	10	8	7	10	9	7	1	1

Convertimos las variables que son factores a numéricas:

for (i in 2:10) BreastCancer[,i]<-as.numeric(BreastCancer[,i])

Quitamos la variable 1 id: *BreastCancer<-BreastCancer[,2:11]*

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
1	5	1	1	1	2	1	3	1	1
2	5	4	4	5	7	10	3	2	1
3	3	1	1	1	2	2	3	1	1
4	6	8	8	1	3	4	3	7	1
5	4	1	1	3	2	1	3	1	1
6	8	10	10	8	7	10	9	7	1

Guardamos en una nueva variable: bc.class la columna class y por último nos quedamos con las variables a utilizar: *BreastCancer<-BreastCancer[,1:9]*

Ahora empezamos con los detalles del cluster:

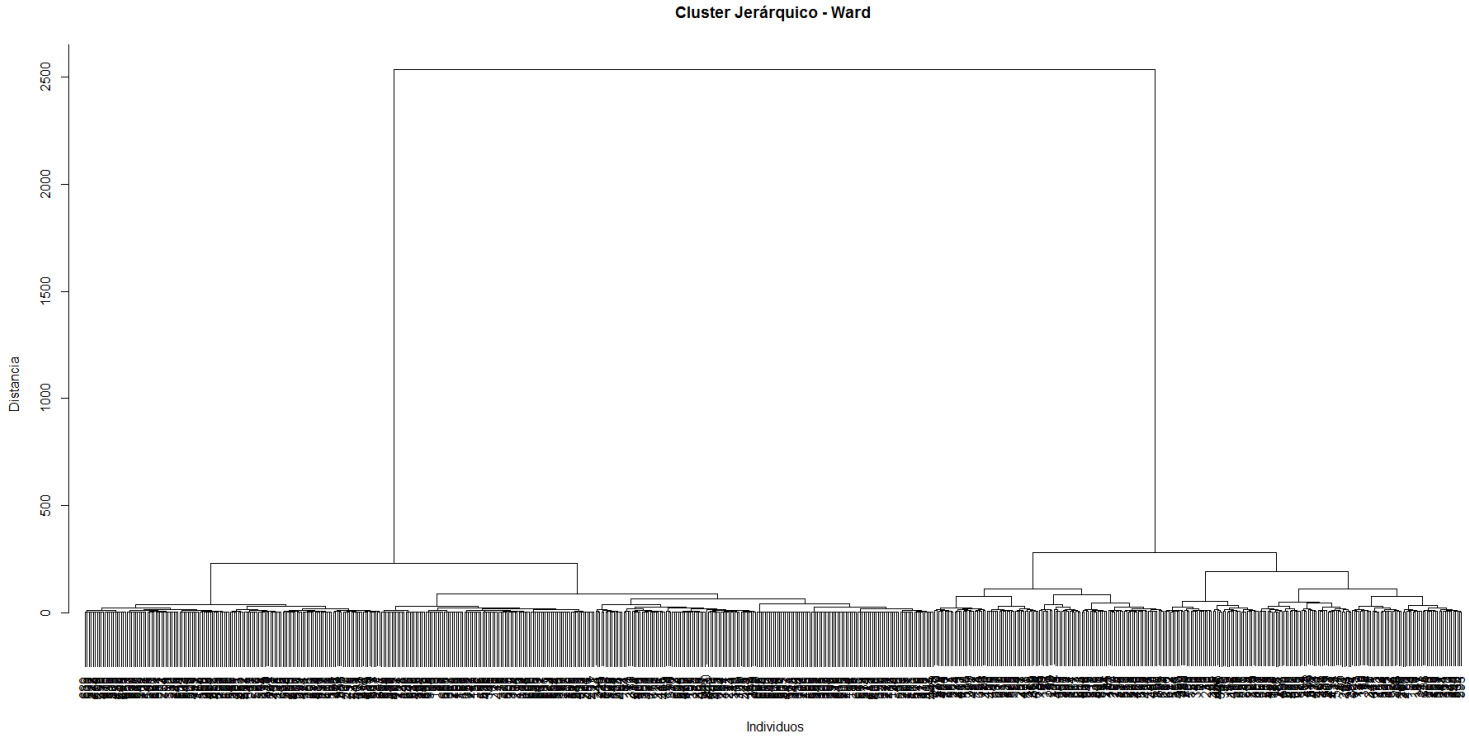
d <- dist(BreastCancer, method = "euclidean"), Matriz de distancias entre individuos la distancia escogida es la Euclidea.

Cluster jerárquico usando el método de Ward:

fit <- hclust(d, method="ward")

Dibujo el Dendograma

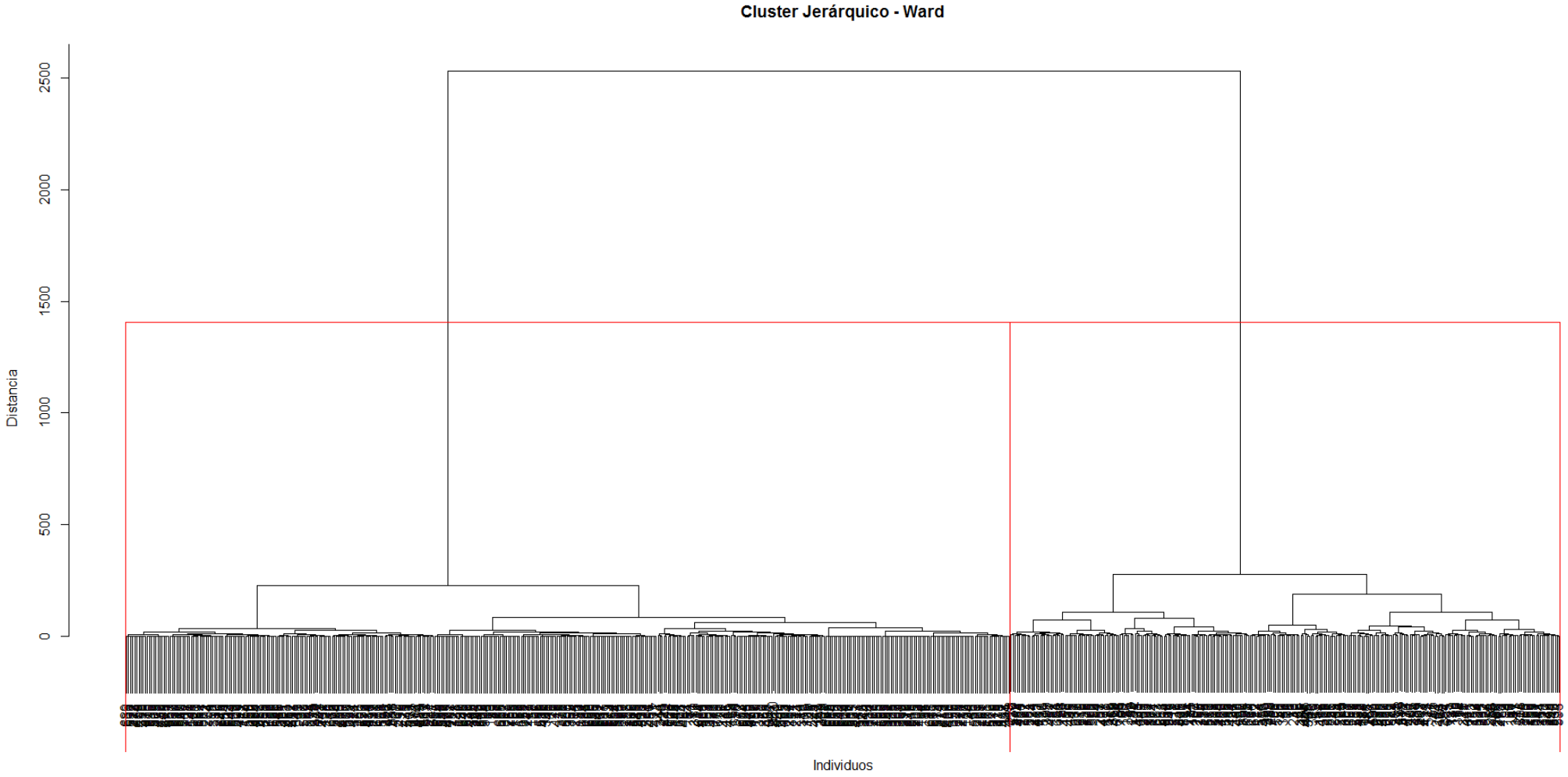
plot(fit,main="Cluster Jerárquico - Ward",xlab="Individuos",ylab="Distancia",sub=NA)



Como se ve en el dendograma hay dos grupos bien diferenciados cuyas líneas verticales representan un gran esfuerzo para realizar la unión (son muy largas). Es aconsejable considerar un corte para dos grupos, ya que son claros. Un corte para 3 grupos estaría muy cerca del corte para 4 o sea que no es un corte claro, el corte para 4 también está cerca del de cinco por ese cúmulo que tienes a la derecha.

Dibujamos el Dendograma con bordes rojos alrededor de los 2 clusters:

rect.hclust(fit, k=2, border="red")



Almacenamos el grupo de pertenencia para 2 clusters. Ahora vamos a comparar los dos clusters con los tipos de cáncer que tenemos almacenados (bc.class) para ver que ha descubierto el análisis cluster.

h.group <- cutree(fit, k=2)

Comparar con la Class original, que guardamos antes como bc.class:

res<-table(bc.class,h.group)

res

prop.table(res,margin=1)

```
> res
           h.group
bc.class      1   2
  benign    420  24
  malignant   1 238

> prop.table(res,margin=1)
           h.group
bc.class             1          2
  benign    0.94594595 0.05405405
  malignant 0.00418410 0.99581590
```

Validación e interpretación:

Observamos una pertenencia muy grande de casi un 95% de los 'benignos' al grupo 1 y de solo un 5,5% al grupo 2, sin embargo respectos a los malignos la pertenencia es casi total al grupo 2 del orden del 99,6%.

El cluster número 1 contiene un 99.78% de 'benignos' y el grupo número 2 un 90.84% de 'malignos'.

Sin emplear la información del grupo de pertenencia, sino empleando sólo el resto de variables con dos grupos estamos "descubriendo" unos grupos pre-existentes. Esto sólo es posible si la información empleada es útil en ese sentido, es decir, tiene que ver con la forma de discernir entre benigno y maligno.

Finalmente queremos saber en qué medida es representativa la estructura final, para ello vamos a utilizar el coeficiente de correlación cofenético:

dfinal=cophenetic(fit)

cor(d,dfinal)

```
> cor(d,dfinal)
[1] 0.7588836
```

El resultado del coeficiente es 0.76, al ser un valor alto, muestra que durante el proceso no ha ocurrido una gran perturbación en lo que concierne a la estructura original de los datos.

4.4. Cluster No Jerárquico: K-means o K-medias

Las dos características claves del k-means:

- La distancia euclídea se usa como una medida y la varianza es usada como una medida de la dispersión de los grupos.
- El número de grupos k es un parámetro de entrada.

Dado un conjunto de observaciones (x1, x2,…, xn), donde cada observación es un vector real de *d* dimensiones, k-means construye una partición de las observaciones en *k* conjuntos (k = n) *E* = {S1, S2,…, *Ek*}

La suma de los cuadrados de las desviaciones sobre un punto de referencia es mínima cuando ese punto de referencia es el centroide del cluster. La suma de los cuadrados de las desviaciones sobre el centroide para el K-ésimo cluster viene dada por:

$$E_k = \sum_{i=1}^{m_k} \sum_{j=1}^{n} (x_{ijk} - \overline{x}_{jk})^2$$

Para una partición dada de un conjunto de individuos en h clusters, la suma de los cuadrados de los errores dentro de los grupos es:

$$E = \sum_{k=1}^{h} E_k$$

y *E* posee un valor característico para dicha partición.

$\sum_{j=1}^{n} (x_{ijk} - \overline{x}_{jk})^2$ es el cuadrado de la distancia euclídea entre el centroide del cluster *K* y el j-ésimo individuo en dicho cluster.

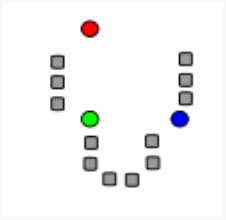
El número de distintas formas en las cuales un conjunto de *m* casos puede ser particionado en *h* clusters es un número de Stirling de segunda especie:

$$\mathbb{S}_n^{(m)} = \frac{1}{m!} \sum_{k=0}^{m} (-1)^{m-k} \binom{m}{k} k^n$$

Por lo que, para no tener que calcular todas esas particiones, consideraremos métodos en los cuales la partición actual es alterada solo si el cambio proporciona una nueva partición con un error total *E* menor. Puesto que cada partición tiene un valor característico *E*, tales métodos no pueden regenerar una partición que haya sido hecha en una etapa anterior y por lo tanto dichos métodos son convergentes.

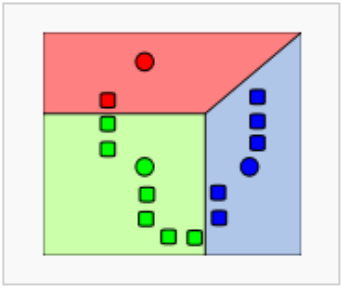
Así pues, un método es convergente si las sucesivas particiones que genera exhiben una sucesión decreciente de valores para *E*.

Demostración:

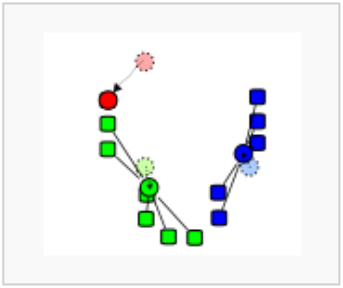


Son generados aleatoriamente dentro de un conjunto de datos (mostrados en color).

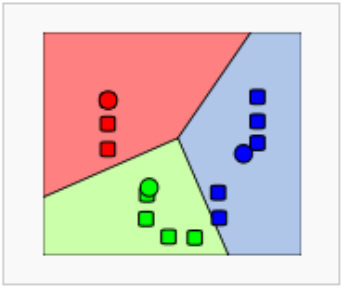
1) *k* centroides iniciales (en este caso *k*=3)



2) *k* grupos son generados asociandole el punto



3) EL centroide de cada uno de los *k* grupos se recalcula.



4) Pasos 2 y 3 se repiten hasta que se logre la convergencia.

Como se trata de un algoritmo heurístico, no hay ninguna garantía de que convergen al óptimo global, y el resultado puede depender de los grupos iniciales. Como el algoritmo suele ser muy rápido, es común para ejecutar varias veces con diferentes condiciones de partida.

4.4.1. Aplicación con software R

Con el mismo ejemplo que el cluster jerárquico nos disponemos a realizar la prueba esta vez con K-means, si lo hacemos de manera independiente debemos volver a cargar librerías, datos, ayuda,… como hicimos anteriormente.

La gran diferencia es que ya especificamos el número de clusters que estamos buscando, indicamos dos centros para que nos busque 2 grupos, ya que a la vista del dendograma parecían distinguirse dos:

km.bc<-kmeans(BreastCancer,centers=2)

Nos indica el tamaño de los grupos (452 y 231), la pertenencia en tanto por ciento a cada tipo de variable, el vector clustering, la suma de los cuadrados por cada cluster y una serie de componentes disponibles que explicaremos a continuación:

km.bc$cluster , número de cluster de pertenencia de cada individuo.

```
> head(km.bc$cluster)
1 2 3 4 5 6
2 1 2 1 2 1
```

km.bc$centers, Centros de los grupos (media de los individuos en cada variable)

```
> km.bc$centers
  Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
1     3.050885  1.296460   1.424779      1.347345     2.095133    1.305310    2.090708        1.250000 1.110619
2     7.164502  6.779221   6.718615      5.731602     5.463203    7.926407    6.095238        6.038961 2.506494
```

km.bc$withinss, Suma de cuadrados dentro de grupos (dividido por *n* sería la varianza dentro de cada grupo). Cuanto más pequeño más compactos son los grupos.

```
> km.bc$withinss
[1]  4298.29 14827.04
```

km.bc$size, Número de individuos asignados a cada grupo.

```
> km.bc$size
[1] 452 231
```

km.bc$totss, Suma total de los cuadrados.

```
> km.bc$totss
[1] 48221.67
```

km.bc$tot.withinss, Suma de cuadrados de los 2 grupos (withinss).

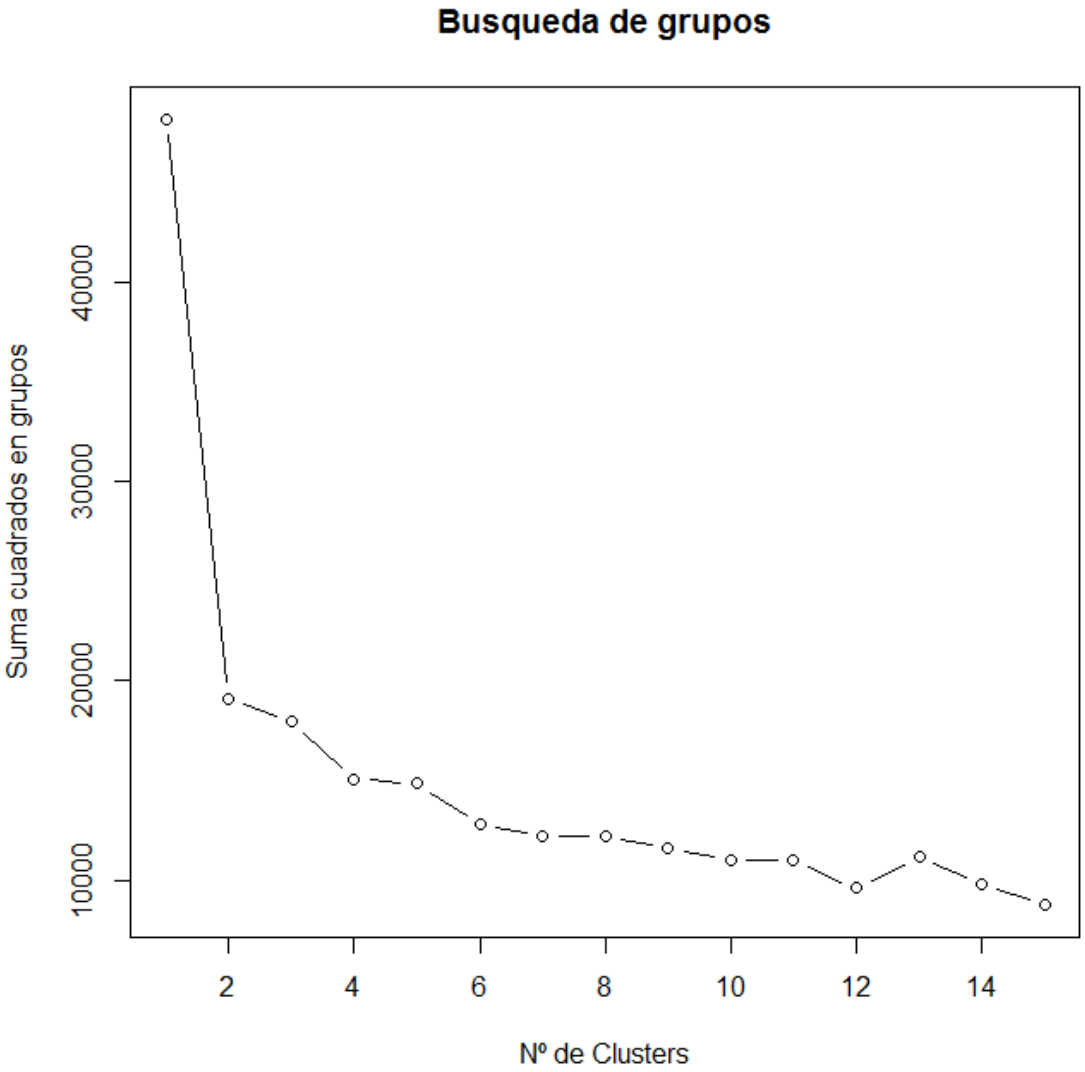
```
> km.bc$tot.withinss
[1] 19125.33
```

km.bc$betweenss, La resta de la suma de cuadrados total menos la suma de cuadrados de los 2 grupos (totss - tot.withinss).

```
> km.bc$betweenss
[1] 29096.34
```

Esta es una forma heurística para tratar de determinar el número de grupos, podemos probar con varios y ver los resultados.

```
wss <- (nrow(BreastCancer)-1)*sum(apply(BreastCancer,2,var))

for (i in 2:15) wss[i] <- sum(kmeans(BreastCancer, centers=i)$withinss)

plot(1:15, wss, type="b", main="Busqueda de grupos",xlab="N de Clusters",
ylab="Suma cuadrados en grupos")
```



Para cada número de clusters se representa el total de la suma de los cuadrados dentro de cada grupo (variabilidad). A mayor número de grupos, menor variabilidad,

así, cuando el número de grupos es igual al número de individuos el valor de la variabilidad será 0.

Por lo tanto, la gráfica, siempre será descendente al aumentar el número de grupos. El objetivo es buscar un punto donde haya un cambio claro de pendiente a partir del cual por añadir un grupo más no se obtiene una gran disminución de esa variabilidad total. En este caso 2 grupos es un claro ejemplo, a partir de ahí no se obtiene gran disminución por poner 3, 4... Grupos.

Elegimos entonces trabajar con 2 grupos, como con el cluster jerárquico:

km.bc<-kmeans(BreastCancer,centers=2)

Podemos tratar de ver la relación entre los grupos encontrados y la clase de tumor, que tenemos guardado en bc.class:

res<-table(bc.class,km.bc$cluster)

res

prop.table(res,margin=1)

```
> res

bc.class      1   2
  benign    435   9
  malignant  17 222

> prop.table(res,margin=1)

bc.class             1          2
  benign    0.97972973 0.02027027
  malignant 0.07112971 0.92887029
```

Como podemos ver el grupo de los tumores 'benignos' tiene una relación del 97.9% con el grupo 1 y un 2% con el 2 grupo del cluster.

Con los 'malignos' pasa al contrario un 7.01% tienen relación con el cluster número 1 y un 92.9% la tienen con el cluster número 2.

Pese a estas relaciones, el método no sirve para clasificar nuevos individuos.

Lo que sí indica esta relación es que las variables sí aportan información sobre la clase. Además, si no dispusiéramos de la información de la "clase" de tumor, el clustering estaría "descubriendo" esa clase subyacente. Luego es tarea del experto/investigador tratar de explicar a que se deben esos grupos.

4.5. Fuzzy K-means o Fuzzy clustering

Es una clase de análisis cluster donde cada individuo tiene un grado de pertenencia difuso a los grupos, es decir no pertenece a un grupo y deja de tener relación con los demás, sino que tiene un porcentaje de cada cluster (se suele dar en tanto por 1).

Este tipo de algoritmos surge de la necesidad de resolver una deficiencia del agrupamiento exclusivo, que considera que cada elemento se puede agrupar inequívocamente con los elementos de su cluster y que, por lo tanto, no se asemeja al resto de los elementos.

Esto se logra representando la similitud entre un elemento y un grupo por una función, llamada función de pertenencia, que toma valores entre cero y uno. Los valores cercanos a uno indican una mayor similitud, mientras que los cercanos a cero indican una menor similitud. Por lo tanto, el problema del agrupamiento difuso se reduce a encontrar una caracterización de este tipo que sea óptima.

Solo funciona para atributos numéricos.

Se han propuesto varios criterios de agrupamiento para obtener la partición difusa óptima para X, pero el más popular hasta el momento está asociado con la función de error mínimo cuadrático:

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d_{ik}^2$$

El valor d_{ik}^2 indica la distancia cuadrada entre los elementos de **X** y los centros de los grupos y puede calcularse utilizando la siguiente fórmula:

$$d_{ik}^2 = ||x_k - v_i||_A^2 = (x_k - v_i)^T A (x_k - v_i)$$

El peso asociado a cada distancia cuadrada, $(u_{ik})^m$, es la *m*-ésima potencia del grado de pertenencia del *k*-ésimo dato al grupo *i*. Cuando $m \to 1$ la partición óptima es cada vez más cercana a una partición exclusiva, mientras que cuando $m \to \infty$ la partición óptima se aproxima a la matriz con todos sus valores iguales a (1/*c*). Los valores de *m* que normalmente se utilizan son valores en el intervalo [1,30]. Cada selección de un valor particular de *m* marca un algoritmo Fuzzy *c*-Means específico.

Teniendo esto en cuenta, el procedimiento general de los algoritmos Fuzzy *c*-Means puede formalizarse en los siguientes pasos:

1. Fijar *c*, *m*, *A* y $||k||_A$. Elegir una matriz inicial $U^{(0)} \in M_{fc}$.
2. Calcular los centros de los grupos con la fórmula $$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}; 1 \leq i \leq c$$.
3. Actualizar la matriz de partición difusa U = [u_{ik}] con: $$u_{ik} = (\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}})^{-1}; 1 \leq k \leq n; 1 \leq i \leq c$$.
4. Si se alcanzó el criterio de parada, finaliza. En caso contrario, regresar al paso 2.

Algunos de los criterios de parada más utilizados son:

- Un número máximo de iteraciones.
- Que la variación en la matriz U sea muy pequeña.

4.5.1. Aplicación con software R

Con el mismo ejemplo que el cluster jerárquico nos disponemos a realizar la prueba esta vez con fuzzy, si lo hacemos de manera independiente debemos volver a cargar librerías, datos, ayuda,… como hicimos anteriormente.

library(cluster, pos=4)

ff<-fanny(BreastCancer,k=2)

ff

Vemos los detalles del cluster:

```
> ff
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.        2
objective     1543.161
tolerance        1e-15
iterations          19
converged            1
maxit              500
n                  683
Membership coefficients (in %, rounded):
    [,1] [,2]
1     84   16
2     34   66
3     89   11
4     34   66
5     83   17
6     23   77
7     55   45
8     89   11
9     76   24
10    89   11
```

Podemos observar detalles, número de grupos, tolerancia para la parada, número de iteraciones… Pero sobre todo lo interesante y diferente de los otros clusters es que cada individuo tiene un porcentaje de pertenencia a cada grupo. Ejemplo:

Individuo 1: [1] 84, [2] 16 → Para el grupo 1.

Individuo 2: [1] 34, [2] 66 → Para el grupo 2.

```
Closest hard clustering:
  1   2   3   4   5   6   7   8   9  10
  1   2   1   2   1   2   1   1   1   1
 61  62  63  64  65  66  67  68  69  70
  2   1   2   2   1   2   1   2   2   1
119 120 121 122 123 124 125 126 127 128
  1   1   1   1   2   2   2   1   2   1
```

Esta es la distribución cuando cada individuo tiene que ir a un cluster, dependiendo de a qué grupo pertenezca más.

Podemos tratar de ver la relación entre los grupos “hard” y la clase de tumor, que tenemos guardado en bc.class:

res<-table(bc.class,ff$cluster)

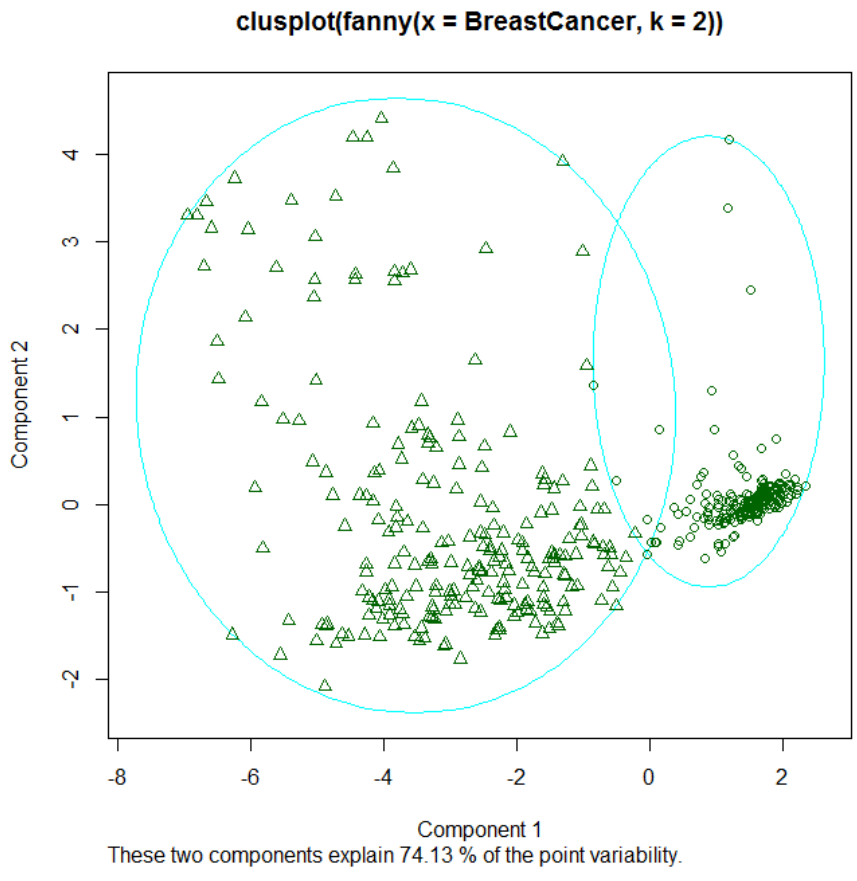
res

prop.table(res,margin=1)

```
> res

bc.class      1   2
  benign    432  12
  malignant   9 230

> prop.table(res,margin=1)

bc.class             1          2
  benign    0.97297297 0.02702703
  malignant 0.03765690 0.96234310
```

Observamos la relación del grupo 1 con los tumores ‘benignos’ de un 97.3% y del grupo 2 con los ‘malignos’ de un 96%.

plot(ff, ask = FALSE, which.plots = NULL, nmax.lab = 40, max.strlen = 5, data = x$data, dist = NULL,stand = FALSE, lines = 2,)



4.6. Diferencias entre ambos métodos

La primera diferencia es que en el cluster jerárquico no sabemos los grupos que se vamos a obtener, en el k-means y en fuzzy k-means por el contrario nosotros le indicamos el número de grupos, aunque podemos buscar el número de grupos gracias al comando $*withinss* en k-means.

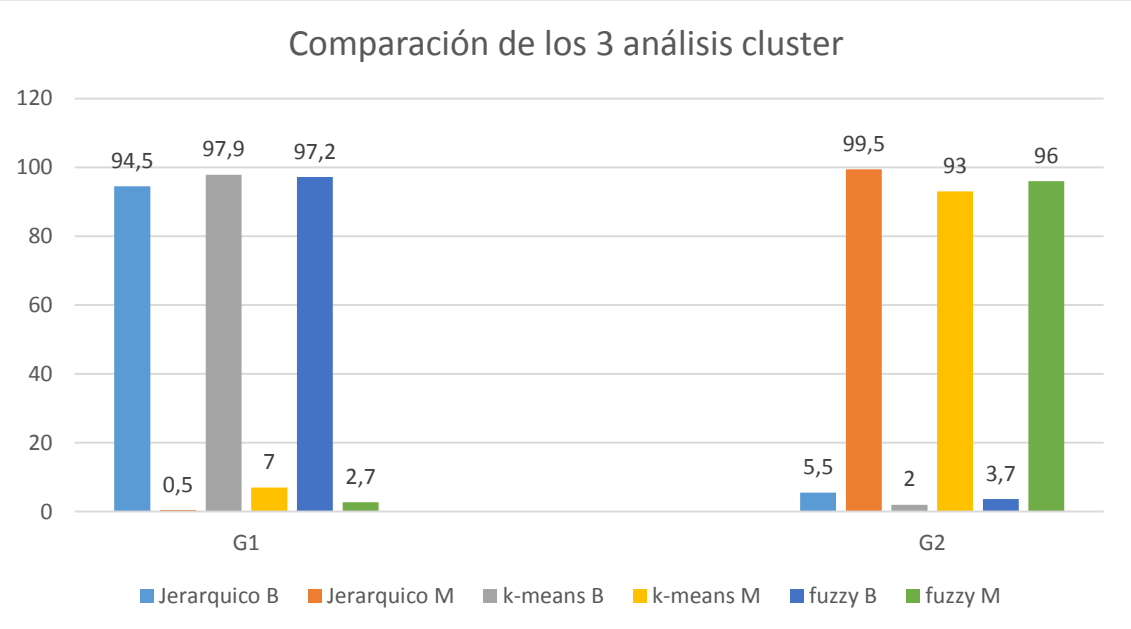
Aquí observamos la primera diferencia entre los tipos de clusters, en el tamaño de los grupos número de individuos varía:

En cluster jerarquizado son 421 y 262,

Y en cluster K-means son 452 y 231.

En Fuzzy k means son 441 y 242.

La relación de los clusters con la class original, que no se ha usado en la clasificación, es similar pero tampoco es igual, como vemos a continuación:



	Jerárquico		k-means		fuzzy	
	B	M	B	M	B	M
G1	94,5	0,5	97,9	7	97,2	2,7
G2	5,5	99,5	2	93	3,7	96

4.7. Ocean Data View

http://odv.awi.de/

Ocean Data View (ODV) es un paquete de software para el análisis y visualización de datos oceanográficos y otros datos georreferenciados. El ODV se puede ejecutar en Windows, Mac OS X, Linux y UNIX.

Además de este software en la página web disponemos de varios apartados en los que tenemos datos, documentación, links, contacto y un foro.

A nosotros nos interesa el apartado de datos, aquí encontraremos bases de datos para nuestro propósito de usarlas para descubrir patrones, grupos que a simple vista no vemos y quizás puedan resultar interesantes.

ODV es utilizado actualmente por más de 25.000 científicos de los principales institutos de investigación en todo el mundo. El proyecto de la UNESCO Ocean Teacher emplea ODV como una de sus principales herramientas de análisis y visualización.

Las bases de datos disponibles vienen en formato *.ODV, estos archivos funcionan en el programa antes comentado. Este programa de uso libre está ya en su 4ª versión y recibe constantes actualizaciones, la última en este 2014.

Como queremos un listado de la base de datos en formato *.ascii o *.txt y los archivos originales están en otro formato, abriremos el programa, cargamos la base de datos, seleccionamos las variables que nos interesen y después exportamos a un *.txt.

Ahora con este nuevo archivo ya podemos trabajar en R.

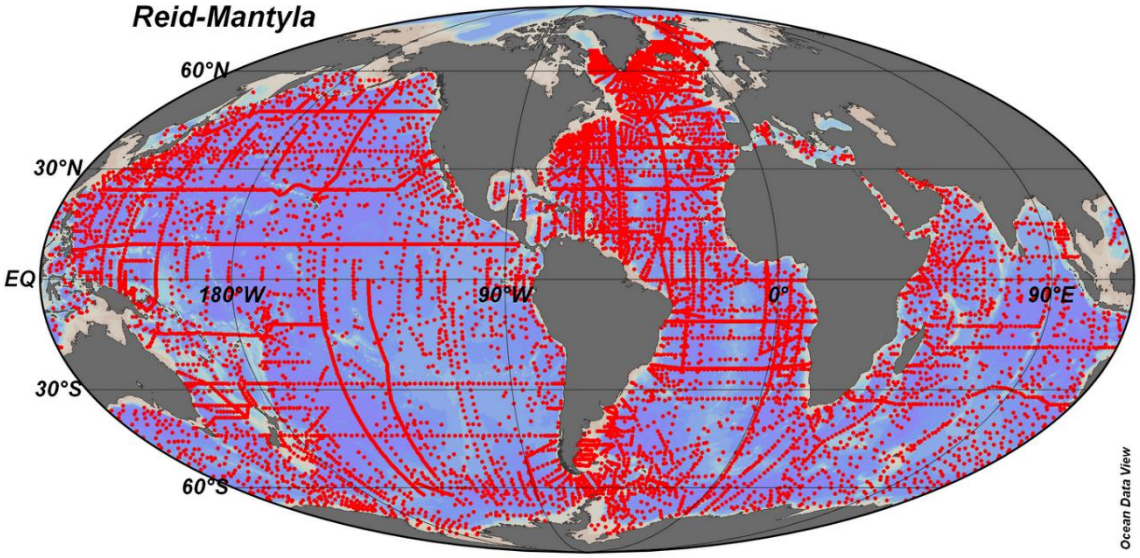
5. Resultados

Con esta base de datos vamos a realizar una prueba, a ver si conseguimos descubrir algún patrón en los datos:

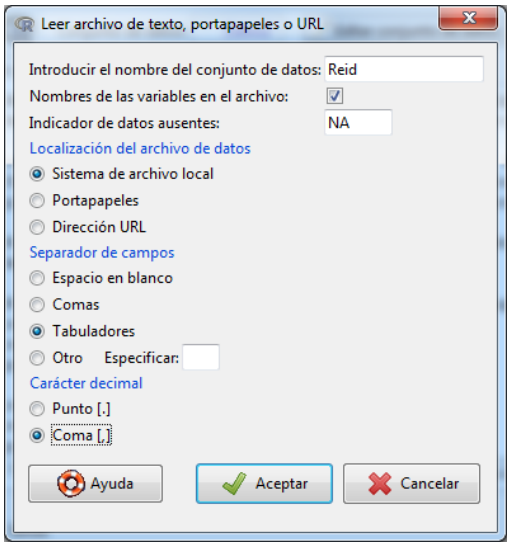
Reid & Mantyla, es una base de datos con cerca de 10.000 estaciones distribuidas a lo largo de los océanos del mundo. Está disponible para cualquier persona que quiera utilizarla, tiene 8 variables con una alta disponibilidad de casi todas ellas:

- Profundidad, temperatura, salinidad, oxígeno, fosfato, silicato, nitrato, nitrito.

Estos datos están recopilados a lo largo del tiempo y de muchas fuentes, aunque la mayoría provienen del NODC, National Oceanographic Data Center, y se recomienda su utilización para estudios globales mejor que para locales.



Abrimos R e importamos los datos desde archivo de texto:



El conjunto de datos Reid tiene 10111 filas y 14 columnas.

Primero vamos a limpiar la base de datos, excluimos los individuos con variables vacías y nos quedamos con que tengan todas, se disminuirá mucho el conjunto. Esto es una deficiencia de estos métodos, los datos "perdidos" no permiten comparar individuos directamente para el cluster por lo que aquellos que tienen algún dato perdido es preferible excluirlos del análisis. También podrían emplearse técnicas estadísticas para "rellenar" esos huecos, considerando para ello que hay cierta "estructura", pero podría tergiversar el análisis.

Tenemos ahora unas 2500 filas y 14 columnas.

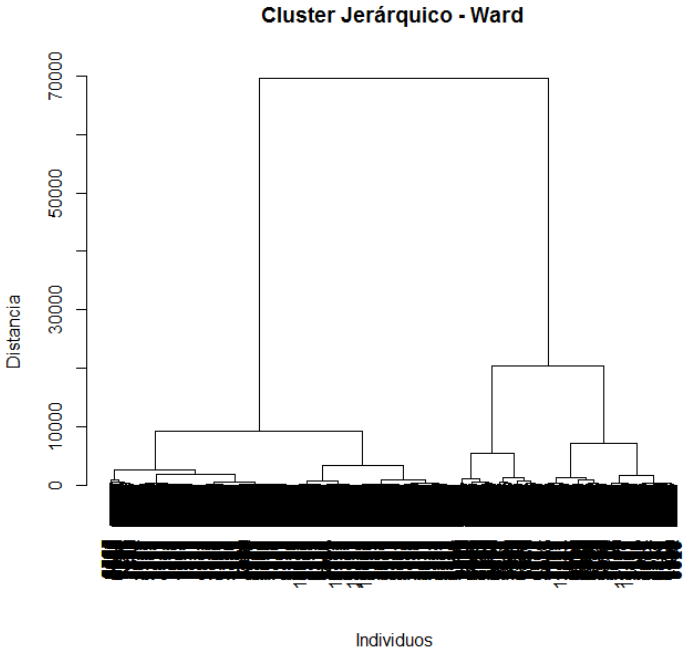
Lo siguiente es pensar que variables vamos a utilizar y quedarnos solo con ellas, en la prueba utilizaremos los componentes del agua medidos en cada estación, están en la misma unidad: $\mu mol/l$ salvo el Oxígeno que es *ml/l.* La conversión para el O_2, 1 ml/l = $10^3/22.391$ = 44.661 *µmol/l*.

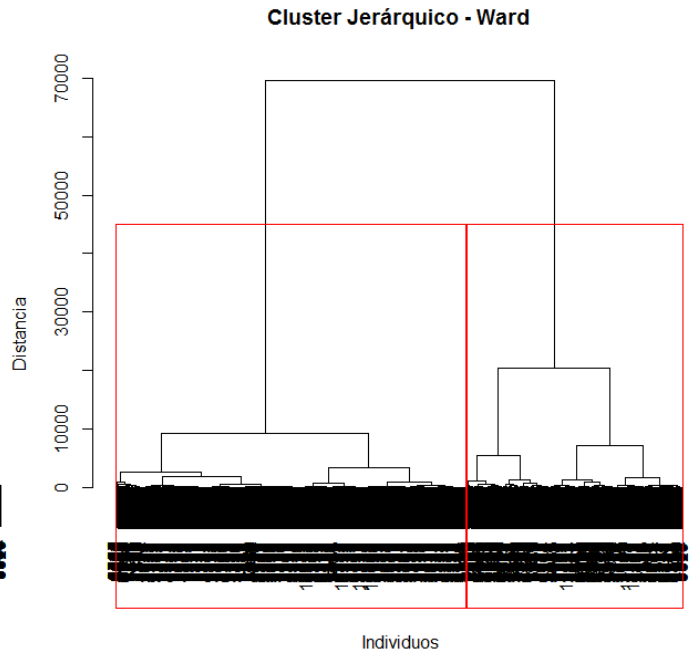
Ahora disponemos de 5 variables y 2528 individuos para realizar el análisis cluster.

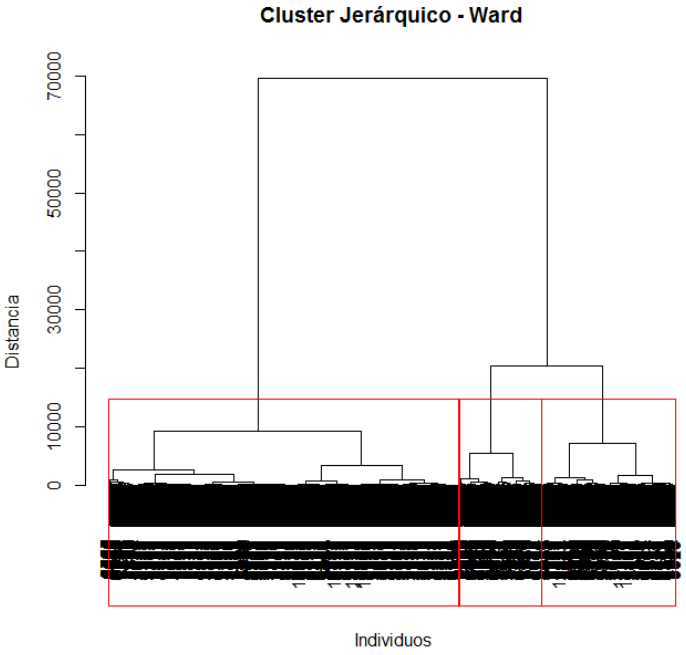
5.1. Cluster Jerárquico

Probamos con distintos grupos de clusters, y vamos observando en los distintos dendogramas de la página siguiente y como los grupos que forma se van formando descienden de manera jerarquizada, al contrario de cómo se forman partiendo desde que todos los individuos son grupos hasta juntarlos en uno solo.

Dendogramas con los clusters diferenciados, k=1, k=2 y k=3.







Vamos a escoger los 2 clusters más representativos que son los análisis con 2 y 3 grupos para continuar y poder compararlos más tarde.

Esa comparación será con otras variables anteriores, como la latitud y la longitud, queremos descubrir si hay zonas donde estos elementos proliferan más.

Para ello vamos a aislar las 2 variables y convertirlas en binarias para que nos muestren:

Latitud: - Norte > 0.

- Sur los restantes o sea los negativos.

Longitud: - Oeste <0.

- Este los restantes o sea los positivos.

Esta clasificación es bastante grosera pero para un primer contacto con los análisis cluster puede darnos algún dato interesante, finalmente:

Latitud: - Norte = 1.

- Sur = 0.

Longitud: -Oeste = 1.

- Este = 0.

El resultado es el siguiente para la LATITUD:

```
> res
        Lat
h.group   0   1
      1 437 527
      2 651 913

> prop.table(res,margin=1)
        Lat
h.group         0         1
      1 0.4533195 0.5466805
      2 0.4162404 0.5837596
```

La relación con la latitud del modo que lo planteamos no nos muestra ningún “descubrimiento” ya que hay casi tantos puntos al Sur como al Norte del Ecuador.

Con 3 clusters:

```
> res
       Lat
h.group   0   1
      1 226 143
      2 211 384
      3 651 913

> prop.table(res,margin=1)
       Lat
h.group         0         1
      1 0.6124661 0.3875339
      2 0.3546218 0.6453782
      3 0.4162404 0.5837596
```

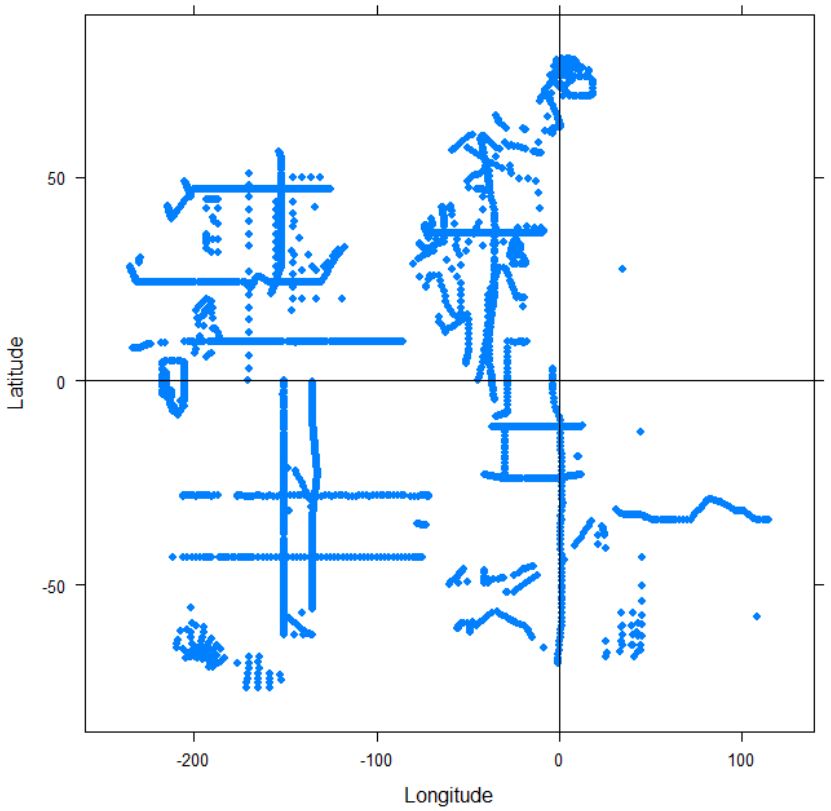
Al igual que el análisis anterior con 2 clusters, con 3 las relaciones siguen parecidas. Por lo que comprendemos que esta no será una fuente de "descubrimiento".

Ahora vamos a ver los resultados de la LONGITUD:

```
> res<-table(h.group,Long)

> res
       Long
h.group    0    1
      1  167  797
      2  156 1408

> prop.table(res,margin=1)
       Long
h.group          0          1
      1 0.17323651 0.82676349
      2 0.09974425 0.90025575
```

Al contrario que en la Latitud, en la Longitud todos los grupos son bastantes parecidos y tienen todos una mayor predominancia por el Oeste. Estos son un 82% a un 90%.

Lo que nos indica que podría ser que la mayoría de los datos fueron tomados al Oeste del meridiano de Greenwich, vamos a ver un gráfico con sus latitudes y longitudes:



Podemos ver claramente que los individuos pertenecientes al Este son menos y obteniendo la cifra exacta en R:

Este = 323 individuos, un 13% del total.

Oeste= 2205 individuos, un 87% del total.

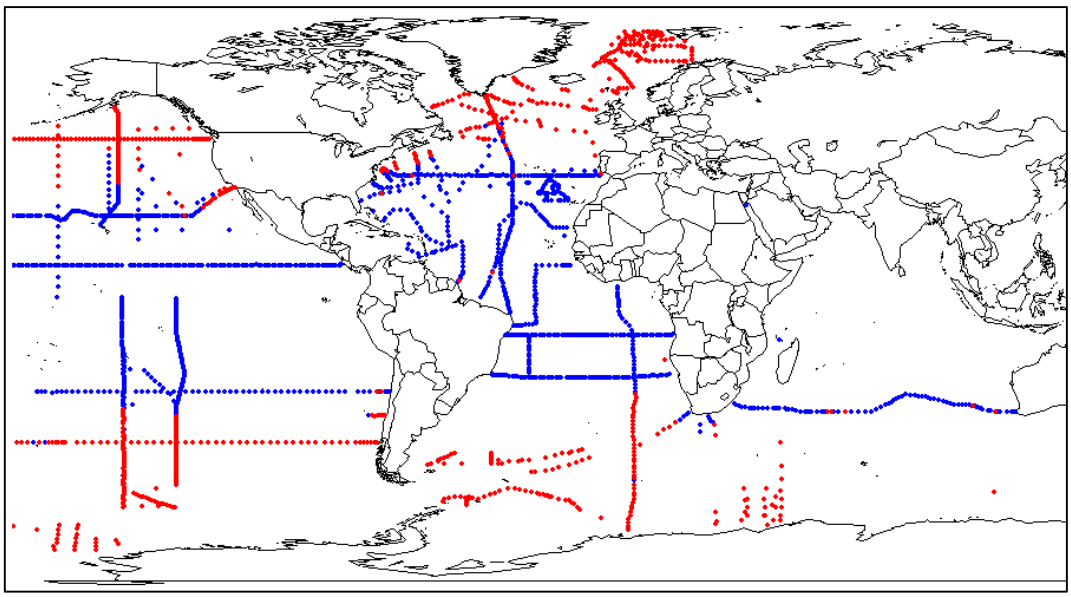
Finalmente queremos saber en qué medida es representativa la estructura final, para ello vamos a utilizar el coeficiente de correlación cofenético:

```
> cor(d,dfinal)
[1] 0.7012272
```

El resultado del coeficiente es 0.70, al ser un valor alto, muestra que durante el proceso no ha ocurrido una gran perturbación en lo que concierne a la estructura original de los datos.

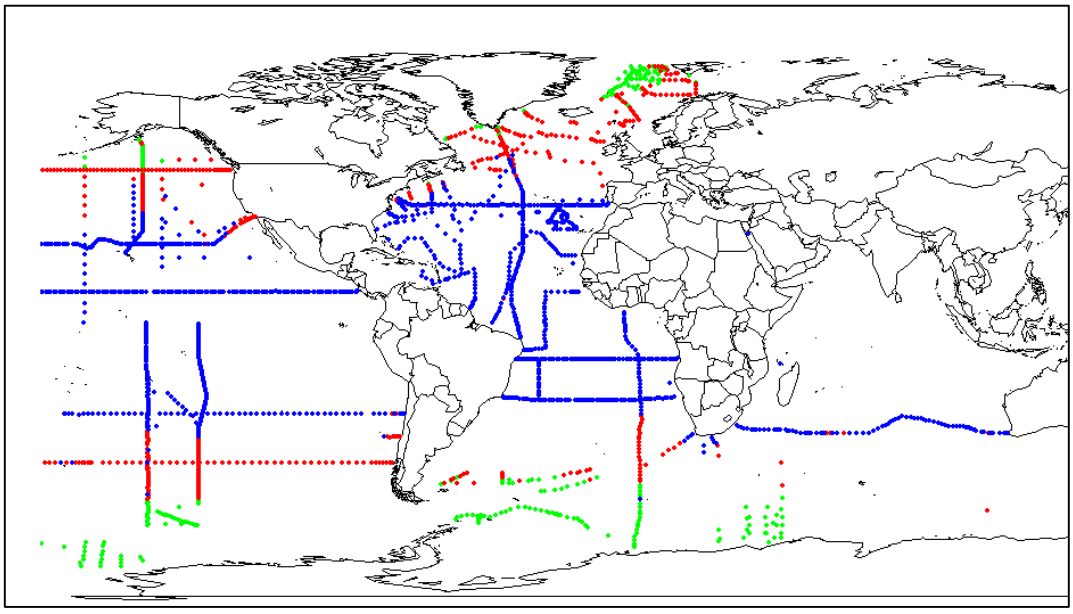
Vamos a proyectar los puntos en un mapa para ver visualmente con 2 y 3 clusters su distribución en un mapa mundo, quizás así descubramos algo interesante:

- **2 clusters:**



Vemos la clara división entre el grupo rojo que está en las zonas polares y el grupo azul que se encuentra entre las latitudes medias y ecuatoriales.

- **3 clusters:**



Nuevamente como el en cluster de 2 grupos vemos una clara división, el grupo verde solo se encuentra cerca de los polos, los del grupo rojo en unas latitudes también muy altas, más de 60º y el grupo azul que se mantiene en el resto.

5.1.1. Código utilizado en R:

```
setwd("C:/Users/kine/Desktop/Dropbox/Master_iñaki/T_F_M/R")

library(mlbench)

library(MASS)

Reid <- read.table("C:/Users/kine/Desktop/Dropbox/Master_iñaki/T_F_M/R/Reid-Mantyla/Reid-Mantyla.txt", header=TRUE, sep="\t", na.strings="NA", dec=",", strip.white=TRUE)

Reid2<-na.exclude(Reid)

Reid2$Oxigeno..umol.l <- with(Reid2, 1*(Oxygen..ml.l.*44.661))

Reid2$Long <- with(Reid2, Longitude<0)

Reid2$Lat <- with(Reid2, Latitude<0)

Reid2$Long<-1*(Reid2$Long=="TRUE")

Reid2$Lat<-1*(Reid2$Lat=="FALSE")

Lat<-Reid2$Lat

Long<-Reid2$Long

Reid3<-Reid2[,11:15]

d <- dist(Reid3, method = "euclidean")

fit <- hclust(d, method="ward")

plot(fit,main="Cluster Jerárquico - Ward",xlab="Individuos",ylab="Distancia",sub=NA)

rect.hclust(fit, k=2, border="red")

h.group <- cutree(fit, k=2)
```

```
# Validación cofenético.
dfinal=cophenetic(fit)
cor(d,dfinal)
# Comparamos con la latitud y longitud.
res<-table(h.group,Lat)
res
prop.table(res,margin=1)

res<-table(h.group,Long)
res
prop.table(res,margin=1)
```

De nuevo con 3 clusters pero solo comparamos con la Latitud.

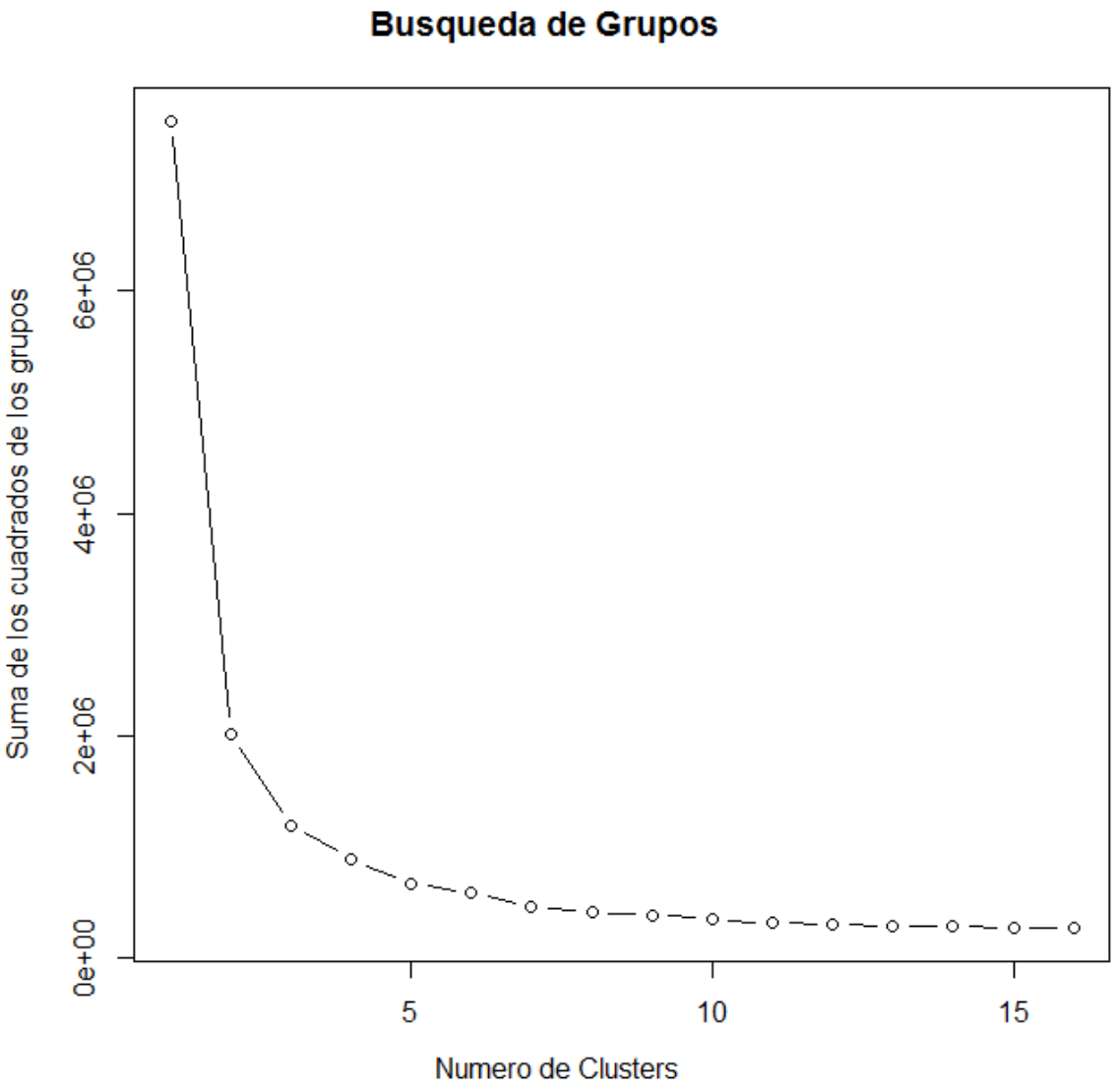
Por ultimo dibujamos los grupos en un mapa mundo:

```
Lat<-Reid2$Latitude
Long<-Reid2$Longitude

library(mapproj)
map("world")



out=mapproject(Long[h.group==1], Lat[h.group==1], projection="", parameters=NULL,
orientation=NULL)
x<-out$x
y<-out$y

points(x,y,col="red",pch=16,cex=.6)
```

5.2. Cluster No Jerárquico: K-means

Vamos a repetir los grupos del anterior. En este análisis cluster tenemos que marcar a priori el número de grupos que estamos buscando. Aunque podemos tratar de determinar ese número de grupos de forma heurística.

Lo vemos a continuación con un Gráfico:



Cuanto más baja sea la suma de los cuadrados más compacto será el grupo, es decir los individuos dentro del grupo tienen unas características parecidas.

Observamos como desde 1 a 3 clusters la bajada de esta suma (withinss) es más pronunciada, y como luego continua pero de forma más estable.

Seguimos con la práctica en R y nos aprovechamos del trabajo realizado anteriormente a la hora de limpiar y preparar la base de datos a utilizar, ya que trabajaremos con los mismos datos, así como de tener las variables para comparar ya guardadas.

Primero con un análisis cluster de 2 grupos:

```
> km.bc$size
[1] 1837  691

> km.bc$withinss
[1]  995406.3 1022280.8

> km.bc$centers
  Phosphate..umol.l. Silicate.umol.l. Nitrate..umol.l. Nitrite..umol.l. Oxigeno..umol.l
1          0.2019162         2.234567         2.099455       0.02140376        228.5267
2          1.0857453        20.055572        14.542836       0.13534009        331.0188
```

Aquí vemos alguna de las características que podemos encontrar en este tipo de cluster, como pudimos ver en el apartado 4.4.1., estos en concreto nos dan información del tamaño de los grupos, de la suma de los cuadrados en cada grupo y de los centros de cada grupo dependiendo del tipo de variable.

Ahora vamos a comparar estos grupos con la LATITUD, para intentar obtener algún tipo de relación:

```
> res
   Lat
      0    1
  1  776 1061
  2  312  379

> prop.table(res,margin=1)
   Lat
            0         1
  1 0.4224279 0.5775721
  2 0.4515195 0.5484805
```

Los resultados vuelven a mostrarnos que hay casi un 50-50 en los 2 hemisferios y que esta relación no dice nada, veremos luego cuando proyectemos los grupos en el mapa.

Seguimos con el análisis de 3 grupos:

```
> km.bc$size
[1] 1615  385  528

> km.bc$withinss
[1] 558556.9 406040.9 233525.2

> km.bc$centers
  Phosphate..umol.l. Silicate.umol.l. Nitrate..umol.l. Nitrite..umol.l. Oxigeno..umol.l
1          0.1715604         2.144396         1.980681       0.01646978        222.9350
2          1.2804675        30.207532        18.411948       0.15654545        353.7453
3          0.6650000         5.435985         6.853030       0.08706439        288.4576
```

Primero vemos los tamaños de los grupos que son bastante dispares, ya que un grupo cuenta con 6 individuos y el otro con 904 individuos.

Seguidamente veremos la relación de los grupos con la LATITUD:

```
> res
   Lat
      0   1
  1 672 943
  2 230 155
  3 186 342

> prop.table(res,margin=1)
   Lat
             0         1
  1 0.4160991 0.5839009
  2 0.5974026 0.4025974
  3 0.3522727 0.6477273
```

Comparamos con la LONGITUD:

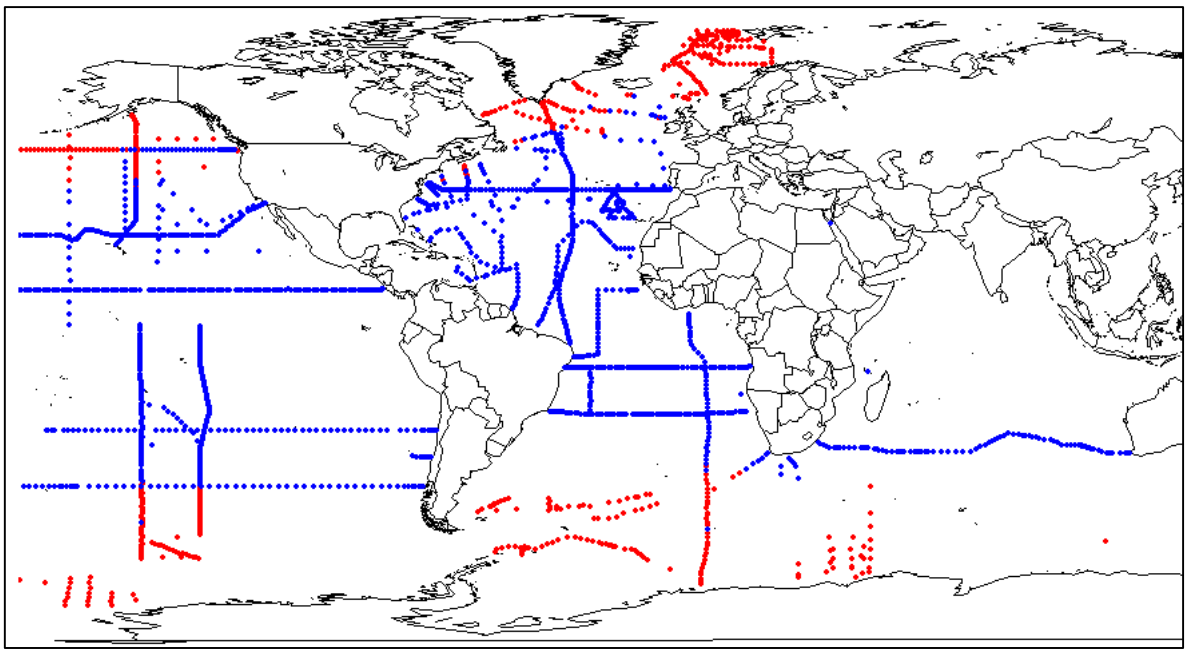
```
> res
   Long
       0    1
  1  176 1661
  2  147  544

> prop.table(res,margin=1)
   Long
              0          1
  1 0.09580838 0.90419162
  2 0.21273517 0.78726483
```

Aquí volvernos a ver algo atípico y es que, como pasaba en el análisis cluster anterior, todos los 2 grupos tienen predominancia al Oeste.

Al igual que dijimos anteriormente la causa se debe a la mayoría de los datos están tomados al Oeste del Meridiano de Greenwich.

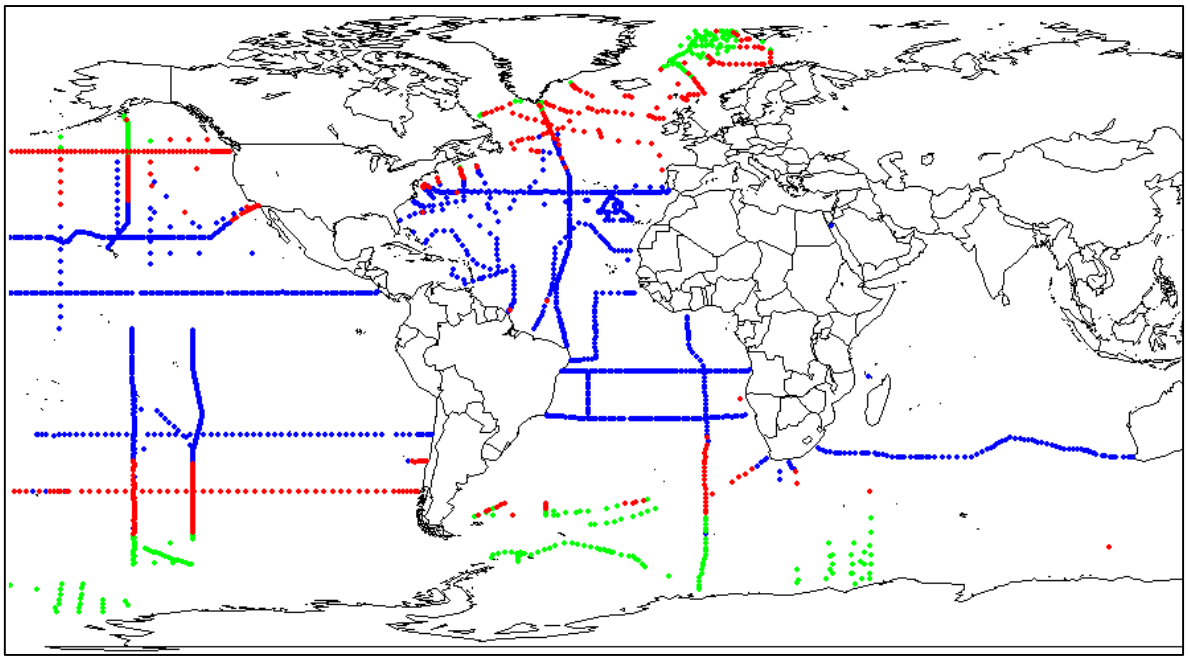
Vamos a proyectar los puntos en un mapa para visualizar con 2 y 3 clusters su distribución en un mapa mundo:

- **2 Clusters:**



Como en el análisis anterior volvemos a ver la misma relación un grupo en la zona polar y el otro en las latitudes medias y ecuatoriales.

- **3 Clusters:**



Igual que el anterior, los 3 grupos están repartidos de la misma manera.

5.2.1. Código utilizado en R:

Igual que el anterior

setwd("C:/Users/kine/Desktop/Dropbox/Master_iñaki/T_F_M/R")

…

Forma heurística de tratar de determinar el número de grupos

*wss <- (nrow(Reid3)-1)*sum(apply(Reid3,2,var))*

for (i in 2:16) wss[i] <- sum(kmeans(Reid3, centers=i)$withinss)

plot(1:16, wss, type="b", xlab="Numero de Clusters", ylab="Suma de los cuadrados de los grupos",main="Busqueda de Grupos")

Le decimos que nos divida en 2 grupos, y que nos muestre los siguientes datos

km.bc<-kmeans(Reid3,centers=2)

km.bc$size

km.bc$withinss

km.bc$centers

Finalmente veremos las relaciones entre los clusters y la Latitud y Longitud:

res<-table(km.bc$cluster,Lat)

res

prop.table(res,margin=1)

res<-table(km.bc$cluster,Long)

res

prop.table(res,margin=1)

Repetimos de nuevo pero esta vez con 3 clusters pero solo comparamos con la Latitud.

Por ultimo dibujamos los grupos en un mapa mundo:

```
library(mapproj)

map("world")


out=mapproject(Long[km.bc$cluster==1], Lat[km.bc$cluster==1], projection="",
parameters=NULL, orientation=NULL)

x<-out$x

y<-out$y


points(x,y,col="red",pch=16,cex=.6)
```

Cambiamos *km.bc$cluster==1,* por 2 o 3 y también la salida de color.

5.3. Fuzzy K-means

Seguimos con la práctica en R y nos aprovechamos del trabajo realizado anteriormente a la hora de limpiar y preparar la base de datos a utilizar, ya que trabajaremos con los mismos datos, así como de tener las variables para comparar ya guardadas.

Cargamos una nueva librería:

library(cluster)

Primero con un análisis cluster de 2 grupos:

```
> ff
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.        2
objective     29567.31
tolerance        1e-15
iterations          25
converged            1
maxit              500
n                 2528
Membership coefficients (in %, rounded):
      [,1] [,2]
1252    84   16
1257    54   46
1270    26   74
1273    73   27
1277    62   38
1278    87   13
1281    83   17
1282     7   93
1283     5   95
```

Observamos el grado de pertenencia de cada individuo dependiendo del grupo, así como varias características, comentadas en el punto 4.5.1.

```
> res
   Lat
      0   1
  1 369 450
  2 719 990

> prop.table(res,margin=1)
   Lat
             0         1
  1 0.4505495 0.5494505
  2 0.4207139 0.5792861
```

Las relaciones con la Latitud, como en los anteriores no marcan ninguna tendencia a reseñar.

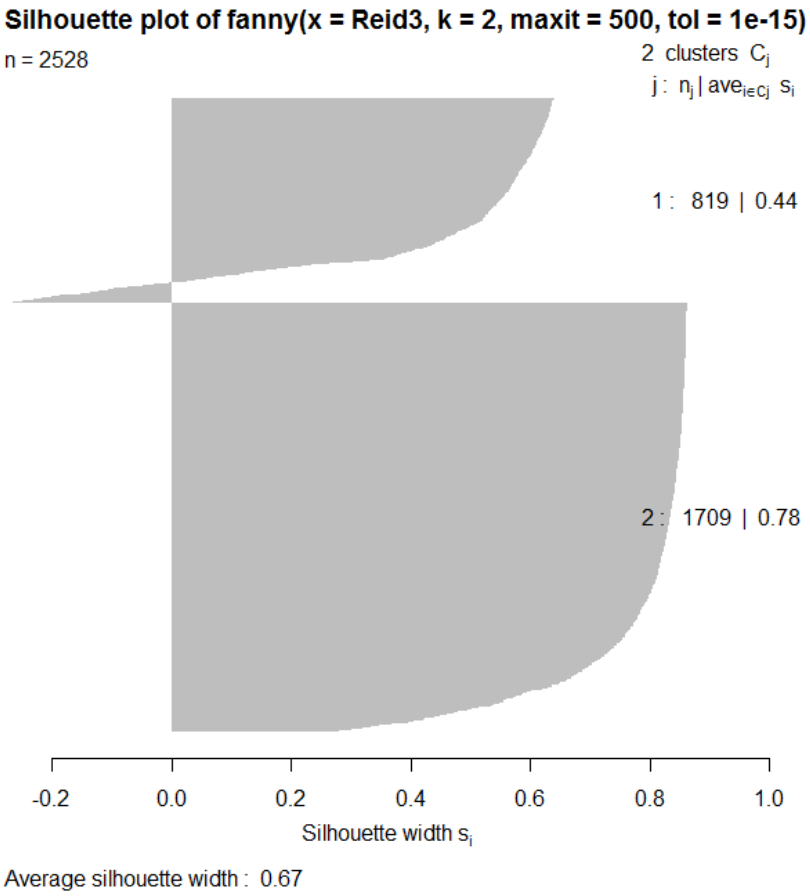


Grafico Silueta de los 2 clusters, vemos el promedio de silueta y el personal.

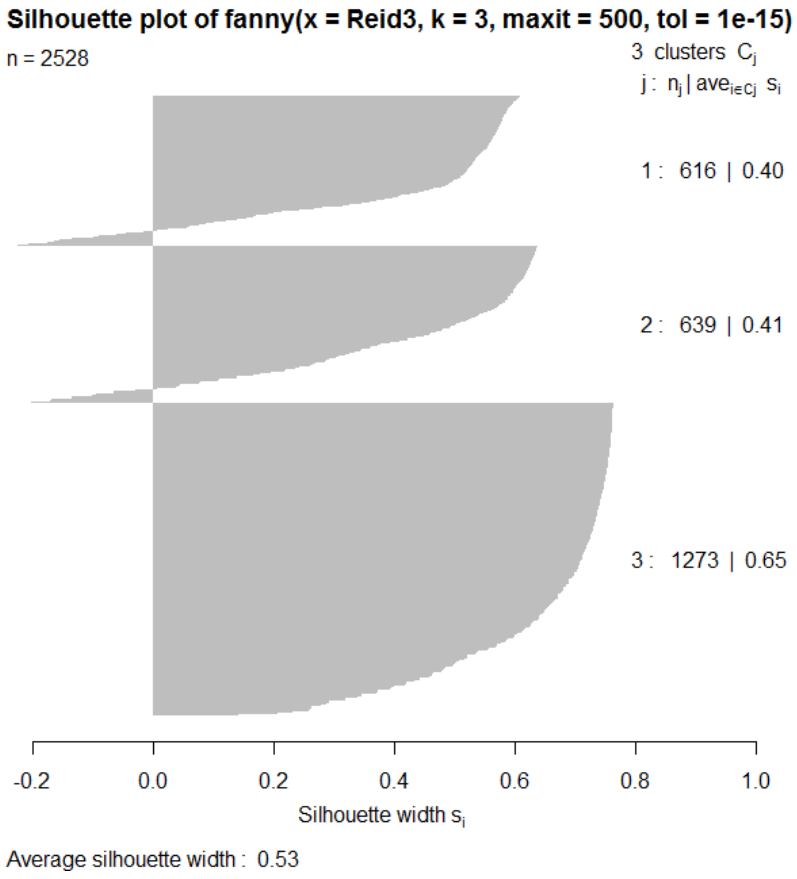
Resultados con 3 clusters:

```
> ff
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.        2
objective     19135.37
tolerance        1e-15
iterations          -1
converged            0
maxit              500
n                 2528
Membership coefficients (in %, rounded):
      [,1] [,2] [,3]
1252    74   16   10
1257    16   67   17
1270     6   76   17
1273    61   22   17
1277    22   60   18
1278    76   15    9
1281    74   16   10
1282     3   10   86
1283     2    8   90
```

Observamos el grado de pertenencia de cada individuo a los 3 clusters y veremos si con más grupos mejora la relación:

```
> res
   Lat
      0   1
  1 283 333
  2 326 313
  3 479 794

> prop.table(res,margin=1)
   Lat
             0         1
  1 0.4594156 0.5405844
  2 0.5101721 0.4898279
  3 0.3762765 0.6237235
```

Nuevamente las relaciones con la latitud no reflejan nada.



Silueta de los 3 clusters.

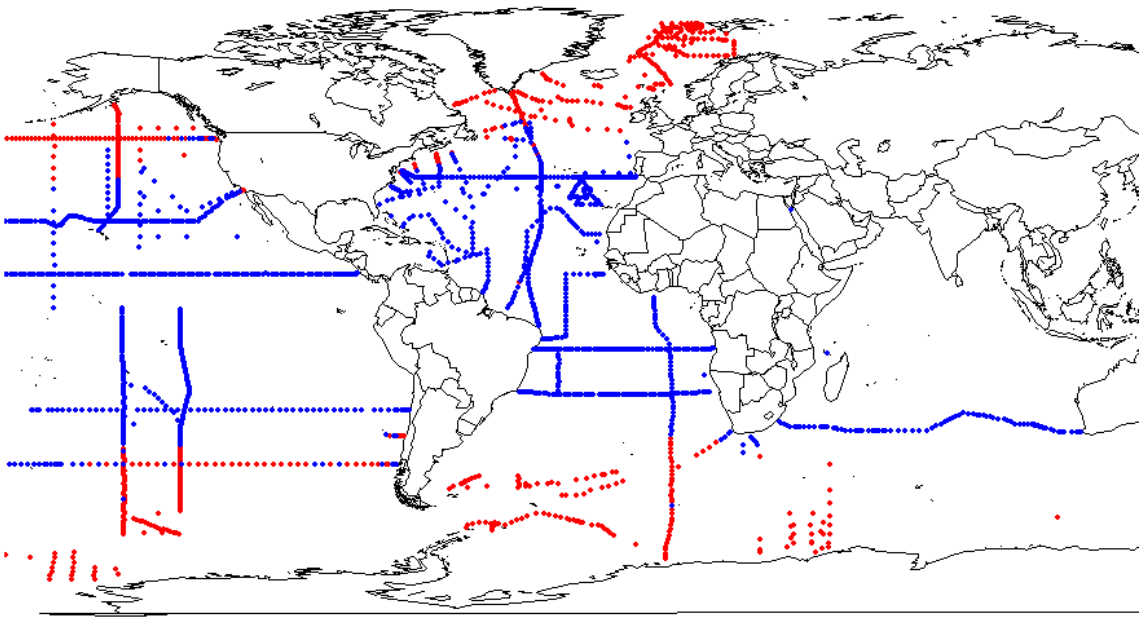
Comparamos con la Longitud:

```
> res
   Long
      0    1
  1  153  666
  2  170 1539

> prop.table(res,margin=1)
   Long
             0          1
  1 0.18681319 0.81318681
  2 0.09947338 0.90052662
```

Como en los otros tipos de análisis cluster la relación con el Oeste es muy marcada, y nuevamente se debe a que las estaciones donde se recogieron las variables están en una zona determinada al oeste del Meridiano de Greenwich.

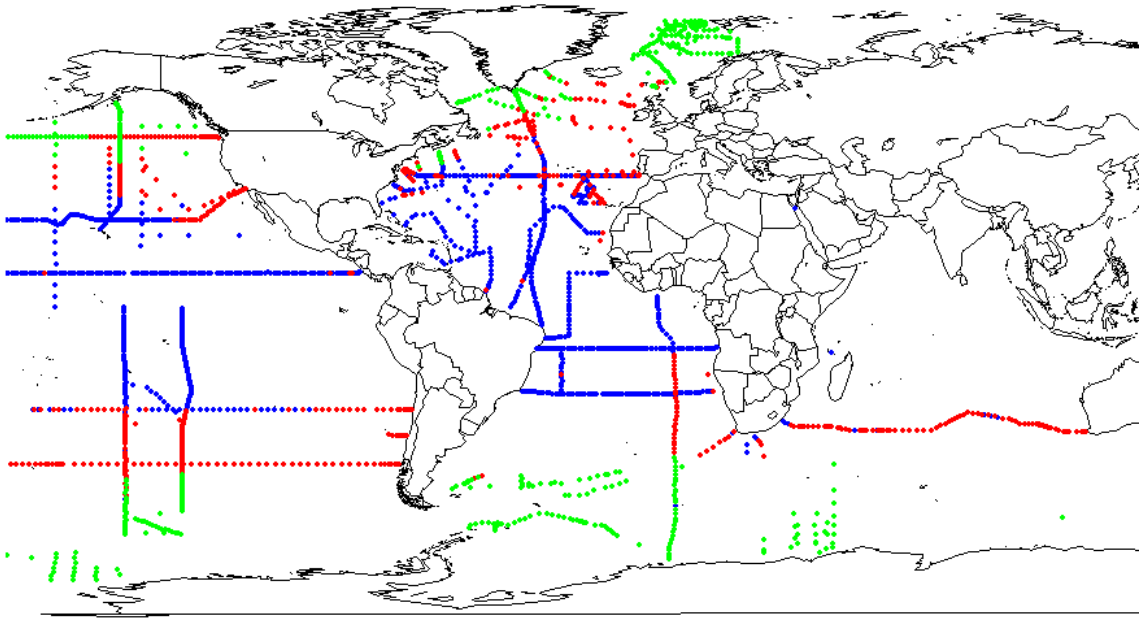
Vamos a proyectar los puntos en un mapa para visualizar con 2 y 3 clusters su distribución en un mapa mundo:

- **2 Clusters:**



Observamos que con el fuzzy seguimos la misma tendencia, un grupo alrededor de las zonas polares y el otro grupo en el resto. Sin embargo se ven unos puntos que se meten en la zona del otro grupo.

- **3 clusters:**



Aquí vemos la variedad de más puntos en los límites de las zonas que ocupan los diferentes grupos, hay zonas incluso en noroeste que se juntan puntos de los 3 grupos.

5.3.1. Código utilizado en R:

```
# Igual que el anterior:
        setwd("C:/Users/kine/Desktop/Dropbox/Master_iñaki/T_F_M/R")
…
        Long<-Reid2$Long
# Nueva librería:
        library(cluster, pos=4)
        ff<-fanny(Reid3,k=2,maxit = 500, tol = 1e-15)
        ff
        plot(ff)
# comparamos con la Latitud y longitud:
        res<-table(ff$cluster,Lat)
        res
        prop.table(res,margin=1)

        res<-table(ff$cluster,Long)
        res
        prop.table(res,margin=1)
# Proyectamos las coordenadas en un mapa mundo para ver la distribución de los
grupos, cambiamos color y numero de grupo.
        library(mapproj)
        map("world")

        out=mapproject(Long[ff$clustering==2], Lat[ff$clustering==2], projection="",
        parameters=NULL, orientation=NULL)
        x<-out$x
        y<-out$y

        points(x,y,col="blue",pch=16,cex=.6)
```

6. Conclusiones

El análisis cluster como forma de clasificación no supervisada de datos, es una manera de encontrar patrones, relaciones o descubrir conocimiento en bases de datos sin poner condiciones a dicha búsqueda.

Con el ejemplo y posterior prueba vemos reflejadas claramente las ventajas e inconvenientes de este tipo de análisis:

- Por una parte podemos descubrir grupos de individuos con una información homogénea, respecto a una serie de variables que sin este tipo de análisis no podríamos ver a simple vista.

- Pero después corre de mano del investigador el encontrar ese porque, así como de encontrarle o no utilidad a estos descubrimientos.

En los resultados contrastamos 3 análisis cluster, con los mismos datos y los mismos números de grupos y representamos esos puntos proyectándolos sobre un mapa mundo. Los 3 dan un resultados muy parecidos, los límites de los grupos están muy marcados salvo en el análisis fuzzy donde se mezclan un poco más en los extremos.

El siguiente paso sería estudiar el porqué de esta concentración de elementos presente en el agua en esas latitudes.

Como decíamos antes los tipos de análisis K-means y Fuzzy K-means, producen resultados más parecidos que los del cluster jerárquico. Al ser Fuzzy k-means una variante del K-means tienen unas pautas parecidas.

7. Bibliografía

- Páginas web:

  http://en.wikipedia.org/wiki/Cluster_analysis

  http://es.wikipedia.org/wiki/Taxonomía

  http://es.wikipedia.org/wiki/R_(lenguaje_de_programación)

  http://en.wikipedia.org/wiki/K-means_clustering

  http://knuth.uca.es/R/doku.php

  http://odv.awi.de/en/home/

  http://www.ugr.es/~gallardo

  http://ocean.ices.dk/Tools/UnitConversion.aspx

  http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html

  http://stat.ethz.ch/R-manual/R-patched/library/stats/html/kmeans.html

  http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/fanny.html

  http://127.0.0.1:24247/library/mapproj/html/mapproject.html

- Base de datos

  World dataset, ODV: J. L. Reid, and A. W. Mantyla.

- Documentos

  Estadística Básica Con R y R–Commander, Autores: Universidad de Cádiz.

  Prácticas de Estadística, EPM 2013-2014, Autores: Ana Colubi y Gil González

  Estadística con R, Universidad de Cantabria, Autor: Fco Javier Glez Ortiz

  Introducción al análisis cluster, Universidad de Granada, José Ángel Gallardo.

- Libros

  Pang--Ning Tan, Michael Michael Steinbach Steinbach & Vipin Kumar: "Introduction Introduction to Data Mining"
  Adisson-Wesley,2006. ISBN 0321321367

  Jiawei Han & Micheline Micheline Kamber: "Data Mining: Concepts Concepts and Techniques"
  Morgan kaufmann, 2006. ISBN 1558609016