UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Doctorado en Estadística Multivariante Aplicada

Tesis Doctoral

CLUSTER NO JERÁRQUICOS

versus CART y BIPLOT

Autor: GONZALO ISAAC CARRASCO OBERTO

Directoras: MARÍA PURIFICACIÓN VICENTE GALINDO

MARÍA CARMEN PATINO ALONSO

2020

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

DEPARTAMENTO DE ESTADÍSTICA

DRA. M.ª PURIFICACIÓN VICENTE GALINDO

Profesora Titular de Universidad del Departamento de Estadística de Universidad de Salamanca. Área de Estadística e Investigación Operativa

y

DRA. M.ª CARMEN PATINO ALONSO

Profesora Titular de Universidad del Departamento de Estadística de Universidad de Salamanca. Área de Estadística e Investigación Operativa

---

CERTIFICAN:

Que **D. Gonzalo Isaac Carrasco Oberto** ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada, que presenta con el título **Cluster no Jerárquicos *versus* Cart y Biplot**, autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a 2 de noviembre de 2020.

M.ª Purificación Vicente Galindo M.ª Carmen Patino Alonso

CLUSTER NO JERÁRQUICOS versus CART Y BIPLOT

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

DEPARTAMENTO DE ESTADÍSTICA

Trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada por la Universidad de Salamanca, presenta:

Gonzalo Isaac Carrasco Oberto

Salamanca

2020

La vida es una aventura, atrévete - Teresa de Calcuta

Con todo mi cariño a:
Julia, *mi madre*,
Gonzalo*†, mi padre*,
Julia y Armando*, mis hermanos*,
Armando, Neythan y Gael, mis sobrinos,
Ivone y Horacio*, mis cuñados*

AGRADECIMIENTOS

Agradezco infinitamente al Departamento de Estadística de la Universidad de Salamanca y muy especialmente a la **Dra. Purificación Galindo Villardón** haberme invitado a participar en el programa de Máster y Doctorado en Estadística Multivariante Aplicada, acogido durante estos años y brindarme su confianza y afecto. Sin su apoyo, trabajo y ejemplo quizás hoy fuera todo diferente.

A la Profesora Directora de esta tesis la **Dra. Mª Purificación Vicente Galindo** por su colaboración y disposición en ayudarme en los requerimientos que he tenido durante la realización de este trabajo.

A la Profesora Directora de esta tesis la **Dra. Mª Carmen Patino Alonso,** le agradezco su paciencia, sus críticas, sus exigencias firmes y sus certeras y oportunas sugerencias que han sido fundamentales para que este trabajo llegue a su fin.

A mis compañeros de la Universidad de Panamá participantes del Programa de Doctorado en Estadística Multivariante Aplicada, Mitzi, Carmen y Estelina por su solidaridad en estos años de estudio.

A la Magister Marisela del C Castillo G. y el Dr. Jose Luis Molina González por su orientación en el tema de Calidad del Agua en la Cuenca Hidrográfica del Canal de Panamá.

A mi familia, por su apoyo durante este tiempo del desarrollo del programa de doctorado. Gracias a Dios por haberme permitido llegar a este punto del camino de la vida.

TABLA DE CONTENIDO

INTRODUCCIÓN

Cada día estamos más inmersos en un mundo en el que los datos crecen y crecen. La minería de datos (MD) muy relacionada con el Descubrimiento de Conocimiento en Bases de datos (KDD -Knowledge Discovery in Databases) nos permite descubrir información de grandes volúmenes de datos y son fundamentales para analizarlos de manera eficaz, a la vez que revelan patrones que no eran conocidos (Holsheimer & Siebes, 1994).

El KDD es un proceso que consta de un conjunto de fases que incluye el preprocesamiento minería y post procesamiento de los datos. La minería de datos es una técnica de Inteligencia Artificial que permite extraer conocimiento útil y comprensible previamente desconocido a partir de grandes volúmenes de datos y consiste en la aplicación de un algoritmo para extraer patrones de datos. Sin embargo, con el fin de analizar los datos enfocados en el descubrimiento del conocimiento se ha ido adaptando y ha surgido lo que se denomina minería de datos espacial (MDE), la cual se considera como el proceso automático de explorar grandes cantidades de datos espaciales con el objetivo de descubrir conocimiento.

En la actividad investigadora resulta de gran interés identificar asociaciones, patrones y reglas. Dentro de las técnicas de MD se encuentra el agrupamiento (Clustering). El agrupamiento de datos es un problema fundamental en una variedad de áreas de la informática y campos relacionados, como el análisis de datos, la compresión de datos y el análisis de datos estadísticos (Aboubi, Drias, & Kamel, 2016). Puede considerarse el

problema más importante de aprendizaje no supervisado tratando de encontrar una estructura de datos no etiquetados (Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999).

Los algoritmos de agrupamiento más conocidos son los métodos jerárquicos y los métodos de partición, aunque existen otros métodos basados en densidades y los métodos basados en Gird. Existen diversas razones por las que las agrupaciones particionadas o de aprendizaje no supervisado son de interés: implementación rápida y convergen rápidamente, permiten categorizar elementos, entre otras. Sin embargo, estos algoritmos sufren inconvenientes en la especificación de los parámetros iniciales no adecuados, que pueden generar una mala convergencia. Se han desarrollado diferentes métodos de agrupamiento que atienden a diversos problemas como costo computacional, sensibilidad a la inicialización, clases desbalanceadas y convergencia a un óptimo local, entre otros. Sin embargo, para la selección de un método, es necesario considerar la naturaleza de los datos y las condiciones del problema con el fin de agrupar patrones similares, de tal forma que se tenga un buen compromiso entre costo computacional y efectividad en la separabilidad de las clases.

Algunos de los algoritmos basados en particiones son el algoritmo K-Medias, el algoritmo K-Medoids, el algoritmo de particionamiento alrededor de Medoids (PAM) y una versión de PAM diseñada para grupos de datos mayores denominado CLARA (Gupta & Panda, 2018). Hay numerosos investigadores que han propuesto algoritmos de K-Medias y K-Medoids (Borah & Ghose, 2009; Dunham, 2002; Han & Kamber, 2006; Khan & Ahmad, 2004; Park, Lee, & Jun, 2006; Rakhlin & Caponnetto, 2007; Xiong, Wu, & Chen, 2009).

La agrupación ha ganado un amplio uso y su importancia ha crecido proporcionalmente debido a la cantidad cada vez mayor de datos y al aumento exponencial en las velocidades de procesamiento de la computadora. La importancia de la agrupación se puede entender

por el hecho de que tiene una amplia variedad de aplicaciones, ya sea en educación o industrias o agricultura o economía. Las técnicas de agrupamiento se han vuelto muy útiles para grandes conjuntos de datos, incluso en redes sociales como Facebook y Twitter (Soni & Patel, 2017). El análisis de conglomerados juega un papel indispensable en la exploración de la estructura subyacente de un conjunto de datos dado, y se usa ampliamente en un variedad de temas de ingeniería y científicos, como, medicina, sociología, psicología y recuperación de imágenes Además en otras áreas, tales como, estudios de segmentación de clientes en el área financiera (Abonyi & Feil, 2007), biología (Der & Everitt, 2005; Quinn & Keough, 2002) , ecología (McGarigal, Cushman, & Stanford, 2000) , entre otros, puesto que la mayoría de las veces no utiliza ningún supuesto estadístico para llevar a cabo el proceso de agrupación (Leiva-Valdebenito & Torres-Avilés, 2010).

A partir de la década de 2000, los algoritmos de agrupación han tenido que abordar millones de patrones. Existen aplicaciones en las que se deben agrupar varios miles de millones de patrones con alta dimensionalidad. El número de grupos es cada vez mayor para muchas aplicaciones, como la recuperación de videos y la clasificación de imágenes. En muchos enfoques, la complejidad del algoritmo es proporcional al número de clústeres. Los costos de abordar grandes conjuntos de datos siempre son importantes, y la lucha entre el tiempo de cálculo y los números de clústeres se vuelve severa, especialmente a medida que aumentan los números de clústeres.

El Análisis de Cluster trabaja con variables tanto cualitativas como cuantitativas y el objetivo es obtener agrupaciones de forma que las observaciones dentro de cada grupo han de ser similares entre sí y los Grupos han de estar bien diferenciados. Los valores atípicos son los datos que se desvían significativamente de la mayoría de los demás datos

del conjunto de datos. Un buen algoritmo de agrupamiento separa los datos en grupos sin ser intervenido por datos atípicos. Además, un buen algoritmo de agrupamiento debe proporcionar una menor complejidad de tiempo y una mayor precisión de agrupamiento, especialmente al procesar una gran cantidad de datos (Yu, Chu, Wang, Chan, & Chang, 2018).

CAPÍTULO 1: K-MEDIAS vs KMEDOIDS: Una Revisión Crítica

1.1.-INTRODUCCIÓN

El término "**Ciencia de Datos**" se ha popularizado, para describir todo esfuerzo para extraer conocimiento a partir de datos. Visto de esta manera, es un término equivalente a "**Minería de Datos**", la cual es el proceso de extracción de conocimiento (relaciones, patrones, anomalías) en un conjunto de datos, usando técnicas de modelización provenientes de otras ramas como la estadística y aprendizaje automático (Witten, Frank & Hall, 2011), de tal manera que convierten los datos en conocimiento e información disponible (Tsiptsis & Chorianopoulos, 2009) y permite a las personas tomar decisiones basadas en información (North, 2012) para aumentar los ingresos, reducir costos, mejorar la relación con los clientes, reducir los riesgos, entre otras aplicaciones (Bucheli & Thompson, 2014).

Mientras que en Estadística las generalizaciones se realizan sobre una muestra representativa de la población, en la minería de datos la generalización se basa en los modelos aplicados sobre los datos disponibles. Si los datos no son representativos de la población puede obtenerse conclusiones erróneas. El investigador debe observar cuidadosamente este aspecto para la generalización.

Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial cuyo objetivo es crear sistemas que aprenden automáticamente a identificar patrones complejos en los datos. **La máquina que aprende es un algoritmo** que revisa los datos y es capaz de predecir comportamientos futuros. Estos sistemas se mejoran sin intervención humana y pueden **detectar patrones de comportamiento** a partir de las variables y descubrir cuáles son las más relevantes. Actualmente se disponible de una amplia gama de algoritmos de aprendizaje para estas tareas, agrupados por el tipo de problema a resolver, así los de clasificación enfocados a "predecir", y los de segmentación o asociación enfocados a "descubrir".

Existen diferentes tipos de Algoritmos de Aprendizaje: los *Supervisados*, orientados a "predecir", los *No supervisados* orientados a "descubrir" y *Por refuerzo* orientados a "aprender por sí mismos que hacer" (Russell & Norvig, 2010). Además, están desarrollándose algoritmos Semi-Supervisados que se sitúan entre los dos primeros (Bucheli & Thompson, 2014).

En el Aprendizaje supervisado, el objetivo es predecir un evento (valor específico de una variable categórica) o estimar valores de una variable continua. Una variable tendrá un rol como objetivo y las demás serán las predictoras (de entrada). Los modelos construidos están "supervisados" por las relaciones evaluadas entre la variable objetivo y las predictoras.

Corresponden a los modelos que en estadística son llamados asimétricos, así pues, las variables predictoras son las variables independientes y la variable objetivo es la variable dependiente.

Entre las técnicas con supervisión están a los árboles de decisión, reglas de decisión, regresión lineal, regresión logística, redes neuronales, máquinas de vectores de soporte (SVM) y las redes bayesianas. Algunas de ellas pueden trabajar con ambos tipos de variables objetivo (categóricas o continuas).

El objetivo de los métodos de Aprendizaje No Supervisado es descubrir patrones no evidentes en el conjunto de datos. Todas las variables juegan el mismo rol, así los modelos construidos no consideran ninguna variable objetivo.

Los modelos construidos agrupan o asocian las unidades taxonómicas y permiten obtener perfiles utilizarlos para tomar decisiones. Corresponden a los modelos que en estadística son llamados simétricos.

En este capítulo vamos a centrarnos en los métodos de clasificación no supervisada de clustering.

Los algoritmos de clustering se pueden clasificar en dos tipos:

- Agrupamiento *jerárquico*: construye un dendograma que representa las relaciones de similitud entre los distintos elementos y pueden ser aglomerativo o divisivo.
    - Agrupamiento *jerárquico aglomerativ*o: se parte de tantos grupos como individuos hay en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo. Dentro de estos métodos se tienen: i) método de encadenamiento simple, (ii) métodos de encadenamiento completo, (iii) método de encadenamiento medio, (iv) método de Ward, y (v) método del centroide entre otros (Hair, Anderson, Tathaman, & Black, 1999). Unos métodos difieren de otros en la forma de calcular las diferentes distancias entre conglomerados (Manhattan, coeficiente de correlación de Pearson, etc.).
    - Agrupamiento *jerárquico divisivo*: se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez más pequeños.

- Agrupamiento *no jerárquico*: también conocidos como *métodos partitivos* o de optimización tienen por objetivo realizar una sola partición de los individuos en k grupos, es decir, el número de grupos se determina de antemano y las observaciones se asignan a los grupos en función de su cercanía. No requiere de procesos de construcción de árboles. Esto conlleva que debe especificar a priori los grupos que deben ser formados. Ésta es probablemente, la principal diferencia respecto de los métodos jerárquicos. La asignación de individuos a los grupos se realiza mediante algún proceso que optimice el criterio de selección. Trabajan con

la matriz de datos originales y no requieren su conversión en una matriz de proximidades. Esencialmente siguen los siguientes pasos:

1. Seleccionar K medias iniciales, siendo K el número de clusters deseados.
2. Asignar cada observación al cluster que le sea más cercano.
3. Reasignar o relocalizar cada observación a uno de los K cluster de acuerdo con alguna regla de parada.
4. Parar si no hay reasignación de los puntos o si la reasignación satisface la regla de parada. En otro caso se vuelve al paso dos.

1.2.- K MEDIAS

El método *K-means clustering* agrupa las observaciones en *K clusters* distintos, donde el número *K* lo determina el analista antes de ejecutar del algoritmo. *K-means clustering* encuentra los *K* mejores *clusters*, entendiendo como *mejor cluster* aquel cuya varianza interna (*intra-cluster variation*) sea lo más pequeña posible. Se trata por lo tanto de un problema de optimización, en el que se reparten las observaciones en *K clusters* de forma que la suma de las varianzas internas de todos ellos sea lo menor posible.

El algoritmo k-medias es unos de los algoritmos más simples y conocidos no solo dentro de los de tipo particional sino de los algoritmos en general, ya que sigue una forma fácil y sencilla para dividir un conjunto de datos en k grupos o clusters conocidos a priori (Forgy, 1965; J. A. Hartigan & Wong, 1979; Lloyd, 1982; MacQueen, 1967).

Pertenece al grupo de algoritmos de partición-optimización también conocido como algoritmo de Lloid en el que se reparten las observaciones en K clusters de forma que la

suma de las varianzas internas de todos ellos sea lo menor posible. El objetivo del método consiste en crear grupos homogéneos en su interior y heterogéneos entre sí.

La idea principal del algoritmo k-medias consiste en los siguientes pasos:

- ✓ *Inicialización:* Consiste en definir los objetos que se van a particionar, el número de grupos y el centroide para cada grupo. Para ello:
  - Se colocan k puntos en el espacio representado por los objetos que se están agrupando. Estos puntos representan los centroides del grupo inicial. Aunque existen varios métodos para definir los centroides iniciales, el más utilizado es la selección aleatoria.
  - Clasificación: Se calcula la distancia hacia todos los centroides para cada objeto y se asigna cada objeto al grupo que tiene el centroide más cercano.
  - Cálculo de centroides: cuando se hayan asignado todos los objetos, se vuelven a calcular las posiciones de los k centroides.
- ✓ *Proceso iterativo:* Mientras los centroides no cambien se procede a calcular la distancia del centroide y volver a distribuir todos los objetos según el centroide más cercano. El proceso se repite hasta que ya no hay cambio en los grupos, es decir, los k centroides no cambian después de una iteración, lo cual es equivalente a decir que el valor de la función utilizada como criterio de optimización no varía.
- ✓ *Criterio de convergencia:* Consiste en establecer el criterio de paro del algoritmo. Es posible converger cuando alcanza un número de iteraciones dado, cuando no exista un intercambio de objetos entre los grupos o converger cuando la diferencia entre los centroides de dos iteraciones

consecutivas es más pequeña que un umbral dado. Si la condición de convergencia no se satisface, se repiten los pasos anteriores del algoritmo.

En este algoritmo se usa como una métrica la distancia euclídea y la varianza como medida de dispersión entre los grupos. El algoritmo tiene como objetivo maximizar una función objetivo, denominada suma de errores cuadráticos (SSE) o también conocida como sumas residuales de cuadrados (RRS). La función objetivo se define como:

$$SSE = J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2$$

Dónde:

$\left\| x_i^j - c_j \right\|^2$ es una medida de distancia elegida entre un punto de datos x^j y el centro de clúster c_j, es un indicador de la distancia de los n puntos de datos de sus respectivos centros de clusters.

Al utilizar la distancia euclídea, SSE se minimiza usando la media aritmética (por cada atributo o variable). Puede ser entendido como un algoritmo que intentará minimizar el factor de inercia del cluster.

Entre las ventajas del método destacan:

- Es eficiente
- La implementación es sencilla; basta con asignar aleatoriamente un número entre 1 y K a cada observación, asignación inicial aleatoria de las observaciones a los *clusters*, e iterar los siguientes pasos hasta que la asignación de las observaciones a los *clusters* no cambie o se alcance un número máximo de

iteraciones establecido por el usuario. Para cada uno de los *clusters* se calcula su centroide y se asigna cada observación al *cluster* cuyo centroide está más próximo.

Como desventajas se pueden señalar:

- Se necesita conocer k de antemano, es decir, requiere que se indique previamente el número de clusters a crear.
- Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides. El algoritmo es significativamente más sensible a los centros de agrupamiento seleccionados al azar inicialmente, por tanto, el resultado obtenido es dependiente de la selección inicial de los centroides de los clústeres y puede converger a óptimos locales. Se recomienda repetir el proceso de clustering entre 25-50 veces y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna.
- Presenta problemas de robustez frente a outliers.
- -No trata datos nominales.

En la figura siguiente se presenta un esquema del algoritmo:

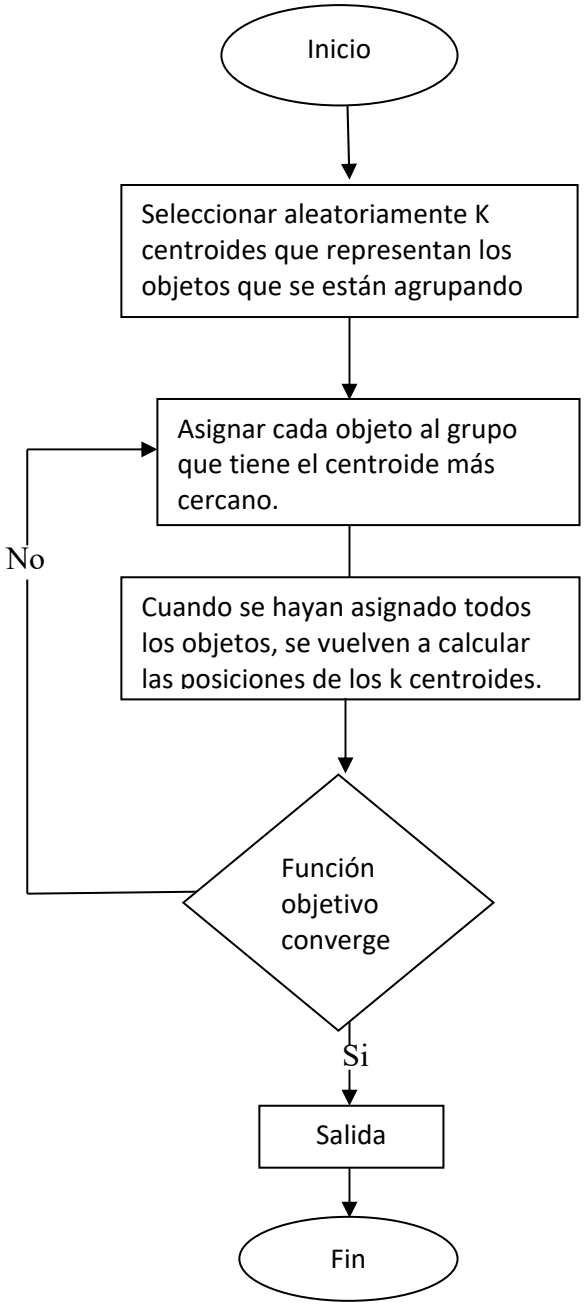

Figura 1. Algoritmo K-Medias

No existe garantía de que la solución encontrada por el algoritmo K-medias sea la más óptima. Solo es una aproximación a la solución del problema matemático (James, Witten, Hastie, & Tibshirani, 2014).

Dado que el K-medias requiere fijar puntos iniciales que son seleccionados de manera aleatoria, repetir la clasificación con puntos iniciales diferentes y analizar la estabilidad de las soluciones, es una buena práctica.

Hay que tener en cuenta que el algoritmo está basado en distancias y puede verse afectado por las unidades de medida de las variables, por lo que puede ser recomendable estandarizar las variables antes de iniciar la búsqueda de los clústeres.

A pesar de ser el algoritmo más usado, hemos puesto de manifiesto que tiene importantes limitaciones, por lo cual vamos a analizar y realizar el estudio crítico de diferentes alternativas que se han publicado posteriormente.

1.3.- K-MEDOIDS

El algoritmo k-Means presenta problemas de robustez frente a *outliers*. La única solución es excluirlos o recurrir a otros métodos de *clustering* más robustos como *K-Medoids (PAM)* que describimos a continuación.

K-medoids es un método de *clustering* similar a *K-means* en cuanto a que ambos agrupan las observaciones en *K clusters*, donde *K* es un valor preestablecido por el analista.

La diferencia es que, en *K-medoids*, cada *cluster* está representado por una observación presente en el *cluster* (*medoid*), mientras que en *K-means* cada *cluster* está representado

por su centroide, que se corresponde con el promedio de todas las observaciones del *cluster* pero con ninguna en particular.

Un medoid puede ser definido como un objeto de un grupo, cuyo promedio de disimilitud a todos los objetos del grupo es mínima, es decir es el punto más céntrico del clúster Park & Jun (2009).

El algoritmo para K-Medoids fue propuesto inicialmente por Vinod (1969), retomado posteriormente por Rao (1971), Massart et al. (1983), Klastoria (1985) y Kaufman y Rousseeuw (1987, 1989) entre otros, siendo su cometido el de minimizar la suma de las distancias entre cada elemento y su correspondiente objeto representativo (medoid).

El algoritmo k-Medoids es un algoritmo de agrupamiento, basado en el uso de objetos del conjunto de datos para ser los representantes de los conglomerados denominados medoids.

La resolución del algoritmo viene determinada por:

1.-La elección de los k objetos representativos, que deberán presentar los diferentes aspectos de la estructura de los datos. En este modelo el objeto representativo de un clusters es el objeto para el que la disimilitud promedio (o equivalentemente la disimilitud total) a todos los objetos del clusters sea mínima. El objeto es llamado medoid del cluster. Para ello, se definirá y_{ij} como una variable de cero-uno, siendo $y_{ij}=1$ solo cuando el objeto i se haya seleccionado como objeto representativo.

2.- La ubicación de cada objeto j junto a uno de los k objetos representativos seleccionados. Se definirá m_{ij} como una variable 0-1, siendo $m_{ij}=1$ sólo cuando el objeto

j haya sido ubicado en aquel cluster cuyo objeto representativo sea i, teniendo un valor igual a cero para cualquier otra circunstancia. A partir de aquí la resolución tendrá que:

$$\text{Minimizar } \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) m_{ij} \tag{1}$$

Sujeto a:

$$\sum_{i=1}^{n} m_{ij} \qquad _{j=1,2,\ldots,n} \tag{2}$$

$$m_{ij} \leq y_i \qquad i,j = 1,2,\ldots,n \tag{3}$$

$$\sum_{i=1}^{n} y_i = k, \;\; k = n\acute{u}meros\ de\ clusters \tag{4}$$

$$y_i, m_{ij} \in \{0,1\} \;\; i,j = 1,2,..,n \tag{5}$$

La restricción (2) expresa que cada objeto j debe ser asignado a un solo medoid. La restricción (2) junto con la restricción (5) expresan que para un j dado uno de los z_{ij} es igual a uno y todos los demás son cero. La restricción (3) garantiza que un objeto j sólo se puede asignar a un objeto i si este último objeto ha sido seleccionado como un medoid.

De hecho, si este no es el caso, entonces y_i es cero y la restricción (3), junto con la restricción (5) implica que todo el z_{ij} son cero. Si i es un medoid, entonces todo los z_{ij} (para este i) puede ser cero o uno. La ecuación (4) expresa que exactamente k objetos son elegidos como medoids. A medida que los clusters se forman mediante la asignación de cada objeto al medoid más similar, habrá exactamente k clusters no vacíos. (En caso de empate, el objeto se asigna al medoid, que se introdujo primero).

La distancia entre un objeto ***j*** y su objeto representativo **i** será:

$$\sum_{i=1}^{n} d(i,j) m_{ij}$$

Como se debe asignar todos los objetos, la disimilitud total viene dada por:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} d(i,j) m_{ij} \qquad (6)$$

que es la función a minimizar la distancia total representada al inicio.

En el algoritmo se pueden diferenciar dos etapas:

A.-En la primera fase tendrá lugar la selección, paso a paso, de cada objeto representativo de los k ***clusters*** especificados, siendo el primer objeto elegido aquél para el cua1 la suma de las distancias al resto de objetos sea mínima. Una vez escogido un objeto representativo *i,* aún no seleccionado, la identificación de cada objeto implicara cuatro movimientos, que se sucederán en el orden siguiente:

a) Para cada objeto j se calculará la diferencia entre su distancia $\boldsymbol{D}_j$ con el objeto más similar a (6), previamente seleccionado, y su distancia con el objeto i recientemente incluido como objeto representativo.

b) Si esta diferencia es positiva, el objeto j estará contribuyendo a la selección del objeto ***i,*** calculándose entonces:

$$C_{ij} = \max\left(D_j - d(i,j), 0\right)$$

c) La mejora total debida a la elección del objeto **i** se obtendrá de:

$$\sum_{j} C_{ji}$$

d) Por último, se seleccionará el objeto **i** que maximice el sumatorio anterior.

B.- En la segunda fase se intentará mejorar el conjunto de objetos representativos, para el1o se tendrán en cuenta todos los pares de objetos (i,h) para los que el objeto **i** ha sido seleccionado pero el objeto ***h*** no. Se observará la modificación que tendría lugar si se eligiera el objeto ***h,*** pero no el objeto **i.** Para verificar el efecto de este cambio se llevarán a cabo 1os dos pasos siguientes:

1. Se tomará un objeto no seleccionado j y se calculará la contribución de Cj, al cambio:

a) Si el objeto j está más alejado de i y de h que de cualquiera de 10s otros objetos representativos, C_{jih}, tendrá un valor igual a cero.

b) Si el objeto j no está más distanciado de i que de cualquier otro objeto representativo se tendrá en cuenta si:

- El objeto j está más próximo de h que del segundo objeto representativo mis cercano

$$d(j,h) < E_j$$

siendo entonces la contribución de j al cambio entre 1os objetos i y h la siguiente:

$$C_{jih} = d(j,h) - d(j,i)$$

En esta situación la contribución de C_{jih}, podría ser tanto positiva como negativa, dependiendo del lugar ocupado por 1os objetos j, h e i. Únicamente si el objeto j está más próximo de i que de h la contribución será positiva, en cuyo caso el objeto j no estaría favoreciendo el cambio.

c) Por último, si el objeto j se encuentra mis alejado del objeto i que al menos uno de 1os objetos representativos, pero más próximo de h que de cualquier otro objeto representativo, la contribución de j al cambio sería:

$$C_{jih} = d(j,h) - D_j$$

2.-Se calculará el cambio total sumando cada una de las contribuciones de C_{jih} y se pasará a decidir si se lleva a cabo dicha modificación. Para ello, se seleccionara el par (i,h) con el fin de de minimizar la suma de las contribuciones de C_{jih}. Si el valor más reducido es negativo, el cambio tendrá lugar, y se volverá al primer paso. En cambio, si el valor más pequeño es positivo o cero el algoritmo se detendrá.

Como la mayoría de los métodos de particionamiento permiten construir k particiones de las observaciones, donde cada partición representará a un cluster, segmento o grupo. La posible manera de seleccionar un valor de k con el fin de encontrar la agrupación más significativa es por medio del coeficiente de silueta de Rousseeuw (1985).

Kaufman y Rousseeuw (1989) destacan las siguientes ventajas del algoritmo:

a) Ser uno de 1os métodos más robustos (las técnicas basadas en la minimización de promedios de distancias, o de residuales) en valores absolutos son más robustas que las basadas en sumas de cuadrados.

b) Ofrecer configuraciones bastante precisas cuando 1os clusters no son excesivamente alargados.

c) Sus agrupaciones no dependen del orden en que han sido introducidos 1os objetos, tal como se ha indicado puede suceder con otras técnicas no-jerárquicas.

d) Es efectivo debido a que es invariable frente a los valores atípicos.

Ventajas del algoritmo K-medoids versus K-medias:

1.-El algoritmo k-medoids es más robusto al ruido y a valores grandes de los datos en comparación con el K-medias, ya que minimiza la suma de diferencias por parejas en lugar de la suma de los cuadrados de las distancias euclidianas.

2.-El algoritmo k-medoids tiene como objetivo minimizar el criterio de error absoluto en lugar del SSE, como en el caso del algoritmo K-medias, sin embargo, procede interactivamente hasta que cada objeto representativo sea en realidad el medoid de la agrupación.

3.-El algoritmo k-medoids utiliza los medoids en lugar de los centroides como en el caso del algoritmo K-medias. K-medoids se basa en el objeto más céntrico de un clúster haciéndolo menos sensible a los valores atípicos en comparación con la agrupación K-medias.

4.-El centro del cluster es parte del conjunto de datos, a diferencia de k-means donde el centro del cluster es basado en el centro de gravedad.

1.4.- K-MEDIAS DIFUSA

El **agrupamiento difuso** (en inglés, ***fuzzy clustering***) es una variación de los algoritmos de agrupación donde cada elemento tiene un grado de pertenencia difuso a los grupos; es decir, a diferencia del k-means 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, tras la introducción de la lógica difusa, surgió una solución para este problema, caracterizando el grado de similitud de cada elemento a cada uno de los clusters. Es decir, no se considera la pertenencia de forma dicotómica. Se trabaja con una 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.

El algoritmo de k-medias difusa es presentado en su forma inicial por Dunn (1974) como una alternativa a los clusters clásicos de k-medias y es completado por Bezdek (1974). La función objetivo se define como:

$$D(X;W;C) = \sum_{J=1}^{k} \sum_{i=1}^{n} \left(w_{ji}\right)^{m} \left\|x_i - c_j\right\|^2 \quad (1)$$

Dónde:

n es el número total de patrones en un conjunto de datos dado y k es el número de clúster. $X = \{x_1, .., x_n\} \subset \mathcal{H} \quad y\ C = \{c_{1,\dots,} c_k\} \subset \mathcal{H}$ son los datos de la característica y los centroides del clúster.

$W = \{w_{ji}\}$ y k × n es una matriz de partición difusa.

La función objetivo no se puede minimizar directamente. Por lo tanto, para minimizar la ecuación (1) se requiere satisfacer las siguientes dos expresiones:

$$W_t = D_w(C_{t-1}) \quad (2)$$

$$C_t = D_c(W_{t-1}) \quad (3)$$

Por lo tanto, se utiliza un algoritmo iterativo que optimiza alternativamente los grados de pertenencia y los parámetros del cluster. Al iterar dos pasos, se alcanza el óptimo conjunto

(aunque no se puede garantizar que se alcance el óptimo global; el algoritmo puede atascarse en un mínimo local de la función objetivo D). Las fórmulas de actualización Dw y Do se deriva estableciendo la derivada de la función objetiva D y los parámetros a optimizar iguales a cero (Höppner, Klawonn, Rudolf, & Runkler, 1999) y se obtiene:

$$w_{ji} = \left[\frac{1}{\sum_{r=1}^{k} \left[\frac{\|x_i - c_j\|}{x_i - o_r} \right]^{2/m-1}} \right] \tag{4}$$

$$c_j = \frac{\sum_{i=1}^{n} (w_{ji})^m x_i}{\sum_{i=1}^{n} (w_{ji})^m} \tag{5}$$

El algoritmo de k-medias difusa termina cuando el cambio relativo en los centros de los conglomerados se vuelve muy pequeño o la función objetivo D ya no puede minimizarse (Bezdek, 1974).

Etapas del algoritmo:

1.-Seleccionar el exponente de ponderación m (m> 1) e inicialice la matriz de partición difusa W= (w_{ji}) al azar.

2.-Mientras no se cumplen las condiciones de terminación se debe:

- Calcular los centros del clúster
- Actualizar la matriz de participación difusa W= (w_{ji})

3.-Fin

Ozdemir & Kaya (2018) exponen que el algoritmo difuso de k-medias detecta grupos en forma de puntos. No es eficaz para encontrar otras formas de racimo.

Ventajas del algoritmo

- Convergencia constante
- Sin supervisión

Desventajas del algoritmo

- La sensibilidad a los puntos de ruido
- La adherencia a los valores iniciales
- El tiempo de cálculo es de larga duración

Sintetizando todo lo anterior, la diferencia entre K-medias y el K-medias difuso es que en el K-medias cada elemento pertenece a un solo cluster, mientras que en el método K-medias difuso, cada elemento puede pertenecer potencialmente a varios clusters; es decir, cada observación tiene asignado un grado de pertenencia a cada uno de los cluster. El algoritmo proporciona información, para cada observación, sobre la probabilidad de pertenecer a cada cluster.

1.5.- K-MEDIAS RECORTADAS

El estudio y desarrollo de los métodos de clúster es un objetivo importante en el análisis de datos (Hartigan, 1975; Kaufman & Rousseeuw, 1990). Entre las técnicas estándar disponibles en estimación robusta, las que se basan en la eliminación de una porción de los datos (procedimientos de recorte) presentan un buen rendimiento, siendo a menudo un punto de referencia obligatorio para comparar nuevos estimadores. Sin embargo, la arbitrariedad en la selección de las zonas para la eliminación de datos es grave inconveniente de tales procedimientos.

Gordaliza (1991a) introdujo una clase de mejores aproximaciones basado en la idea de recorte imparcial, Los recortes dependen sólo de la estructura conjunta de los datos y no de zonas seleccionadas arbitrariamente para eliminar los datos. Por lo tanto, son especialmente adecuados en el caso multivariante (Gordaliza, 1991b). El procedimiento de k-medias recortado se basa en el algoritmo de k-medias y en la metodología de recorte imparcial descrito por Cuesta-Albertos et al. (1997).

Los pasos para llevar a cabo el algoritmo k-medias recortadas fijando un nivel de recorte α son:

1.-Determinar aleatoriamente un conjunto inicial de k centroides.

2.-Asignar cada observación al centroide más cercano.

$$d_i = \min_{j=1,2,\dots,k} \|x_i - c_j\|, \quad \text{i=1,2,,,n}$$

$\{c_1, c_2, \dots, c_k\}$: k centroides fijados en el paso anterior. Por cada observación se calcula la distancia euclídea al k centroide más cercano.

Para determinar las observaciones a recortar se determina la condición conocida como:

"radio óptimo": $d_i \geq d_{([n(1-\alpha)])}$

Las (n* α) con mayor distancia d_i se recortan y se asignan a:

$$H = \{i: d_i \geq d_{([n(1-\alpha])}\}, \text{ i=1, 2,…,n}$$

Basándose en la distancia cada observación se asigna al centroide más cercano y dejando fuera los pertenecientes a H.

3.-Calcula los nuevos centroides como la media de las observaciones que han sido asignadas al grupo y no han sido recortadas:

$$c_j = \frac{1}{n_j} \sum_{i \in H_j} x_{i,} \quad j=1,\dots,k$$

Es necesario inicializar varias veces el algoritmo desde la etapa inicial para encontrar soluciones y utilizar aquella que minimice la función objetivo.

El procedimiento de k-medias recortadas es muy similar al k-Medias pero con la diferencia de que las observaciones recortadas no se consideran en el cálculo de la función objetivo.

1.6.- K-ARMÓNICA MEDIAS

La agrupación K-Armónica Medias (KHM) es un método basado en la agrupación de centro propuesto por Zhang et al. (2000) y modificado por Hammerly & Elkan (2002). Este algoritmo utiliza la media armónica de la distancia de cada elemento a los centroides.

La media armónica se define como $MA(\{a_{1,\dots,}a_k\}) = \frac{K}{\sum_{k=1}^{k}\frac{1}{a_k}}$.

Cuando asignamos los objetos a los clusters se utiliza el mínimo de las distancias en el algoritmo k-medias y el promedio armónico en el algoritmo k-armónica medias:

$$MIN\{\|x-c\|^2/c \in C\} \rightarrow HA\{\|x-c\|^2 \mid c \in C\} = \frac{|C|}{\sum_{c\in C}\frac{1}{\|x-c\|^2}}$$

La función de rendimiento para el algoritmo K-medias es, entonces:

$$Perf_{KM}(X,C) = \sum_{i=l}^{N} MIN\{\|x_i - c_l\|^2 \; l = 1,\dots,k\}$$

Por tanto, la función objetivo del algoritmo es:

$$KAM(X,C) = \sum_{i=l}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}}$$

Función de peso de KAM: $W_{KAM}(x_i) = \frac{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}{\left(\sum_{j=1}^{k} \|x_i - c_j\|^{-p}\right)^2}$

Al reemplazar la función MIN () en $Perf_{KM}$ (X, C) con HA () obtenemos:

$$Perf_{KAMp}(X,C) = \sum_{i=l}^{N} HA\{\|x_i - c_l\|^2 | l = l, \dots, K\} = \sum_{i=l}^{N} \frac{K}{\frac{1}{\|x_i - c_l\|^2}}$$

de rendimiento Y la función para el algoritmo de K-armónica medias es entonces:

$$KAM(X,C) = \sum_{i=l}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}}$$

Se encontró que KAM funciona mejor con valores de p> 2.

Pasos del algoritmo:

1.- Elegir un punto arbitrario de X como primer centroide y el resto de centroides se escogen siguiendo una estrategia en la que el elemento elegido en la i-ésima iteración es aquel cuyo elemento más cercano de los i − 1 ya elegidos se encuentra más alejado.

2.-Se realiza una selección aleatoria del primer centroide c_1 del conjunto X y el segundo centroide c_2 se obtiene como el elemento que presenta la máxima distancia entre c_1 y X − c_1.

3.-A partir del cálculo de los dos primeros centroides para mejorar la convergencia del algoritmo se aplica una etapa de inicialización basada en el criterio Max-Min.

Para abordar el problema de sensibilidad de inicialización Qin & Suganthan, (2005) propusieron un algoritmo de cuantificación vectorial de aprendizaje (LVQ) llamado algoritmo (H2N-LVQ). Para resolver problema de óptimos locales, existen métodos heurísticos como la búsqueda tabú (Glover & Laguna, 1997), recocido simulado de Kirkpatrick et al. (1983), los métodos evolutivos, métodos de optimización nube (Swarm) son ejemplos de estos métodos heurísticos.

La agrupación en clústeres es una forma eficaz de obtener información a partir de datos sin procesar y k-medias es un método básico para ello. Aunque es fácil de implementar y comprender, k-medias tiene serios inconvenientes. Algunos se enfocan en crear buenos métodos de inicialización, mientras que otros buscan encontrar valores óptimos para k. Güngör & Ünler (2007) proponen un algoritmo de agrupamiento de datos de k medias armónicas con heurística de recocido simulado "simulated annealing heuristic" que llamamos SAKHMC. En el estudio desarrollado por Güngör & Ünler (2007) el algoritmo (SAKHMC) se ha implementado y probado en varios conjuntos de datos y a resultado que el algoritmo (SAKHMC) supera a k-medias y k-medias armónica desde el punto de valor de rendimiento en la mayoría de los casos.

1.7.- K-MEDIAS SPARSE

Witten & Tibshirani (2010) propusieron una alternativa al método clásico k-medias llamado K-Medias Sparce (sparse k-means - SK-means) el cual de forma simultánea encuentra los clusters y las variables importantes en la agrupación.

Sparse explota el hecho de que las medidas de disimilitud comúnmente utilizadas (por ejemplo, distancia euclidia al cuadrado) pueden descomponerse en términos p, cada uno de ellos dependiendo únicamente de una única variable. Es decir, dada una partición del clúster C = (C_1, C_2,…..,C_K), la medida de disimilaridad asociada entre cluster A (C) se denota como:

$$A(C) = \sum_{j=1}^{p} A_j(C)$$

A_j(C): depende de las j^{th} variables.

Dado un vector de pesos no negativo:

$$A(v, C) = \sum_{j=1}^{p} v_j A_j\,(C) = v' A(C) \qquad (1)$$

Dónde: $A(C) = \begin{pmatrix} A_1(C) \\ A_2(C) \\ . \\ . \\ . \\ A_p(C) \end{pmatrix}$

K-Medias Sparse busca para el par (v^*, C^*) que maximice

$$A(C) = \sum_{j=1}^{p} A_j(C)$$

sujeto a: $\sum_{j=1}^{p} v_j^2 \leq 1 \;\; and \sum_{j=1}^{p} v_j \;\leq l$

para $1 < l < \sqrt{p}$ SK-means realiza una versión regularizada k-medias (tipo LASSO-*"***L**east **A**bsolute **S**hrinkage and **S**election **O**perator*"*), propuesto por Tibshirani (1996). SK-means no produce una partición razonable cuando los datos contienen incluso una proporción muy pequeña de valores atípicos (por ejemplo, el 1% de las observaciones que tienen un valor atípico en sólo una de los cientos de características). Para remediar esta falta de robustez. Kondo et al. (2012) proponen una alternativa robusta llamada Robust and Sparse K-means (RSK-means).

1.8.- K-MEDIAS SPARSE ROBUSTO (RSK-MEANS)

El algoritmo de agrupación de k-medas encuentra K agrupaciones C_1,. . .,C_k que minimizan la suma de cuadrados dentro de los grupos:

$$\min_{C_{1,\dots,}C_k} \left\{ \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i'} \right\}, \tag{2}$$

Dónde C_1,C_2,…,C_k son los conjuntos disjuntos de índices de conglomerados, n_k es el número de observaciones en el k-ésimo conglomerado, y

$d_{i,i'} = \sum_{j=1}^{p} d_{i,i''}, j$: es la medida de disimilitud (aditiva) entre las observaciones i y i′. Cuando nuestras observaciones son vectores X_1,X_2,…,$_n \in \Re^p$ y $d_{i,i'}$ es la distancia euclidiana al cuadrado entre el i-ésimo y i′-ésimo puntos tiene $d_{i,i',j} = \left(x_{ij} - x_{i'j}\right)^2, j = 1,2,\dots,p$ *y*

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} \sum_{j \in C_k} \left\|x_j - \mu_k\right\|^2,$$

donde $\boldsymbol{\mu_k}$ es la media muestral de las observaciones en el k-ésimo grupo. Un algoritmo para encontrar soluciones locales a (2) es el de Lloyd (1982) . Primero se seleccionan K "centros" iniciales al azar $\mu_1, \ldots, \mu_k$ e iterar los siguientes dos pasos hasta la convergencia:

i) Dados los centros de los conglomerados $\mu_1, \ldots, \mu_k$, asigne cada punto al grupo con el centro más cercano.

ii) Dada una asignación de conglomerados, actualice los centros de conglomerados para que sean la media muestral de las observaciones en cada conglomerado.

La función objetivo en cada iteración, puede quedar atrapado en diferentes mínimos locales. Por lo tanto, se inicia varias veces y se devuelve la mejor solución.

Cuesta-Albertos et al. (1997) propuso una modificación de este algoritmo con el fin de obtener clústeres robustos para valores atípicos. La idea principal es reemplazar el paso ii) por:

ii´) Dada una asignación de conglomerados, recorte las observaciones de $\alpha 100\%$ con la mayor distancia a sus centros de conglomerados y actualice los centros de conglomerados para que sean la media muestral de las observaciones restantes en cada conglomerado.

El parámetro de ajuste α regula la cantidad de recorte y es seleccionado por el usuario. Dado que la suma total de cuadrados:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{i`=1}^{n} d_{i,i`}$$

no depende de las asignaciones de grupos, minimizar la suma de cuadrados dentro del grupo es equivalente a maximizar la suma de cuadrados entre grupos:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{i`=1}^{n} d_{i,i`} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i`\in C_k} d_{i,i`} = \sum_{j=1}^{p}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{i`=1}^{n} d_{i,i`} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i`\in C_k} d_{i,i`,j}\right\}$$

El algoritmo SK-means de Witten & Tibshirani (2010) introduce pesos no negativos v_j, j=1,…,p para cada característica y luego se soluciona:

$$\max_{C_{1,\ldots,}C_{k,V}} \sum_{j=1}^{p} v_j \left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{i`=1}^{n} d_{i,i`} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i`\in C_k} d_{i,i`,j}\right\} \quad (3)$$

Sujeto a $\|v\|_2 \leq 1,\ \|v\|_1 \leq l\ y\ v_j > 0, j = 1, \ldots., p,$ donde $l > 1$ determina el grado de escasez (en términos de pesos distintos de cero) de la solución. El problema de optimización en (3) se puede resolver iterando los siguientes pasos:

a. Dados los pesos v y los centros de los conglomerados $\mu_1, \ldots, \mu_k$ resolver

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j=1}^{p} v_j \left(x_{ij} - \mu_{kj}\right)^2,$$

que se obtiene asignando cada punto al grupo con el centro más cercano usando distancias al cuadrado euclidianas ponderadas.

b. Dados los pesos w y las asignaciones de grupos $C_1,\ldots, C_k$, actualiza los centros de los conglomerados para que sean la media muestral ponderada de las observaciones de cada conglomerado.

c. Dadas las asignaciones de clúster $C_1,\ldots, C_k$ y centros $\mu_1,\ldots, \mu_k$ resuelve:

$$\max_{\|v\|_2\leq 1,\|v\|_1\leq l} \sum_{j=1}^{p} v_j\, A_j(C_1,\ldots,C_k),$$

Dónde:

$$A_j(C_1, C_2, \ldots, C_k) = \frac{1}{n}\sum_{i=1}^{n}\sum_{i`=1}^{n} d_{i,i`,j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i`,j}$$

1.9.- K-MEDIANA

No siempre la media es la mejor medida de síntesis de un conjunto de datos, por ejemplo, en presencia de datos muy extremos. Algo análogo sucede con los métodos de cluster basados en cálculo de medias.

El algoritmo K-Mediana es una variación del algoritmo K-Medias. El enfoque de k-medianas es un primer intento de obtener algoritmos de agrupamiento más robustos, dado que la mediana es una medida más robusta que la media ya que no se ve influenciada por los valores extremos. Fue sugerido por MacQueen (1967) y desarrollado por. Kaufman & Rousseeuw (1990).

La función por minimizar está dada por la siguiente expresión:

$$P(V, C_e) = \sum_{l=1}^{k}\sum_{X\in C_l} V_l^{'}\,|X - C_{el}|$$

Donde C_{el} es el vector de medianas del l-ésimo conglomerado y $V = \{V_1, \dots, V_k\}$ es una matriz de pesos con dimensión n×k, cuyos vectores columnas son $V_l = \{v_{1l,\dots,} v_{kl}\}'$, para i=1,…,k, con $\sum_{l=1}^{k} v_{il} = 1, \text{donde } v_{il} \in (0,1)$ para todo i=1,…,n , j=1,…,k.

Desarrollo del algoritmo:

1.-Seleccionar de forma arbitraria K objetos que serán los centroides iniciales de los cluster. Es un algoritmo sensible a la selección de los centroides iniciales.

2.-Se asigna cada punto al cluster con el centroide más cercano. No se utiliza para ello, la distancia euclídea como medida de disimilitud, sino que, en este caso, se utiliza la distancia de Manhattan. Esta distancia define la distancia entre dos puntos *a* y *b* como el sumatorio de las diferencias absolutas entre cada dimensión. Esta medida se ve menos afectada por *outliers* (es más robusta) que la distancia euclídea debido a que no eleva al cuadrado las diferencias.

$$d_{man}(a,b) = \sum_{i=1}^{n} |(a_i - b_i)|$$

3.-A continuación, se calcula el nuevo conjunto de centroides de los clusters, calculando la mediana de los nuevos grupos.

4.-Se lleva a cabo un proceso iterativo hasta que minimice la función objetivo o los centroides no se modifique.

En resumen, se trata de un algoritmo que:

- Usa como centros las medianas y no las medias.

- No se ve afectado por valores outliers.
- Es más robusto.

1.9.1.- GRADIENTE ESTOCÁSTICO K-MEDIANAS

MacQueen (1967) y Hartigan (1975) propusieron por primera vez algoritmos de agrupamiento secuencial. Cardot et al. (2013) han estudiado las propiedades de los algoritmos de gradiente estocástico que pueden dar estimadores recursivos eficientes de la mediana geométrica en espacios de alta dimensión, y formularon una estrategia recursiva que es capaz de estimar los centros de los conglomerados. Una de las principales ventajas de este enfoque recursivo, es que puede manejar grandes conjuntos de datos y es más robusto que el algoritmo k-medias. Que, por su naturaleza recursiva, otra característica importante es que permite la actualización automática y no necesita almacenar todos los datos. Un parámetro de ajuste clave en este algoritmo es el valor del paso de descenso.

Definiciones:

Sea (Ω, A, P) un espacio de probabilidad. Supongamos que tenemos una secuencia de repeticiones independientes $Z_1,. ..,Z_n$ de un vector aleatorio Z tomando valores en $\Re^d$. Se trata de particionar Ω en un número finito de k clúster $\Omega_1,\ldots, \Omega_k$. Cada cluster de Ω_i es representado por su centro, que es un elemento de $\Re^d$ denotado por θ^i. Desde el punto de vista de la población, los algoritmos de las k-medias y k-medianas tienen como objetivo la búsqueda de mínimos locales de una función g de $\Re^{dk}$ a $\Re$ y se definen de la siguiente manera. Sea $x= (x^1, ..., x^k)'$ para todo i , $x^i \in \Re^d$.

$$g(x) \overset{def}{=} E(\underbrace{min}_{r=1,\ldots,k} \emptyset(\|z - x^r\|)) \qquad (1)$$

Donde ϕ es una función real, positiva, continua y no decreciente y la norma $\|.\|$en $\Re^d$ tiene en cuenta la dimensión d de datos, para cada z $\in \Re^d$, $\|z\|^2 = d^{-1}\sum_{j=1}^{d} z_j^2$.

El algoritmo k-medias recursivo propuesto por MacQueen (1967) comienza con k grupos arbitrarios, cada uno con un solo punto $X^1{}_1.,,,.\ X^k{}_1$. Luego, en cada iteración, los centros de los conglomerados se actualizan de la siguiente manera.

$$x_{n+1}^r = x_{n-}^r a_n^r I_r(Z_n; X_n)(X_n^r - Z_n)$$

En donde por $n \geq 2, a_n^r = (1+n_r)^{-1}\ y\ n_r = 1 + \sum_{l=1}^{n-1} I_r\ (Z_l;\ X_l)$ es el número de elementos asignados al clúster r hasta la iteración n-1. Para n =1, permite $a_1^r = \frac{1}{2}$. Esto también significa que $X^r{}_{n+1}$ es el centroide de los elementos asignados al clúster r hasta la iteración n.

$$X_{n+1}^r = \frac{1}{1+\sum_{l=1}^{n} I_r(Z_l; X_l)}\left(X_1^r + \sum_{l=1}^{n} I_r\ (Z_l; X_l)Z_l\right)$$

1.9.2.- AJUSTE DEL GRADIENTE ESTOCÁSTICO K-MEDIANAS Y SU VERSIÓN PROMEDIO

Suponiendo Z tiene una distribución absolutamente continua tenemos:

$P\left(\left\|Z - x^i\right\| = \left\|Z - x^j\right\|\right) = 0$, para cualquier i≠j y $x^i \neq x^j$

Entonces, el enfoque k-medianas se basa en la búsqueda de los mínimos, que pueden ser locales, de la función g que se escribe de la siguiente manera, para cualquier x tal que $x^i \neq x^j$ cuando i≠j.

$$g(x) = \sum_{r=1}^{k} E[I_r\ (Z; x)\|Z - x^r\|] \qquad (2)$$

El gradiente estocástico del algoritmo k-medianas, se presenta al tener un conjunto de inicialización de k puntos distinto en $\Re^d$, $X^1{}_1,\ldots,X^k{}_1$, en donde el conjunto de los centros de los clusters se actualiza en cada iteración de la siguiente manera. Para r=1,…,k, y n≥1,

$$X_{n+1}^r = X_n^r - a_n^r I_r(Z_n; X_n) \frac{X_{n-}^r Z_n}{\|X_n^r - Z_n\|}$$

$$= X_n^r - a_n^r \nabla_r g(X_n) - a_n^r V_n^r \qquad (3)$$

El comportamiento de algoritmo (3) depende de la secuencia de pasos $a^r{}_n$, r∈{1,…,k} y el vector de inicialización X_1.

A partir de una muestra de n realizaciones Z_1, ..., Z_n de Z y un conjunto de puntos de inicialización del algoritmo, la estimación seleccionada de los centros de los conglomerados es la que minimiza el siguiente riesgo

$$R(X_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{k} I_r\,(Z_i; X_n) \|Z_i - X_n^r\| \qquad (4)$$

Denotemos por $n_r = 1 + \sum_{l=1}^{n-1} I_r\,(Z_l; X_l)$ el número de pasos para la actualización del clúster r, hasta la iteración n-1, para r∈{1,…,k}.

Un clásico de los pasos del descenso puede ser dado por

$$a_n^r = \begin{cases} a_{n-1}^r & Si\ I_r(Z_n; X_n) = 0 \\ \frac{C_\gamma}{(1+C_\alpha n_r)^\alpha} & otro\ caso \end{cases} \qquad (5)$$

Donde $C_\gamma, C_\alpha \quad 1/2 \leq \alpha \leq 1\ control\ de\ ganancia$

En (5) son muy sensible a los valores de los parámetros C_γ y C_α que tienen que ser elegido con mucho cuidado (Cardot et al., 2012).

El estimador promedio de los centros de los clusters, que sigue siendo recursivo, se define como sigue, sea r∈{1,…,k}, n ≥1 y el valor de X^r_{n+1} obtenido mediante la combinación de (3) y (5),

$$\bar{X}^r_{n+1} = \begin{cases} \bar{X}^r_n & si\ I_r(Z_n; X_n) = 0 \\ \dfrac{n_r \bar{X}^r_n + X^r_{n+1}}{n_r + 1} & otro\ caso, \end{cases}$$

1.10.-MÉTODOS DE PARTICIONAMIENTO

Entre los algoritmos de partición se encuentran:

- PAM (Partitioning Around Medoids)
- CLARA (Clustering Large Applications)
- CLARANS (Clustering Large Applications based upon RANdomized Search).

1.10.1.-PAM (Partitioning Around Medoids)

El algoritmo Partitioning Around Medoids (PAM) es un método tipo k-medoids que intenta determinar k particiones de n objetos determinando los objetos representativos de cada conglomerado (Kaufman & Rousseeuw, 1990a).

Para encontrar los k medoids, PAM empieza con una selección arbitraria de k objetos representativos. Este objeto representativo, llamado *medoid*, es el que se encuentra localizado más al centro dentro del *cluster*. Una vez que los *medoids* han sido seleccionados. cada objeto no seleccionado es agrupado con el *medoid* al cual es más similar. En cada iteración hace un intercambio entre un objeto seleccionado, Q_i, y uno no seleccionado Q_h, si y solo si el intercambio mejora la calidad del agrupamiento. El efecto de tal intercambio entre Q_i y Q_h se mide a través de una función de costo, es decir, el algoritmo calcula los costos C_{jih} para todos los objetos no seleccionado Qj.

Si denotamos por $Q = \{Q_1, \dots, Q_N\}$al conjunto de N objetos, se trata de dividir Q en K grupos, $Cl_1, \dots, Cl_k$ de forma que:

$$\cup_{j=1}^{k} Cl_j = Q$$

$$Cl_i \cap Cl_j = \phi \; para \; i \neq j$$

El objeto representativo, llamado *medoid*, es reconocido como el objeto mejor ubicado cerca de la parte central del conglomerado.

Una vez que los medoids han sido seleccionados, cada objeto no seleccionado es agrupado con el medoid con el que guarda más similitudes.

Denotamos por:

Q_j: Objeto no seleccionado

Q_m: medoid seleccionado

De manera más precisa, si Qj, es un objeto no seleccionado y Q_m es un medoid (seleccionado), decimos que Qj pertenecerá al conglomerado representado por Q_m si:

$$d(Q_j ; Q_m) = \min Oe\ d(Q_j ; Oe)$$

donde la notación min denota el mínimo entre todos los medoids Oe y la notación d (Q_1 ; Q_2) denota la disparidad o distancia entre los objetos Q_1 y Q_2. Todos los valores de disparidad están dados como entradas a PAM. Finalmente, la calidad de las conglomeraciones es medida por un promedio de disparidad entre el objeto y el medoid de su conglomerado. Para hallar los k-medoids, PAM comienza con una selección arbitraria de *k* objetos. Luego, en cada paso, se realiza un intercambio entre el objeto seleccionado Q_m *y* el no seleccionado Qp, siempre y cuando ese intercambio resulte en una mejora de la calidad del conglomerado. Para establecer el efecto del intercambio entre Q_m y Q_p, PAM computa los costos *Cjmp* para todos los objetos no-medoids Qj Oj.

El costo total de reemplazar a Q_m con Q_p está dado por: $TC_{mp} = \sum_j C_{jmp}$

La estructura del algoritmo PAM es la siguiente:

1.-Seleccionar arbitrariamente k objetos representativos, los cuales serán los k-medoids iniciales.

2.-Ingresar el *TCmp* para todos los pares de objetos Q_m; Q_p *donde* Q_m ya se encuentre seleccionado, y Q_p *no*.

3. Seleccionar el par (Q_m; Q_p) que corresponde a minQ_m;Q_p TCmp .

4. Si el mínimo *TCmp* es negativo, reemplace Q_m con Q_p y vuelva al paso 2. De lo contrario, para cada objeto no-seleccionado, se debe encontrar el objeto representativo más similar.

Los resultados experimentales muestran que PAM trabaja satisfactoriamente para bases de datos pequeñas (como por ejemplo 100 objetos en 5 conglomerados). Pero, no es eficiente el manejar bases de datos medias o grandes.

El nivel de complejidad de una interacción en PAM es elevada y muy costoso en tiempo computacional para gran cantidad de valores. Este análisis motiva el desarrollo de CLARA.

Schubert y Rousseeuw (2019) proponen una modificación del algoritmo PAM cambiando el orden de anidación de los bucles en el algoritmo. Proponen encontrar el mejor intercambio para cada medoid y ejecutar tantos como sea posible en cada iteración, lo que reduce el número de iteraciones necesarias para la convergencia sin pérdida de calidad.

1.10.2.-CLARA (Clustering Large Applications)

Clara fue introducido por Kaufman y Rousseeuw (1990) y es un método de partición utilizado para tratar conjuntos de datos mucho más grandes (más de varios miles de observaciones) con el fin de reducir el tiempo de computación y el problema de almacenamiento RAM.

La diferencia entre PAM y CLARA es que el segundo se basa en muestreos. Solo una pequeña porción de los datos totales es seleccionada como representativa de los datos y los *medoids* son escogidos de la muestra usando PAM. La intuición es si la muestra es seleccionada de manera aleatoria, entonces es representativa del conjunto total de datos, y los objetos representativos escogidos (*medoids*), serán similares tal y como si hubieran sido escogidos del conjunto total de datos.

Es un método que permite trabajar con matrices más grandes que el algoritmo PAM.

A continuación, se presenta el algoritmo CLARA con *n* muestras de tamaño *t*.

Pasos del algoritmo:

1. Para $i = 1$ hasta n, repetir los siguientes pasos:
2. Seleccionar una muestra de *t* objetos de forma aleatoria del conjunto total de datos, y llamar al algoritmo PAM para encontrar *k medoids* de la muestra.
3. Para cada objeto Q_j *del* conjunto total de datos, determinar cuál de los *k medoids* es el más similar a Q_j.
4. Calcular la disimilaridad promedio del agrupamiento obtenido en el paso anterior. Si este valor es menor al mínimo actual, usar este valor como el mínimo actual y retener los *k medoids* encontrados en el paso (2) como el mejor conjunto de *medoids* obtenidos.
5. Retornar al paso (1) para comenzar la próxima iteración.

El CLARA al aplicar PAM encuentra los medoids en una muestra del conjunto de datos. Si las muestras son suficientemente aleatorias, los medoids de la muestra se aproximan a los medoids del conjunto de datos. Funciona bien para conjuntos de grandes de datos.

1.10.3.-CLARANS (Clustering Large Applications based upon Randomized Search).

Es una mezcla de PAM y CLARA. Las búsquedas las realiza sobre un subconjunto del conjunto de datos y no se limita a ninguna muestra. Mientras CLARA tiene una muestra

fija en cada etapa de la búsqueda, CLARANS forma una muestra aleatoria en cada etapa de la búsqueda.

Pasos del algoritmo:

1. Dar como datos de entrada los parámetros *numlocal* y *maxneighbor*. Inicializar *i* a 1, y *mincost* a un número mayor.
2. Establecer *current* a un nodo arbitrario en *Gn,k*.
3. Establecer *j* a 1.
4. Considerar un vecino aleatorio *R* de *current*, y basado en la ecuación (5), calcular el costo diferencial de los 2 nodos.
5. Si *R* tiene un costo menor, establecer *current* a *R*, e ir al paso (3).
6. De lo contrario, incrementar *j* en 1. Si $j \leq maxneighbor$, ir al paso (4).
7. De lo contrario, cuando $j > maxneighbor$, comparar el costo de *current* con *mincost*. Si éste es menor a *mincost*, establecer *mincost* al costo de *current*, y establecer *bestnode* a *current*.
8. Incrementar *i* en 1. Si $i > numlocal$, el resultado es *bestnode* y terminar. De lo contrario, ir al paso (2).

El principio por la cual se realiza este algoritmo es encontrar una muestra con una cierta aleatoriedad en cada paso de la búsqueda. El agrupamiento obtenido después de sustituirlo a un solo medoid se denomina el *vecino del agrupamiento actual*. Si en el camino el objeto encuentra un mejor vecino, CLARANS lo mueve al nodo del vecino y el proceso comienza de nuevo; si ya no lo encuentra entonces el agrupamiento actual para y se produce un óptimo local (Cluster).

1.11.-DBSCAN

La mayoría de las técnicas tradicionales como el K-Means requiere especificar el número de clusters e incluyen valores atípicos. Sin embargo, los algoritmos de agrupación en clústeres basados en densidad son muy eficaces para encontrar regiones de alta densidad y la detección de anomalías.

Los algoritmos para clustering basado en densidad identifican regiones de alta densidad que están rodeadas de áreas poco densas. Cada una de las regiones densas identificadas se corresponde con un cluster. Son adecuados cuando los clusters no tienen una forma geométrica definida y localizan regiones de alta densidad que están separadas una de la otra por regiones de baja densidad. En este contexto se entiende el concepto de densidad como el número de puntos dentro de un radio determinado.

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) es un algoritmo propuesto por Ester et al. (1996) basado en la densidad, especialmente diseñado para la identificación de agrupaciones en un contexto espacial.

DBSCAN para obtener clusters útiles requiere que se fijen adecuadamente dos parámetros:

- Puntos mínimos (MinPts): número mínimo de puntos que queremos para que una región se considere densa, es decir, la cantidad de datos alrededor de una observación para definir un cluster.

  MinPts $=\sum_{i=1}^{n} d_i$

donde d_i es el i-ésimo valor de cada densidad medida en la vecindad de un contenido puntual en un conjunto de datos específico D con n número de experimentos

- Radio (épsilon): se define como la distancia al k-ésimo punto más cercano. Cuando incluye suficientes puntos dentro de él se considera "área densa".

Pasos del algoritmo:

1.-Se fija el valor de los puntos mínimos y radio.

2.-Se selecciona al azar un punto de partida en su área de vecindad, y se determina usando el radio. Si hay un número mínimo de puntos en su área comienza la formación del grupo. Cada punto se clasifica como central (core), frontera (border) o ruido (noise):

- Un punto de datos es central si hay al menos un número mínimo de puntos (incluido el punto en sí) en su área circundante con radio épsilon.
- Un punto es fronterizo si es accesible desde un punto central y hay menos del número de puntos fijados dentro de su área circundante
- Se considera un punto de ruido aquel que no pertenecen a ningún grupo. Esto ocurre si el punto arbitrario escogido (eps) tiene menos del mínimo de puntos en su círculo de radio.

3.-Se elige aleatoriamente otro punto entre los puntos que no han sido visitados en los pasos anteriores y se aplica el mismo procedimiento.

4.-El proceso finaliza cuando todos los puntos han sido visitados.

Resumiendo, un cluster se forma con al menos un punto central, todos los puntos centrales accesibles y todos sus puntos fronterizos.

DBSCAN ofrece un rendimiento superior al de técnicas conocidas como K-Means o algoritmos jerárquicos, dado que no es necesario determinar a priori el número de clusters deseados. Los parámetros de ajuste de DBSCAN son mínimos, ya que está diseñado para realizar un agrupamiento no supervisado como una búsqueda exploratoria de características en grandes conjuntos de datos; además, fue desarrollado para mantener un bajo costo computacional (Perafán-López & Sierra-Pérez., 2020).

DBSCAN no presupone clusters convexos, sino que se basa en la densidad de las muestras para identificar los clusters, por este motivo, los clusters identificados pueden ser de cualquier forma.

Es un algoritmo útil para la detección de valores atípicos dado que considera como "agrupados" todos aquellos puntos de las zonas más densas (normalmente puntos válidos) y considerará como anormales aquellos puntos alejados y en zonas poco densas.

1.12.-AGRUPAMIENTO EN CLUSTER: REPRESENTACIÓN GRÁFICA

La salida de un algoritmo de agrupación de partición como K-medias, K-medoids entre otros es una lista de clusters y sus objetos que pueden ser difíciles de interpretar. Por tanto, sería útil disponer de una representación gráfica que describe los objetos, y al mismo tiempo muestre los clusters. Esto nos permitiría imaginar el tamaño y la forma de los clusters, así como su posición relativa.

De las gráficas útiles para la interpretación y validación del análisis de cluster de partición, tenemos la llamadas CLUSPLOT de Kaufman & Rousseeuw (1990) y las SILUETAS.

1.12.1.- CLUSPLOT

La salida de un método de partición es simplemente una lista de clusters y sus objetos, que pueden ser difíciles de interpretar. Por tanto, sería útil disponer de una representación gráfica que describe los objetos con sus interrelaciones, y al mismo tiempo muestra los clusters, lo que permitiría imaginar el tamaño y la forma de los clusters, así como su posición relativa.

ClusPlot es una nueva forma de representar clusters en el cual los objetos son representados como puntos en un gráfico bidimensional y los cluster como elipses de varios tamaños y formas (Pison, Rousseeuw, & Struyf, 1999).

Existen varios algoritmos en la formación de ClusPlot. Kaufman y Rousseeuw (1990) estudian para algoritmos de tipo jerárquico: AGNES, DIANA y MONA y para algoritmos de tipo no jerárquico: PAM, CLARA, FANNY. Según los autores ClusPlot es una representación gráfica que usa la salida que le proporciona un algoritmo de datos particionados para poder visualizar los individuos y los clusters formados por dichos individuos según sus variables sobre un gráfico en dos dimensiones.

El ClusPlot utiliza la partición resultante, así como los datos originales, para producir la siguiente figura:

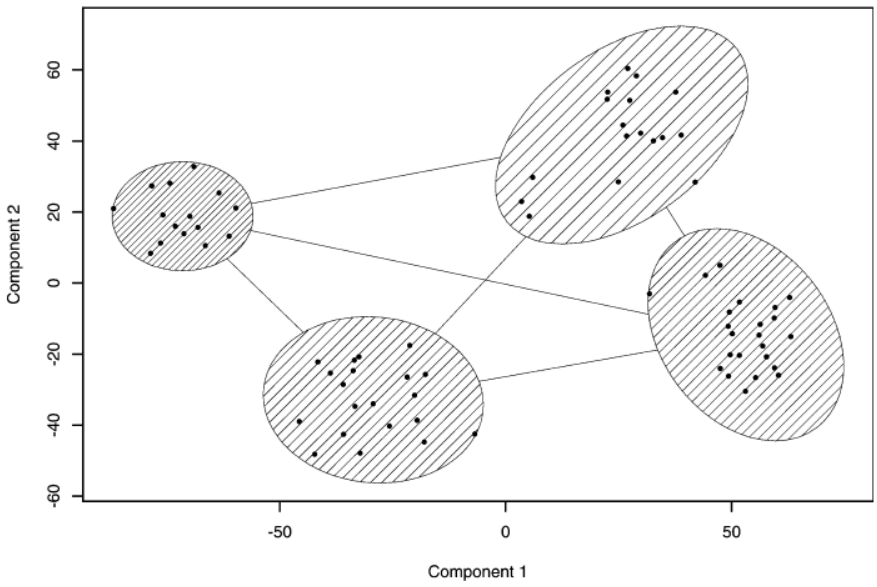

Figura 2. Clusplot de la tabla de Rusipini de 75 elementos y 4 clúster. (Fuente: Pison et a1999).

Las elipses se basan en la media y la matriz de covarianza de cada cluster, y su tamaño es tal que contienen todos los puntos de su clúster, lo que explica por qué siempre hay un objeto en el límite de cada elipse. También es posible dibujar la elipse de expansión de cada grupo, es decir, la elipse más pequeña que cubre todos sus objetos. La elipse de expansión se puede calcular con el algoritmo de Titterington (1976).

Para tener una idea de las distancias entre los clusters, podemos dibujar segmentos de las líneas entre los centros del clúster. En el gráfico, la intensidad de sombreado es proporcional a la densidad del conglomerado, es decir, su número de objetos dividido por el área de la elipse.

Para dimensiones superiores para ($p > 2$) $X=(X_{i1},X_{i2},\ldots,X_{ip})$, $i=1,\ldots,n$, podemos reducir la dimensión de los datos a través de las componentes principales que produce una primera componente con varianza máxima, luego una segunda componente con varianza máxima entre todos los componentes perpendiculares al primero, y así sucesivamente.

Los principales componentes se encuentran en las direcciones de los vectores propios de una matriz de dispersión, que puede ser la matriz de covarianza clásica o la matriz de correlación correspondiente. Otra posibilidad es partir de una matriz de dispersión robusta (que puede resistir el efecto de valores atípicos), como el elipsoide de volumen mínimo y los estimadores determinantes de covarianza mínima de Rousseeuw (1984).

Pison et al. (1999) desarrollan la función ClusPlot en el software S-Plus mediante un gráfico bidimensional, mejorando las herramientas anteriores que incluían la distancia(Chen, Gnanadesikan, & Kettenring, 1974) o gráficos de silueta (Rousseeuw, 1987). El algoritmo CLUSPLOT puede también ser visto como una versión generalizada y automatizada de mapas taxométricos los cuales representaban los clusters pero no los objetos (Carmichael & Sneath, 1969).

Los gráficos que proporciona la función ClusPlot pueden tener aspectos diferentes dependiendo del algoritmo que se utilice, así por ejemplo, si se usa el algoritmo Eddy (1977) se reemplaza cada elipse por una forma convexa de todos los puntos que están en el cluster o por el bugplot del cluster (Rousseeuw & Ruts, 1997). Otra posibilidad es dibujar elipses basadas en el estimador del Determinante de Covarianza Mínima, que es fácilmente obtenido con la función cov.mcd (Rousseeuw & Van Driessen, 1997). E incluso la expansión de la elipse puede ser computada con el algoritmo de Titterington (1976).

Casos particulares de ClusPlot:

a) U*n clúster solo contiene a un individuo*, entonces se dibuja un círculo alrededor de dicho individuo.

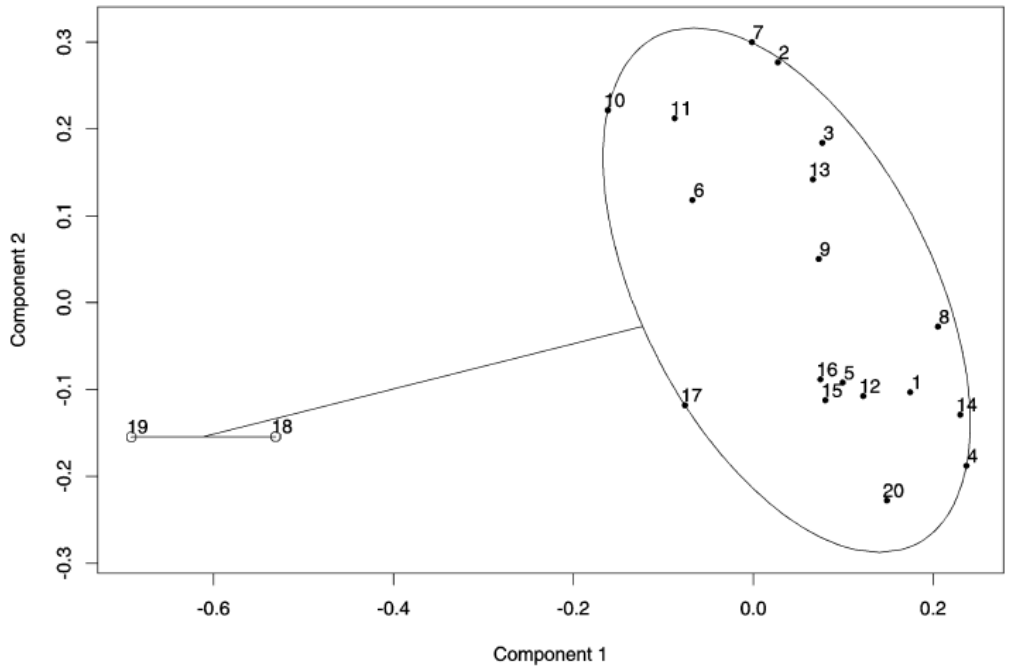

Figura 3. Clusplot de datos de disimilitud de Abbot y Perkins (1978) con un cluster de solo dos puntos. (Fuente: Pison et al. 1999).

b) ***Cuando los individuos de un clúster caen sobre una line recta.*** Con el argumento "span = FALSE", se obtiene una estrecha elipse alrededor de la línea, y si se utilizará "span = TRUE", se obtendría una línea recta exacta.

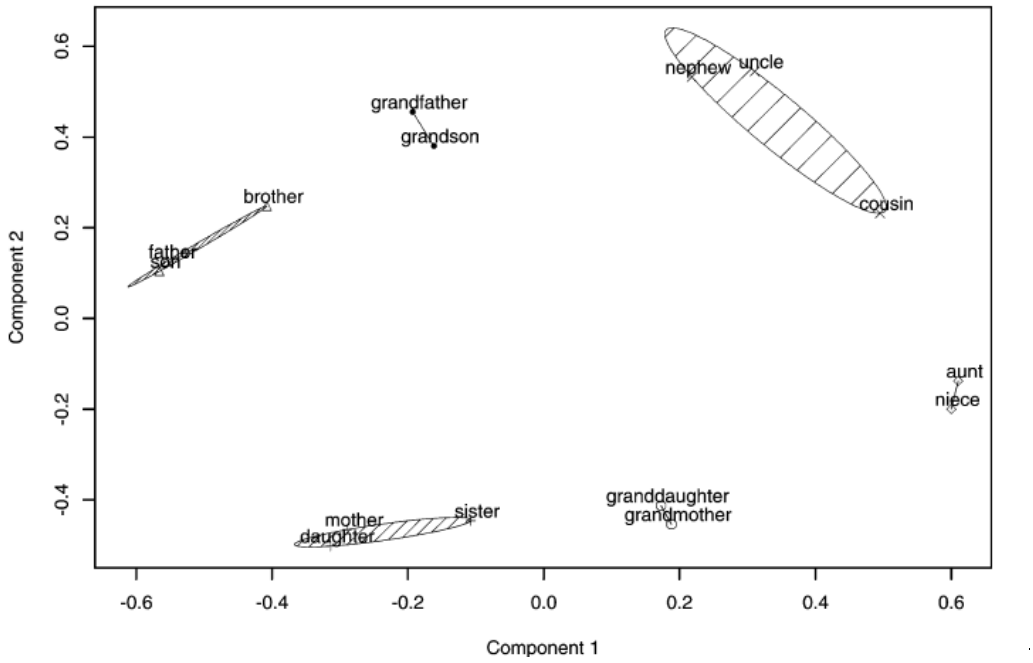

Figura 4. Clusplot de los datos de disimilitud de Rosenberg (1982), con muchos cluster de pocos objetos. (Fuente: Pison et al. 1999).

Cálculo de la Matriz de Disimilaridades

Otra posibilidad de aplicación del ClusPlot es a través del método de escalamiento multidimensional (MDS), cuando los datos se presentan como una matriz de disimilitudes.

Las disimilitudes son números no negativos d (i, j) que son pequeños cuando i y j son cerca el uno del otro y que se hacen grandes cuando i y j son muy diferente. Tienen dos propiedades:

$$d(i,i) = 0$$

$$d(i,j) = d(j,i)$$

Las disimilitudes entre dos objetos pueden calcularse de diferentes maneras, por ejemplo, cuando los datos contienen variables nominales u ordinales, sino también las medidas subjetivas de discordancia están permitidas. Se dice que la matriz $D = (d(i,j); 1 \leq i,j \leq$

n) es métrica si, además de las dos propiedades anteriores, también sostienen la desigualdad triangular:

$$d(i,j) \leq d(i,h) + d(h,j)$$

En ese caso, las disimilitudes son llamadas distancia.

En general, un método de escalado multidimensional (MDS) construye un conjunto de n puntos, que se caracteriza por sus coordenadas en relación con algunos ejes, de tal manera que las distancias euclidianas entre estos n puntos se aproximan a las disimilitudes originales. La aproximación será mejor cuando la matriz de disimilitud D sea métrica. Tenga en cuenta también que un método MDS produce componentes de tal manera que la primera componente explica tanto la variabilidad como sea posible, el segundo componente explica como gran parte de la variabilidad restante como sea posible y así sucesivamente.

ClusPlot se aplica en un método de MDS a D (que puede o no ser métrico), y a continuación, muestra los dos primeros componentes. El porcentaje de variabilidad explicada por estos dos componentes (en relación con todos los componentes) aparece por debajo del siguiente gráfico.

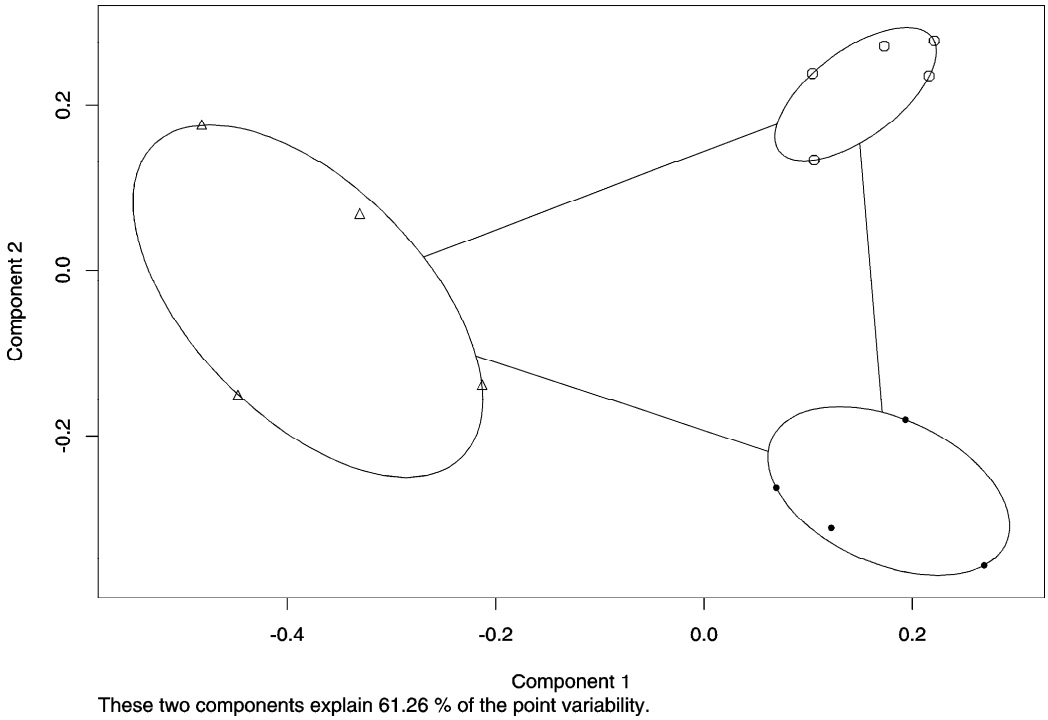

Figura 5. Clusplot de los datos de disimilitud de Harman (1967). (Fuente: Pison et al., 1999).

1.12.2.-SILUETAS

Silueta se refiere a un método de interpretación y validación de consistencia dentro de grupos de datos . Proporciona una representación gráfica sucinta de lo bien que cada objeto se encuentra dentro de su grupo. Se trata de una métrica para evaluar el buen funcionamiento de los algoritmos de aprendizaje no supervisado.

En los algoritmos de aprendizaje no supervisado el número de clústeres puede venir determinado de dos formas:

- Puede ser un parámetro de entrada del algoritmo (k-. medias).
- Puede determinarse automáticamente por el algoritmo (DBSCAN).

Los rangos de silueta de -1 a 1, en donde un valor alto indica que el objeto está bien adaptado a su propio clúster y mal adaptado a agrupaciones vecinas. Si la mayoría de los

objetos tienen un valor alto, entonces la configuración de agrupamiento es apropiada. Si muchos puntos tienen un valor bajo o negativo, entonces la configuración de agrupación puede tener demasiados o demasiado pocos grupos.

El coeficiente de silueta es indicador del número ideal de clústeres. Se puede calcular con cualquier distancia métrica, tal como la distancia euclidiana o la distancia Manhattan .

El coeficiente de Silueta para un conjunto está dato como la media del coeficiente de Silueta de cada objeto de la muestra, s(i). El coeficiente de Silueta para un objeto es:

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & si\ a(i) < b(i) \\ 0 & si\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & si\ a(i) > b(i) \end{cases}$$

Dónde:

a: es la distancia promedio entre el objeto y todos los otros objetos de la misma clase.

b: es la distancia promedio entre el objeto y todos los otros objetos del

Incluso es posible escribir esto en una fórmula:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

s(i) varía entre -1 y 1. Si está cerca de 1, se puede decir que el objeto está "bien clasificado". Si está cerca de 0, no está claro si debe pertenecer al clúster o a su vecino. Un valor negativo sugiere que se ha clasificado erróneamente.

Proceso del trazado de la silueta

Una vez calculado las cantidades s(i) de las similitudes o disimilitudes, se procede a construir la representación gráfica. La silueta del cluster A es un gráfico de s(i), clasificado en orden decreciente, para todos los objetos i en el cluster A. En una impresora de línea, se representan los s(i) por una fila de asteriscos, la longitud de los cuales es proporcional a s(i).

Con el fin de obtener una visión general, la silueta de los diferentes clusters se imprimen una debajo de la otra. De esta manera toda la agrupación se puede visualizar por medio de un solo gráfico, lo que nos permite distinguir entre los clusters bien definidos de los débiles.

Las dimensiones de una silueta son la anchura y la altura. La anchura lo determina los valores de s(i). Una amplia silueta indica grandes valores s(i), y por lo tanto un clúster pronunciado y una estrecha silueta es indicativo de pequeños valores s(i) y un cluster no muy abultado. La otra dimensión de una silueta es su altura, que es igual al número de objetos en A.

Ventajas de la silueta

Las siluetas se utilizan para:

- ✓ Mejorar los resultados del análisis de cluster.

- ✓ La salida de diferentes algoritmos de agrupación aplicados a los mismos datos.
- ✓ Interpretar y validar los resultados del análisis de cluster.

Las siluetas solo dependen de la partición real de los objetos y no del algoritmo de agrupación que se utilizó para obtenerlas. Como consecuencia, las siluetas podrían usarse para mejorar los resultados del análisis de conglomerado o para comparar la salida de diferentes algoritmos de conglomerado aplicados a los mismos datos. Sin embargo, la principal utilidad de las siluetas radica en la interpretación y validación de los resultados del análisis de conglomerados. Para obtener una visión general, las siluetas de los diferentes grupos se imprimen una debajo de la otra. De esta manera, la agrupación completa se puede mostrar mediante un solo gráfico, lo que nos permite distinguir las agrupaciones "claras" de las "débiles".

1.13.-SOFTWARE DE ALGORITMOS DE AGRUPAMIENTO

El lenguaje de programación permite especificar de manera precisa sobre qué datos debe operar un software específico, cómo deben ser almacenados o transmitidos dichos datos, y qué acciones debe tomar el software bajo una variada gama de circunstancias. El representar los datos por una serie de clusters, conlleva la pérdida de detalles, pero consigue la simplificación de los mismos. Agrupamiento es una técnica más de Aprendizaje Automático, en la que el aprendizaje realizado es no supervisado. Desde un punto de vista práctico, el agrupamiento juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como

GIS o datos procedentes de astronomía), aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, y muchas otras.

En los últimos años han surgido una gran variedad de software que desarrollan algoritmos de agrupamiento, a continuación, proyectamos nuestra mirada a un lenguaje de código abierto llamado R.

1.13.1.-LENGUAJE DE PROGRAMACIÓN R

R de Ihaka & Gentleman (1996) es un sistema para ´análisis estadísticos y gráficos y es un lenguaje orientado a objetos. Una ventaja de R es que es un lenguaje interpretado que no requiere de la realización de un programa que se compilará después. Los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir ejecutables. El número de paquetes que se le pueden añadir a R crece muy rápidamente, pues la comunidad de R es muy dinámica. Proporciona acceso a una amplia variedad de técnicas estadísticas y graficas.

Los resultados de análisis estadísticos se muestran en la pantalla, y algunos resultados intermedios se pueden guardar, exportar a un archivo, o ser utilizados en análisis posteriores.

R posee muchas funciones para análisis estadísticos y gráficos; estos últimos pueden ser visualizados de manera inmediata en su propia ventana y ser guardados en varios formatos.

Al lenguaje R se le pueden instalar diferentes librerías, cada una de las cuales tiene una serie de funciones, que producen resultados específicos.

Las funciones disponibles están guardadas en una librería localizada en el directorio R HOME/library (R HOME es el directorio donde R está instalado). Este directorio contiene *paquetes* de funciones, las cuales a su vez están estructuradas en directorios. El paquete denominado base constituye el núcleo de R y contiene las funciones básicas del lenguaje para leer y manipular datos, algunas funciones gráficas y algunas funciones estadísticas (componentes principales y análisis de cluster). Cada paquete contiene un directorio denominado R con un archivo con el mismo nombre del paquete (por ejemplo, para el paquete base, existe el archivo R HOME /library/base/R/base). Este archivo está en formato ASCII y contiene todas las funciones del paquete.

1.13.2.-PAQUETES EN R PARA EL DESARROLLO DE ALGORITMOS DE AGRUPAMIENTO

En la siguiente tabla se presenta información sobre paquetes en el entorno estadístico de R para el desarrollo de algoritmo de agrupamiento no jerárquico.

Tabla 1. Paquetes que desarrollan algoritmos de agrupamiento no jerárquico en el lenguaje *R*

Nombre del método	Paquete	Fecha de creación y de modificación	Breve Descripción
K-Medias	stats versión 3.6.2	12-12-2019	Realiza agrupaciones de k-medias en una matriz de datos.
K-Medoids	Cluster_Medoids versión 1.2.1	28-11-2019	Partición alrededor de medoids
K-Media Difusa	cmeans versión 1.7-3	25-11-2019	Genera agrupaciones difusas, en donde cada elemento tiene una probabilidad de pertenecer a cada grupo
K-medias Sparse Robusto	RSKC versión 2.4.2	11-08-2016	Contiene una función RSKC que ejecuta el robusto algoritmo de agrupamiento de K medias dispersos
PAM (Partitioning Around Medoid s)	pam versión 2.1.0	07-06-2019	Partición de los datos en k agrupaciones "alrededor de medoides", una versión más robusta de K-medias.
CLARA Medoids	Clara_Medoids versión 1.2.1	28-11-2019	Agrupaciones a partir de matrices de tamaño grande
Gradiente estocástico K-Mediana	kGmedian versión 1.2.5	02-03-2020	Técnica de agrupación k-mediana realizada de forma recursiva
DBSCAN	dbscan versión 1.1-5	29-10-2019	Algoritmos basados en densidad de la familia DBSCAN para datos espaciales
ClusPlot	clusplot versión 2.1.0	07-06-2019	Gráfico de conglomerados bivariados
Silhouette	silhoutte versión 2.1.0	07-06-2019	Calcular información de silueta de la agrupación
Funciones extras	dentroextras Versión 0.2.3.	24-01-2018	Proporciona funciones adicionales para cortar, etiquetas y colocar grupos de dendogramas

A continuación, se presenta los argumentos y parámetros que se utilizan en el desarrollo de cada uno de los paquetes.

- **Paquete stats**

El paquete stats versiòn 3.6.2, realiza agrupaciones de k-medias en una matriz de datos. El argumento que se utiliza es:

```
kmeans(x, centers, iter.max = 10, nstart = 1,
    algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                "MacQueen"), trace=FALSE)
```

dónde:

x	matriz numérica de datos
centers	número de conglomerados, digamos k, o un conjunto de centros de conglomerados iniciales (distintos)
iter.max	el número máximo de iteraciones permitido
nstart	número de conjuntos aleatorios que deben elegirse
algorithm	carácter: puede abreviarse. Tenga en cuenta que "Lloyd" y "Forgy" son nombres alternativos para un algoritmo
trace	número lógico o entero, actualmente solo se utiliza en el método predeterminado ("Hartigan-Wong"): si es positivo (o verdadero), se genera información de seguimiento sobre el progreso del algoritmo. Los valores más altos pueden producir más información de rastreo.

- **Paquete pam**

El paquete pam versión 2.1.0, genera una partición de los datos en k agrupaciones "alrededor de medoids", una versión más robusta de K-medias. El argumento que se utiliza es:

```
pam(x, k, diss = inherits(x, "dist"),
```

metric = c("euclidean", "manhattan")

dónde:

x	matriz numérica de datos
K	entero positivo que especifica el número de conglomerados, menor que el número de observaciones
diss	Si es VERDADERO (predeterminado para objetos de disimilitud o dist), entonces x se considerará como una matriz de disimilitud. Si es FALSO, entonces x se considerará como una matriz de observaciones por variables
metric	cadena de caracteres que especifica la métrica que se utilizará para calcular las diferencias entre las observaciones

- **Paquete Cluster_Medoids**

El paquete Cluster_Medoids versión 1.2.1, crea una partición alrededor del medoids.

El argumento que se utiliza es:

```
Cluster_Medoids(data, clusters, distance_metric = "euclidean",
  minkowski_p = 1, threads = 1, swap_phase = TRUE, fuzzy = FALSE,
  verbose = FALSE, seed = 1)
```

dónde:

data	matriz o marco de datos. El parámetro de datos también puede ser una matriz de disimilitud, donde la diagonal principal es igual a 0.0 y el número de filas es igual al número de columnas
clusters	número de clúster
distance_metric	una cadena que especifica el método de distancia. Puede ser, euclidiana, manhattan, chebyshev, canberra, braycurtis, pearson_correlation, simple_matching_coefficient, minkowski, hamming, jaccard, coseno, Coefficient, Rao_coefficient, mahalanobis
minkowski_p	un valor numérico que especifica el parámetro minkowski en caso de que distance_metric = "minkowski"
threads	número entero que especifica el número de núcleos que se ejecutarán en paralelo
swap_phase	VERDADERO o FALSO. Si es VERDADERO, entonces se llevarán a cabo ambas fases ('construcción' e 'intercambio'). El 'swap_phase' se considera más intensivo en computación
fuzzy	VERDADERO o FALSO. Si es VERDADERO, las probabilidades de cada grupo se devolverán según la distancia entre las observaciones y los medoids
verbose	VERDADERO o FALSO, que indica si el progreso se imprime durante la agrupación
seed	valor entero para generador de números aleatorios (RNG)

- **Paquete Clara_Medoids**

  El paquete Clara_Medoids versión 1.2.1, crea agrupaciones a partir de matrices de tamaño grande.

El argumento que se utiliza es:

```
Clara_Medoids(data, clusters, samples, sample_size,
distance_metric = "euclidean", minkowski_p = 1, threads = 1,
swap_phase = TRUE, fuzzy = FALSE, verbose = FALSE, seed = 1)
```

dónde:

data	matriz o marco de datos. El parámetro de datos también puede ser una matriz de disimilitud, donde la diagonal principal es igual a 0.0 y el número de filas es igual al número de columnas
clusters	número de clúster
Sample	número de muestras para extraer del conjunto de datos
Sample size	fracción de datos para extraer en cada iteración de muestra. Debe ser un número flotante mayor que 0.0 y menor o igual a 1.0
distance_metric	una cadena que especifica el método de distancia. Puede ser, euclidiana, manhattan, chebyshev, canberra, braycurtis, pearson_correlation,simple_matching_coefficient, minkowski, hamming, jaccard, coseno, Coefficient, Rao_coefficient, mahalanobis
minkowski_p	un valor numérico que especifica el parámetro minkowski en caso de que distance_metric = "minkowski".
threads	número entero que especifica el número de núcleos que se ejecutarán en paralelo.
swap_phase	VERDADERO o FALSO. Si es VERDADERO, entonces se llevarán a cabo ambas fases ('construcción' e 'intercambio'). El 'swap_phase' se considera más intensivo en computación
fuzzy	VERDADERO o FALSO. Si es VERDADERO, las probabilidades de cada grupo se devolverán según la distancia entre las observaciones y los medoids
verbose	VERDADERO o FALSO, que indica si el progreso se imprime durante la agrupación
seed	valor entero para generador de números aleatorios (RNG)

- **Paquete clusplot**

El paquete clusplot versión 2.1.0, dibuja un "clusplot" bidimensional (gráfico de agrupamiento). La función genérica tiene un método predeterminado y de partición. El argumento que se utiliza es:

```
clusplot(x, ...)

## S3 method for class 'partition'
```

clusplot(x, main = NULL, dist = NULL, ...)

dónde:

x	un objeto de la clase "partición"
main	título del grafico cuando es NULL (por defecto), se construye un título, usando x $ call
dist	cuando x no tiene una disimilitud ni un componente de datos, por ejemplo,para pam (dist (*), keep.diss = FALSE), dist debe especificar la disimilitud para el clusplot.
. . .	clusplot.default (excepto el método diss) también se pueden suministrar a esta función

- **Paquete silhoutte**

El paquete silhoutte versión 2.1.0, calcule la información de la silueta según una agrupación determinada en k agrupaciones. Se establece que para cada observación i, el ancho de la silueta s (i) se define de la siguiente manera: Ponga a (i) = disimilitud promedio entre i y todos los demás puntos del conglomerado al que pertenece i (si i es la única observación en su grupo, s (i): = 0). Para todos los demás grupos C, ponga d (i; C) = disimilitud promedio de i para todas las observaciones de. El más pequeño de estos d (i; C) es b (i): = min$_C$ d (i; C), y puede verse como la disimilitud entre i y su grupo "vecino", es decir, el más cercano al que no pertenece.

- **Paquete dentroextras**

El paquete cluster:: Dendroextras versión 0.2.3, proporciona funciones adicionales para cortar, etiquetar y colorear grupos de dendrogramas.

Función colour clusters

Dendroextras proporciona colour_clusters para colorear todos los bordes que forman grupos cortados por altura o número de grupos. El argumento que se utiliza es:

colour_clusters(d, k = NULL, h = NULL, col = rainbow,

groupLabels = NULL)

color_clusters(d, k = NULL, h = NULL, col = rainbow, groupLabels = NULL)

dónde:

d	un dendograma o un objeto de árbol hclust
k	número de grupos
h	altura a la que cortar el árbol
col	función o vector de colores
groupLabels	Si es VERDADERO agregue la etiqueta de un grupo numérico, examine detalles para la opción.

Dendroextras proporciona labels y labels<-. Para obtener y establecer las etiquetas del clúster. El argumento que se utiliza para labels es:

labels(object, ...)

dónde:

object	cualquier objeto R: la función es genérica
. . .	parámetros adicionales pasados a métodos específicos

El argumento que se utiliza para labels<- es

Labels(x,…)<- value

x	objeto sobre el que se pone etiquetas
. . .	parámetros adicionales pasados a métodos específicos
value	nuevas etiquetas

- **Paquete RSKC**

  El paquete RSKC versión 2.4.2 desarrollo el algoritmo robusto de k -medias Sparse Robusto. El argumento que se utiliza en el paquete RSCK es:

  RSKC (d, ncl, alpha, L1 = 12, nstart = 200, silent=TRUE, scaling = FALSE, correlation = FALSE)

dónde:

d	una matriz numérica de datos, N por p, donde N es el número de casos y p es el número de características. Los casos se dividen en grupos ncl. Se aceptan los valores perdidos.
ncl	el número especificado de antemano de las agrupaciones.
alpha	0 <= alfa <= 1, la proporción de casos que se recortarán en K-medias robustas y dispersas. Si alfa> 0 y L1> = 1, RSKC realiza K-medias robustas y dispersas. Si alfa> 0 y L1 = NULL, RSKC realiza K-medias recortadas. Si alpha = 0 y L1> = 1 entonces realiza RSKC K-medias dispersas (con el algoritmo de Lloyd (1982)). Si alpha = 0 y L1 = NULL entonces RSKC realiza K-means (con el algoritmo de Lloyd)
L1	un solo límite L1 en pesos (los pesos de característica). Si L1 es pequeño, pocas características tendrán pesos distintos de cero. Si L1 es grande, todas las características tendrán pesos distintos de cero. Si L1 = NULL, RSKC realiza un agrupamiento no disperso (ver alfa).
start	el número de conjuntos iniciales aleatorios de centros de conglomerados en cada paso (a) que realiza K-medias o K-medias recortadas.
silent	Si es VERDAD, entonces la etapa de procesamiento no se imprime.
scaling	Si es VERDAD, RSKC resta cada entrada de la matriz de datos por la media de la columna correspondiente y la divide por la SD de la columna correspondiente
correlation	Si es TRUE, RSKC centra y escala las filas de datos antes de que se realice la agrupación. es decir, trans.d = t (escala (t (d))). La distancia euclidiana al cuadrado entre casos en el conjunto de datos transformado trans.d es proporcional a la medida de disimilitud basada en la correlación entre los casos en el conjunto de datos d

- **Paquete cmeans**

  El paquete cmeans versión 1.7-3, crea agrupaciones difusas en donde cada objeto tiene una probabilidad de pertenecer a cada grupo. El argumento que se utiliza es:

```
cmeans(x, centers, iter.max = 100, verbose = FALSE,
dist = "euclidean", method = "cmeans", m = 2,
rate.par = NULL, weights = 1, control = list())
```

dónde:

x	la matriz de datos donde las columnas corresponden a variables y las filas a observaciones
centers	número de grupos o valores iniciales para centros de grupos
iter.max	número máximo de iteraciones
verbose	Si es VERDAD, haga alguna salida durante el aprendizaje
dist	debe ser uno de los siguientes: Si "euclidean", el error cuadrático medio, si "manhattan", se calcula el error absoluto medio
method	Si "cmeans", entonces tenemos el método de agrupamiento difuso cmeans, si "ufcl" tenemos la actualización en línea
m	el grado de fuzzificación. Se define para valores superiores a 1
rate.par	el parámetro de la tasa de aprendizaje
weigts	un vector numérico con ponderaciones de los casos no negativos. Reciclado al número de observaciones en x si es necesario.
control	una lista de parámetros de control

- **Paquete kGmedian**

  El paquete kGmedian versión 1.2.5. genera agrupaciones de k-medianas basada en algoritmos de gradiente estocástico promediado recursivo. El argumento que se utiliza es:

kGmedian (X, ncenters=2, gamma=1, alpha=0.75, nstart=10, nstartkmeans=10, iter.max=20)

dónde:

X	matriz, con n observaciones (filas) en la dimensión d (columnas)
ncenters	O el número de conglomerados, digamos k, o un conjunto de centros de conglomerados iniciales (distintos). Si es un número, los centros iniciales se eligen como la salida de la kmeans función calculada con el MacQueen algoritmo
gamma	Valor de la constante que controla los pasos de descenso
alpha	Tasa de disminución de los pasos de descenso
nstart	Número de veces que se ejecuta el algoritmo, con conjuntos aleatorios de centros de inicialización elegidos entre las observaciones.
nstartkmeans	Número de puntos de inicialización en la kmeans función para elegir el punto de partida kGmedian
iter.max	Número máximo de iteraciones consideradas en la kmeans función para elegir el punto de partida de kGmedian

Paquete dbscan versión 1.1-5 una reimplementación rápida de varios algoritmos basados en densidad de la familia DBSCAN para datos espaciales. El argumento que se utiliza es:

dbscan(x, eps, minPts = 5, weights = NULL, borderPoints = TRUE, ...)

donde:

X	una matriz de datos o un objeto dist.
eps	tamaño de la vecindad épsilon.
minPts	número de puntos mínimos en la región eps (para puntos centrales).
weights	pesos para los puntos de datos
borderPoints	lógico; deben asignarse puntos fronterizos. El valor predeterminado es VERDADERO para DBSCAN normal. Si es FALSO, los puntos fronterizos se consideran ruido
object	un objeto de agrupación DBSCAN.
data	el conjunto de datos utilizado para crear el objeto de agrupación DBSCAN.
newdata	nuevo conjunto de datos para el que se debe predecir la pertenencia al clúster
…	los argumentos adicionales se pasan al algoritmo de búsqueda de vecino más cercano de radio fijo

1.14.-CONTRIBUCIÓN AL ESTUDIO DE LOS ALGORITMOS NO JERÁRQUICOS

El algoritmo **K-Medias** (Forgy, 1965) es uno de los métodos más utilizados en la investigación científica y destaca por la sencillez, velocidad de su algoritmo y su implementación en una gran variedad de software estadísticos. Sin embargo, entre sus limitaciones como ha sido señalado presenta problemas importantes:

- Convergencia a un óptimo local
- Es muy sensible a la inicialización
- Falta de robustez frente a valores atípicos

Está basado en los valores medios y, por consiguiente, son muy sensibles a los valores atípicos. Tales valores atípicos, que pueden ser comunes en muestras grandes pueden deteriorar significativamente el rendimiento de estos algoritmos, aunque sólo representen una pequeña fracción de los datos.

Con el fin de conseguir un algoritmo más robusto surge el **K-Medoids** propuesto por Vinod (1969) que utiliza medoid en vez de medias para limitar la influencia de los valores atípicos. Es un método muy similar a K-Medias en cuanto a que ambos agrupan las observaciones en n conglomerados, donde n es un valor preestablecido por el investigador. La diferencia entre ambos métodos está en que en K-Medoids cada observación está representado por una observación presente en el medoid mientras que el K-Medias cada conglomerado está representado por su centroide que se corresponde con el promedio de todas las observaciones del conglomerado. Al ser un método más robusto que K-Medias es más adecuado cuando el conjunto de datos contiene valores atípicos.

Sin embargo, sigue presentando el inconveniente de que necesita muchos recursos computacionales para grandes conjuntos de datos.

El algoritmo más empelado para aplicar K-Medoids se conoce como **PAM** propuesto por Kaufman & Rousseeuw (1990), el cual es efectivos para conjuntos de datos pequeños, pero presenta el inconveniente de que no es adecuado para conjunto de datos grandes, con lo que se propusieron para solventar esta limitación los algoritmos CLARA y CLARANS. El algoritmo **CLARA** aplica PAM a cada muestra extraída y devuelve la partición con menor error cuadrático. Aunque tiene la ventaja de ser un algoritmo óptimo para grandes conjuntos de datos presenta el inconveniente de que si la muestra está sesgada.

El algoritmo **CLARANS** es una mezcla de PAM y CLARA. Las búsquedas las realiza sobre un subconjunto del conjunto de datos con cierta aleatoriedad en cada paso de la búsqueda. Mientras CLARA tiene una muestra fija en cada etapa de la búsqueda, CLARANS forma una muestra aleatoria en cada etapa de la búsqueda y no tiene ningún requisito sobre la naturaleza de la función de distancia.

En la misma línea fue propuesto el **K-Medianas** como una variante del K-Medias. Usa como centros las medianas y no las medias siendo la mediana de un conglomerado la instancia más centrada. Por tanto, no se ve afectada por outliers y es un método robusto al ruido. Con frecuencia K-Medianas es confundida con K-Medoids, sin embargo, hay una diferencia importante en el en K-Medoids el punto central tiene que ser una de las observaciones.

La mayoría de las técnicas tradicionales como el K-Means requiere especificar el número de clusters e incluyen valores atípicos. Sin embargo, **DBSCAN** propuesto por Ester et al. (1996) es un algoritmo de agrupación en clústeres basados en densidad son muy eficaces para encontrar regiones de alta densidad y la detección de anomalías.

En 1997 fue propuesto el método **K-Medias Recortada (Trimmed K-Means)** propuesto por Cuesta-Albertos et al. (1997) muy similar al procedimiento K-Medias a diferencia de que las observaciones recortadas no se consideran en el cálculo de la función objetivo.

Zhang et al. (1999) con el fin de solventar una de las limitaciones del K-Medias propusieron el algoritmo **K-Medias Armónica** que es insensible a la inicialización de los centroides utilizando en lugar de la media aritmética, la media armónica.

Dado el avance tecnológico de los últimos años cada día se generan más datos lo que ha llevado al resurgimiento del Big Data que son conjunto de datos de gran variedad, que llegan en volúmenes cada vez mayores. En un gran número de investigaciones no solo es importante identificar posibles agrupaciones en los datos sino también analizar qué número relativamente pequeño de esas variables determinan esa partición. Para abordar estos problemas, Witten & Tibshirani (2010) propusieron una alternativa a las K-Medias llamada en este sentido **K-Medias Sparse** que simultáneamente encuentra los conglomerados y las variables de agrupamiento importantes.

En muchas situaciones en las que el interés reside en la identificación de los grupos podría ocurrir que no todas las variables contengan información sobre estos grupos. Además, la calidad de los datos (por ejemplo, los valores atípicos o faltantes) podría presentar un

problema grave y a veces difícil de evaluar en el caso de conjuntos de datos grandes y complejos por lo que una proporción de observaciones atípicas podría tener graves efectos adversos en las soluciones encontradas por el algoritmo de agrupación dispersa de Witten & Tibshirani (2010). Para solventar Kondo et al. (2012), proponen una robustificación de del algoritmo de K-Medias Sparse basado en el algoritmo de K-Medias Recortado de Cuesta-Albertos et al.(1997).

Análisis comparativo de las características computacionales de los algoritmos no jerárquicos

A continuación, en la tabla 2 se presenta una comparación desde el punto de vista computacional de los diferentes algoritmos de agrupamiento no jerárquico.

Tabla 2. Aspectos computaciones de algoritmos de agrupamiento no jerárquico

Algoritmo de Agrupamiento	Ventajas	Desventajas
K-Medias	Es fácil de implementar en grandes bases de datos	*Es muy sensible a la inicialización *Converge a un óptimo local
K-Medoids	Minimizar la suma de las distancias entre cada objeto y su correspondiente objeto representativo. Llamado medoid Sus agrupaciones no dependen del orden en que han sido introducidos 1os objetos. Es efectivo debido a que es invariable frente a los valores atípicos	Necesita que se especifique de antemano el número de clusters que se van a crear.
K-Media Difusa	La convergencia constante	El tiempo de cálculo de larga duración
K-Media Recortada	Las observaciones recortadas no se consideran en el cálculo de la función objetivo	Arbitrariedad en la selección de la zona de eliminación de datos
K-Armónica Media	Utiliza la media armónica de la distancia de cada punto a los centroides en lugar de usar el criterio de la mínima suma de cuadrados.	*Es muy sensible a la inicialización. *La función rendimiento puede quedar atrapado en diferentes mínimos locales
K-Medias Sparse	De forma simultánea encuentra los clusters y las variables importantes en la agrupación	Se ve afectado por presencia de valores atípicos
K-Medias Sparse Robusto (RSK-Means)	Robustez antes valores atípicos.	La función objetivo en cada iteración, puede quedar atrapado en diferentes mínimos locales. Por lo tanto, se inicia varias veces y se devuelve la mejor solución
K-Medias Sparse Robusto (RSK-Means)	Robustez antes valores atípicos.	La función objetivo en cada iteración, puede quedar atrapado en diferentes mínimos locales. Por lo tanto, se inicia varias veces y se devuelve la mejor solución
K-Mediana	*Utiliza como centroide al valor de la mediana, logrando algoritmos de agrupamiento más robustos *No se ve afectado por valores atípicos	Es un algoritmo sensible a la selección de los centroides iniciales
Gradiente estocástico K-Mediana	Manejar grandes bases de datos	
PAM	Es un método tipo K-Medoid que intenta determinar k particiones de n objetos trabaja satisfactoriamente para bases de datos pequeñas	El nivel de complejidad de una interacción en PAM es elevada y muy costoso en tiempo computacional para gran cantidad de valores
CLARA	El CLARA aplica el PAM en muestras del conjunto de datos, Funciona bien para conjuntos de datos grandes	La eficiencia depende del tamaño de la muestra

Continuación tabla 2:

Algoritmo de Agrupamiento	Ventajas	Desventajas
CLARANS	Es una mezcla de PAM y CLARA más eficiente. CLARANS forma una muestra aleatoria en cada etapa de la búsqueda	
DBSCAN	No es necesario determinar a priori el número de clúster deseados	*Solo son adecuados cuando los clústers no tienen una forma geométrica definida *Converge a un mínimo local

La tabla 3 se presenta una comparación desde el punto de vista algebraico de los diferentes algoritmos agrupamiento no jerárquico.

Tabla 3. Aspecto algebraico de algoritmos de agrupamiento no jerárquico

Algoritmo de agrupamiento	Función objetivo	Centroide	Métrica
K-Medias	$SSE = J = \sum_{J=1}^{k}\sum_{i=1}^{n}\left\|x_i^j - c_j\right\|^2$	Media	Distancia euclidiana
K-Medoid	$\sum_{i=1}^{n}\sum_{j=1}^{n} d(i,j)m_{ij}$	Medoid	Distancia euclidiana
K-Media Difusa	$D(X;W;O) = \sum_{J=1}^{c}\sum_{i=1}^{n}(w_{ji})^m \left\|x_i - o_j\right\|^2$	$c_j = \frac{\sum_{i=1}^{n}(w_{ji})^m x_i}{\sum_{i=1}^{n}(w_{ji})^m}$	Métrica normalizada
K-Media Recortada	$d_i = \min_{j=1,2,\ldots,k}\left\|x_i - c_j\right\|$, i=1,2,,,n	Media	Distancia euclidiana
K-Armónica	$KAM(X,C) = \sum_{i=l}^{N}\frac{k}{\sum_{j=1}^{k}\frac{1}{\left\|x_i - c_j\right\|^p}}$	Media armónica	Distancia euclidiana
K-media Sparce	$A(C) = \sum_{j=1}^{p} A_j(C)$	Media	Distancia euclidiana al cuadrado
K-medias Sparce Robusto	$\min_{C_1,\ldots,C_k}\left\{\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i'}\right\}$	Media	Distancia euclidiana al cuadrado
K-Mediana	$P(W,M_e) = \sum_{l=1}^{k}\sum_{X\in C_l} W_l' \lvert X - M_{el}\rvert$	Mediana	Distancia de manhattan
Gradiente Estocástico K-Mediana	$g(x) \overset{def}{=} E(\underbrace{min}_{r=1,\ldots,k} \emptyset(\left\|z - x^r\right\|))$	$\frac{(X_1^r + \sum_{l=1}^{n} I_r(Z_l;X_l)Z_l)}{1+\sum_{l=1}^{n} I_r(Z_l;X_l)}$	
DBSCAN	La densidad local denotada densidad (x_i) en la vecindad i-esima $x_i \in$ de un conjunto de datos X	Radio (épsilon)	Distancia euclidiana

CAPÍTULO 2: CLUSPLOT vs CLUSTER HJ-BIPLOT

2.1.-INTRODUCCIÓN

Los métodos biplots introducidos por Gabriel (1971) son una representación gráfica de datos multivariantes. Un diagrama de dispersión presenta la distribución conjunta de dos variables, un Biplot representa tres o más variables (K. Gabriel & Odoroff, 1990). Permiten representar las filas (individuos) y columnas (variables) de una matriz de datos en un subespacio de dimensión reducida. Se trata de una matriz de datos constituida por "i" individuos con "j" variables:

$$X = \begin{pmatrix} x_{11} & x_{12} \dots\dots\dots x_{1j} \dots\dots x_{1J} \\ x_{21} & x_{22} \dots\dots\dots x_{2j} \dots\dots x_{2J} \\ . & . \ . \ . \ . \ . \ . \ . \ . \ . \\ . & . \ . \ . \ . \ . \ . \ . \ . \ . \\ . & . \ . \ . \ . \ . \ . \ . \ . \ . \\ x_{i1} & x_{i2} \dots\dots\dots x_{ij} \dots\dots x_{iJ} \\ . & . \ . \ . \ . \ . \ . \ . \ . \\ . & . \ . \ . \ . \ . \ . \ . \ . \\ x_{I1} & x_{I2} \dots\dots\dots x_{Ij} \dots\dots x_{IJ} \end{pmatrix}$$

La *fundamentación teórica* de los biplots clásicos desarrollado por Gabriel (1971) se basa en la aproximación de la matriz de datos $\mathbf{X_{(IxJ)}}$, por una de menor rango q, siendo q < r, a través de la descomposición en valores singulares de **X**. Se realiza una factorización en matrices de marcadores filas y de marcadores columnas de manera tal que el producto escalar entre los marcadores aproxime "lo mejor posible" los valores de **X**.

Si consideramos los marcadores $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_I$ como filas de una matriz **A** y los marcadores $\mathbf{b}_1$, $\mathbf{b}_2$,..., $\mathbf{b}_j$ como filas de una matriz **B**, entonces podemos escribir: $\mathbf{X} \cong \mathbf{AB}'$.

Tanto los marcadores $\mathbf{a}_i$ para las filas, como los marcadores $\mathbf{b}_j$ para las columnas estarán representados en un espacio de dimensión $q \leq r$, siendo q el número de ejes retenidos y r el rango de **X**.

El método está basado en la descomposición en valores y vectores singulares de la matriz **X** de datos: **X** = **UDV**´, donde U la matriz cuyas columnas contienen los vectores propios de **XX'**, **V** la matriz cuyas columnas contienen los vectores propios de **X'X** y D es una matriz diagonal que contiene a los valores propios de **X**.

Debe cumplirse que **U'U**=**V'V**=**I**, es decir, las columnas de **U** y **V** son ortonormales, esta propiedad asegura la unicidad de la factorización.

Siendo:

$$\mathbf{X} = \mathbf{AB'} = \mathbf{UDV'}$$

La selección de distintas factorizaciones conduce a distintos tipos de marcadores y por lo tanto a distintos tipos de Biplots.

En el caso de la *métrica identidad*, se pueden elegir los marcadores de acuerdo a los distintos valores de **y** en la siguiente descomposición:

$$\mathbf{A}=\mathbf{UD} \quad \text{y} \quad \mathbf{B}=\mathbf{VD}^{1-y}$$

Dependiendo del valor seleccionado para y (y = 0, 1) tenemos los Biplots Clásicos de Gabriel (1971): GH-Biplot y JK-Biplot.

Cuando en la expresión $\mathbf{A}=\mathbf{UD}^{y}$ $\mathbf{B}=\mathbf{VD}^{1-y}$ y = 0 se trata del GH-Biplot, luego tendremos lo siguiente:

$$\mathbf{X}=\mathbf{A\,B'} = (\mathbf{U}\,\mathbf{D}^{y})\,(\mathbf{D}^{1-y}\,\mathbf{V'}) = (\mathbf{U})\,(\mathbf{D\,V'})$$

La factorización en el GH-BIPLOT corresponde a la elección de los marcadores tal que:

$$\mathbf{A = U\ B = V\ D}$$

Como trabajamos con la métrica identidad en el espacio de las filas, **A** debe verificar que: **A' A=I** (por lo que la representación es única, excepto por rotaciones).

Luego:

$$\mathbf{X'\ X = (AB')'\ AB' = BA'AB' = BB'}.$$

Es decir:

$$\mathbf{X'\ X = BB'}$$

Por lo tanto, este BIPLOT preserva la métrica euclídea usual entre las columnas, pero no entre las filas, obteniéndose así para las columnas, una alta calidad de representación.
El nombre de este BIPLOT (**GH-BIPLOT**) se debe a que GABRIEL adoptó la notación **G** para los marcadores fila y **H** para los marcadores columna.

$$\mathbf{G = U} \qquad \mathbf{H = V\ D}$$

Al introducir un factor de escala, la matriz [(1/n-1) **X'X**] coincide exactamente con la matriz de covarianzas, y si se designan los marcadores como:

$$\mathbf{A} = \sqrt{n-1}\ \mathbf{U} \quad \mathbf{B} = (1/\sqrt{n-1})(\mathbf{V\ D})$$

Así obtenemos el biplot denominado Biplot de Componentes Principales, donde los productos escalares entre los marcadores columna reproducen la estructura de las covarianzas entre variables.

Si y = 1 en la expresión $\mathbf{A}=\mathbf{U}\mathbf{D}^{y}$ $\mathbf{B}=\mathbf{V}\mathbf{D}^{1-y}$ se define el JK-BIPLOT, luego tendremos lo siguiente:

$$\mathbf{X}= \mathbf{A}\,\mathbf{B}' = (\mathbf{U}\,\mathbf{D})\,(\mathbf{V}')$$

La elección de los marcadores es:

$$\mathbf{A} = \mathbf{U}\,\mathbf{D} \quad \mathbf{B} = \mathbf{V}$$

En este tipo de Biplot se impone la métrica $\mathbf{B}'\mathbf{B} = \mathbf{I}$ en el espacio de las filas de la matriz $\mathbf{X}_{(IxJ)}$.

Considerando el tipo de factorización y la métrica tenemos:

$$\mathbf{XX'}= \mathbf{AB'(AB')'}= \mathbf{AB'BA'}= \mathbf{AA'}$$

Luego:

$$\mathbf{X}\,\mathbf{X}' = \mathbf{A}\,\mathbf{A}'$$

Este Biplot preserva la métrica euclídea usual entre las filas, pero no entre las columnas, obteniéndose alta calidad de representación para las filas. A este biplot Gabriel lo denominó JK-biplot porque utilizó **J** para denotar la matriz de marcadores fila y **K** para la matriz de marcadores columna.

$$\mathbf{J=UD} \qquad \mathbf{K=V}$$

Gabriel describió esencialmente dos tipos de biplots: biplot CMP (preservación métrica de columna), que conduce a una alta calidad para las variables, y biplot RMP (preservación métrica de filas) que conduce a una alta calidad para las filas.

El HJ-Biplot (Galindo, 1986) es una técnica de representación simétrica y simultánea similar de algún modo al análisis de correspondencia propuesto por Benzécri (1973) pero no restringida a datos de frecuencia. Este método permite una representación simultánea de filas (individuos) y marcadores de columnas (variables) y uno puede estudiar la relación entre marcadores de columnas y grupos en dimensión reducida.

El HJ-Biplot (Galindo, 1986) es una extensión de los biplots clásicos introducidos por Gabriel (1971). Es un método de análisis de datos exploratorio que busca patrones ocultos en la matriz de datos. Es, por tanto, una técnica de representación de datos que consiste en visualizar una matriz de datos multivariantes X_{nxp} usando vectores como puntos llamados marcadores g_1, g_2, …, g_n para cada fila y vectores llamados marcadores h_1, h_2, …, h_p para cada columna. Cada fila representa un sujeto y cada columna una variable, de modo que ambos conjuntos de marcadores se pueden superponer en el mismo sistema de referencia con la máxima calidad de representación.

Si las filas de una matriz A se describen como marcadores g_1, g_2, ..., g_n y matriz B como marcadores h_1 , h_2,…, h_p se obtiene $X = AB^T$.

Los marcadores se obtienen de la descomposición de valores singulares (SVD) de la matriz de datos. La descomposición de valores singulares (SVD) de la matriz X se define por $X = UDV^T$, donde U es la matriz cuyas columnas son los vectores propios de XX^T, V es la matriz cuyas columnas son los vectores propios de X^TX y D es la diagonal de la matriz de valores singulares λ_i de X. A y B son las matrices de las dos primeras columnas de UD y VD respectivamente.

Tomando como base la información que arroja la representación geométrica multidimensional HJ-Biplot (Galindo, 1986) de los diferentes clusters posibles, es viable elegir aquellos que sean conceptualmente interpretables. El criterio propuesto por Vicente-Tavera (1992) es una extensión del "criterio de inercia o de varianza" propuesto por Benzécri (1973), que se basa en la representación del HJ-Biplot.

A través del método de inercia basado en una representación HJ-Biplot presentado por Vicente-Tavera (1992) es posible interpretar los diferentes grupos en función de las variables que más los afectan logrando aglomeración jerárquica y gráficos espaciales en los diferentes planos factoriales del HJ-Biplot, lo que permite la interpretación de los diferentes grupos de acuerdo con la especificación de sus variables más importantes. Esta característica de la técnica no se encuentra en los métodos clásicos de aglomeración jerárquica de agrupamiento. Cada plano factorial representa las muestras y variables en función de dos ejes dados y ofrece una perspectiva diferente de las asociaciones obtenidas.

2.2.- CLASIFICACIÓN JERÁRQUICA

El término Clasificación sirve para designar un sistema de clases jerárquico. La representación gráfica correspondiente a una clasificación jerárquica se lleva a cabo, generalmente, mediante un diagrama en forma de árbol de tal manera que en la base del árbol aparezcan las n unidades taxonómicas a clasificar: de ellas parten ramas que se unen en un nudo (de cada nudo solamente salen dos ramas), estos nudos a su vez se unen con otros nudos y así en forma sucesiva llegamos al tronco general que engloba a todas las ramas.

El árbol resume el proceso de clasificación. Los individuos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de semejanza o diferencia entre los individuos.

En todo árbol jerárquico se pueden establecer dos tipos de lecturas: descendiente y ascendente.

Seguir en el árbol un camino descendente es considerar una secuencia de nudos o individuos: n_1, n_2, ..., tal que $n_{(p+1)}$ sea el descendiente inmediato del n_p.

El algoritmo descendente exige conocer perfectamente las variables mientras que el algoritmo ascendente parte de un conjunto de elementos individuales sobre los cuales en principio no es necesario tener ningún tipo de información a priori y mediante un proceso matemático objetivo se establece la clasificación ascendente.

El algoritmo ascendente es el algoritmo más conocido debido a su sencillez en la implementación a su bajo costo computacional, y está implementado en casi todos los paquetes comerciales de análisis multivariado de datos.

La complejidad del algoritmo descendente es de al menos 2^{n-1}, si n es el nnúmero de individuos, por motivos computacionales son menos usados que los algoritmos ascendentes.

El interés que pueda tener una partición viene en función de las cualidades que tengan las clases que definen:

- Clases compactas: si los elementos que forman cada clase son coherentes y bien caracterizados. La partición idónea sería aquella donde las clases fuesen de un sólo elemento ya que forman un todo coherente y están perfectamente caracterizadas.
- Clases separadas: si las clases de la partición se diferencia claramente las unas de las otras.

2.2.1.-ANÁLISIS DE LAS COMPONENTES PRINCIPALES Y COORDENADAS PRINCIPALES EN CLUSTERS AGLOMERATIVOS

El análisis de componentes principales (PCA) realizado sobre una matriz de datos n × p produce variables aleatorias transformadas linealmente que tienen propiedades especiales en términos de varianzas. En efecto, transformar la variable del vector original en el vector de componentes principales equivale a una rotación de ejes de coordenadas a un nuevo sistema de coordenadas. Los componentes principales resultan ser los vectores característicos de la matriz de covarianza. Por tanto, se puede considerar que el estudio de los componentes principales pone en términos estadísticos los desarrollos habituales de los valores propios y los vectores propios para matrices semidefinidas positivas.

Sin embargo, un PCA de una matriz de covarianza no es apropiado cuando las variables no se miden en unidades comparables ya que los componentes principales no son invariantes de escala, mientras que un PCA de la matriz de correlación podría ser suficientemente significativo, ya que la matriz de correlación de la muestra es invariante bajo cambios de escala. En este caso se pueden utilizar diferentes formas de estandarización (Milligan & Cooper, 1988; Schaffer & Green, 1996).

El uso del PCA antes de aplicar algoritmos de agrupamiento no es un método inadecuado para la recuperación de verdaderos clústeres si se considera la pérdida de información al reducir grandes dimensiones a pequeñas dimensiones. Sin embargo, la tendencia de las capacidades de recuperación de los algoritmos de agrupamiento es diferente dependiendo de las estrategias de estandarización y la configuración estructural de los parámetros y, diseñado en X.

El análisis de coordenadas principales (POA) implica proyectar los puntos en un espacio definido por un pequeño número de ejes principales. Las distancias entre el i-ésimo y el j-ésimo objeto podrían aproximarse usando ejes principales. Cuando todas las distancias entre los objetos i-ésimo y j-ésimo de n muestras son conocidas, sus coordenadas, denominadas ejes principales, se encuentran mediante un conjunto de condiciones necesarias y suficientes para que exista una solución en el espacio euclidiano real.

El PCA y POA se definen como **duales** entre sí cuando ambos conducen a un conjunto de n puntos con las mismas distancias entre objetos. Aunque POA no es un método asociado para incluir información sobre las variables (Gower & Harding, 1988), las coordenadas principales pueden usarse para asegurar la identificación de objetos.

Consideramos la matriz XX' para un POA, donde X es la matriz de datos estandarizada n × p, mientras que para un PCA la matriz se requiere la matriz $X'X$. Supongamos que la matriz $X'X$ tiene un valor propio λ y el vector propio correspondiente w. Así obtenemos:

$$X'Xw = \lambda w$$

Si la matriz XX' tiene un valor propio α y el vector propio correspondiente u, entonces:

$$XX'u = \alpha u$$

De las relaciones anteriores obtenemos:

$$XX' (Xw) = \lambda(Xw),$$

de modo que:

$$\lambda=\alpha \text{ y } Xw = ku$$

donde k es una constante que relaciona la escala de los dos conjuntos de autovectores. Los cosenos de dirección para las componentes principales se normalizan como $w'w = 1$, por tanto:

$$k^2u'u = w'X'Xw = \lambda w'w= \lambda$$

Si los autovectores w están normalizados, entonces k = 1 y u=Xw.

Los elementos b_{ij} de la matriz XX' n × n vienen dados por:

$$b_{ii} = \sum_{k=1}^{p} x_{ik}^2 \quad y \quad b_{ij} = \sum_{k=1}^{p} x_{ik}\, x_{jk}$$

y las distancias entre coordenadas entre los objetos i y j se presentan mediante:

$$b_{ii} + b_{jj} - 2b_{ij} = \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2 = d_{ij}^2$$

Esto está asociado con la matriz simétrica Fn × n con elementos, d_{ij}, i = 1, 2, ..., n − 1, y j = i + 1, ..., n, que son la distancia euclidiana entre i-esimo y j-esimo objetos. Si la matriz Fn × n es semidefinida positiva, el POA que opera en XX´ o F es un dual del PCA en X´X. Por tanto, es posible calcular las coordenadas principales de cualquier matriz de distancia euclidiana sin estar en posesión de la matriz de datos original o de una matriz de varianza covarianza de los datos.

Seong y William (2006) concluyeron que el análisis de coordenadas principales es una técnica más poderosa que el análisis de componentes principales para asegurar la identificación en grupos de objetos.

2.3.-CRITERIO DE LA INERCIA

La utilización de métodos de representación simultánea de datos multidimensionales se ha visto reducida al Análisis de Correspondencias de Benzécri (1973), técnica íntimamente relacionada con el análisis de Componentes Principales, que permite visualizar la posible relación entre un par de variables categóricas, y entre sus respectivas categorías, pero pensada para trabajar con matrices de frecuencias.

Para obtener clases lo más homogéneas posibles y tal que estén suficientemente separadas, se introducen las nociones de **Inercia Dentro de las clases** y otra **Inercia Entre Las Clases**.

La idea sobre la que se fundamenta este método consiste en ir descomponiendo la inercia total del sistema en cada partición, en dos tipos de inercia, una denominada **Inercia Dentro de las Clases** y otra **Inercia Entre las Clases** que forman la partición.

Sea esta partición P= $\{Q_1,\ldots,Q_K\}$, de un conjunto de datos donde los $c_1,\ldots,c_K$ son los centros de gravedad de las clases:

$$c_j = \frac{1}{n}\sum_{i\in Q_j} x_i$$

c es el centro de gravedad total:

$$c = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Se denomina $I_D(P) = \frac{i}{n}\sum_{j=1}^{K}\sum_{i\in Q_j}\left\|x_i - c_j\right\|^2$ la **Inercia Dentro $I_D(P)$** de una partición **P** como la suma de las inercias internas de las clases que componen dicha partición.

Se denomina $I_E(P) = \sum_{j=1}^{K}\frac{|Q_j|}{n}\left\|c_j - c\right\|^2$ **Inercia Entre $I_{(E)}(P)$** las clases de la partición **P** a la suma de la inercia de estos K puntos (centro de gravedad de las clases).

El paso de una partición $\mathbf{P_k}$ a otra $\mathbf{P_{k+1}}$ se hace mediante la agregación de las dos clases que menor Inercia Entre tengan.

La suma de la "inercia entre" clases para una partición más la inercia dentro de las clases es constante sea cual sea la partición puesto que esa suma es siempre igual a la inercia

total de la nube. Es sólo el reparto entre la inercia dentro y entre clases la que varía con la partición.

Una clase es tanto más compacta cuanto más baja sea la inercia dentro de la clase con respecto a su centro de gravedad.

De entre dos particiones que conlleven el mismo número de clases preferíamos aquella que sea más compacta.

Dada una partición, al reagrupar dos clases en una sola, la "inercia dentro" de la misma se modifica. La inercia de la nueva clase con respecto a su centro de gravedad es igual a la suma de la inercia entre y de la inercia dentro de las clases que se han agrupado; por tanto, agrupar dos clases en una sola aumenta la inercia dentro. Por esta razón, es por lo que se toma como criterio agregar primero aquellas clases que proporcionen un aumento mínimo.

Este criterio tiene la ventaja de propiciar particiones que tienen clases homogéneas y bien separadas unas de otras:

-Homogéneas puesto que la agregación se hace de manera que la inercia alrededor del centro de la clase sea lo más pequeña posible y esto es así para todas las clases de la partición, por tanto, la inercia dentro de la partición es baja.

-Bien separadas puesto que las nubes de los centros de gravedad de las clases de una partición están dispersa ya que la inercia entre es grande (Vicente-Tavera, 1992).

2.4.- LA TÉCNICA DE CLUSTERING BASADA EN EL HJ-BIPLOT

El HJ-Biplot es una representación gráfica multivariante basado en la Descomposición de Valores Singulares (SVD) de la matriz X, mediante marcadores para las filas (j_1, …j_n) y para las columnas (h_1, …., h_p) supuestos en el mismo sistema de referencia y con la máxima calidad de representación, de forma que nos permite interpretar las filas, las columnas y las relaciones entre ambos. El método parte de una matriz X cuyos elementos X_{ij} representan los valores que toman j variables sobre i individuos (unidades taxonómicas).

$$X = U_{nxl} \sum_{l} V_{lxp}^{T}$$

U_{nxl}: representa la matriz que contiene los vectores propios de XX´

V_{lxp}: representa la matriz que contiene los vectores propios de X´X

$\sum_l$: matriz diagonal de los valores singulares de X (λ_1, …, λ_l)

El objetivo del HJ-Biplot no es reproducir exactamente los valores de la matriz original X, sino representar, en un mismo sistema de referencia, las filas y las columnas con la misma calidad de representación.

Clustering HJ-Biplot

El análisis de agrupamiento basado en el HJ-Biplot es una extensión del criterio de la varianza propuesto por Benzécri (1973), el cual está basado en la representación de un HJ-Biplot en lugar de un análisis de correspondencia (Vicente-Tavera, 1992; Vicente-Tavera, Molina-Ballesteros, Vicente, & Garcia-Talegon, 1999).

Utiliza el criterio de agrupamiento jerárquico aglomerativo, utilizando la distancia euclídea como medida de distancia. El algoritmo jerárquico aglomerativo parte de tantos clusters como datos tiene la muestra y en cada paso se van juntando siguiendo algún criterio especificado hasta obtener un único cluster con todos los datos.

El algoritmo se desarrolla de la siguiente manera:

A.-En primer lugar, se realiza un HJ-Biplot, y se guardan las coordenadas mediante la descomposición en valores y vectores singulares de la matriz. Se preserva la métrica de las filas y de las columnas, consiguiéndose así la misma bondad de ajuste para ambas. De esta manera los individuos quedan representados por puntos, en un espacio multidimensional, donde el número de dimensiones estará determinado por los factores retenidos en el análisis.

B.-A continuación, se construye la matriz de distancias al cuadrado entre individuos, agrupándose los individuos entre los cuales existe una distancia muy pequeña, mediante la **Inercia Interna Dentro** de una partición **($I_D(P)$).** El algoritmo produce una partición en **q** conglomerados o clases del conjunto de individuos.

La **$I_D(P)$** se define por la suma de las inercias internas de las C clases que componen dicha partición (**P**), es decir, se calculan las coordenadas del centro de gravedad de la clase C como la media de las coordenadas de los objetos que la componen.

Sabiendo que:

P^q: Partición del conjunto de datos a clasificar, formada por q conglomerados

$F_{\propto(S_i)}$ son los $\propto -ésimas$ coordenadas de cada elemento que componen el conglomerado

g(C): coordenadas del centro de gravedad de la clase C

q: número de factores retenidos en el HJ-Biplot

O: centro de gravedad donde se han retenido q factores

La inercia interna de una clase C formado por n_i objetos vendría definido por:

$$F_{\alpha([g(C)]} = \frac{F_{\propto(S_1)} + F_{\propto(S_2)+\cdots+F_{\propto(S_i)}} + \cdots + F_{\propto(S_q)}}{n_q}$$

$$d^2(i_1, i_2) = \sum_{\alpha=1}^{q} \left[F_{\alpha(S_{i_1})} - F_{\alpha}(O_{i_1,i_2})\right]^2 + \sum_{\alpha=1}^{q} \left[F_{\alpha(S_{i_2})} - F_{\alpha}(O_{i_1,i_2})\right]^2$$

$$d^2(S_t, S_V) = I_E(\{S_t, S_v\})$$

C.- Se calcula la distancia o separación entre las clases que componen la partición **P** mediante la **Inercia Entre** definida como la suma de la inercia de los Q puntos (centro de gravedad de las clases) siendo grande en el caso de que la nube de los **Q** puntos esté dispersa entorno al centro de gravedad del conjunto total.

D.-Se repiten los pasos B y C hasta que todos los individuos son agrupados.

2.5.- ESTUDIO COMPARATIVO CLUSPLOT vs ANÁLISIS DE CLUSTER SOBRE LAS COORDENADAS DEL HJ-BIPLOT

En un conjunto de datos bivariantes es fácil representar los clusters, por ejemplo, rodeándolos manualmente o separándolos por líneas. Pero muchos conjuntos de datos tienen más de dos variables, o vienen en forma de disimilitudes entre objetos. Existen métodos para dividir un conjunto de datos de este tipo en grupos, pero la división resultante no es visual por sí misma. De aquí la construcción de una nueva visualización gráfica llamada Clusplot.

Clusplot es una representación grafica de los clusters y sus objetos particularmente de una salida de un método de partición, en el cual los objetos son representados como puntos en un gráfico bidimensional y los clusters como elipses de varios tamaños y formas. La función Clusplot desarrollada por Pison et al. (1999) proporciona gráficos que pueden tener aspectos diferentes dependiendo del algoritmo que se utilice.

El algoritmo Clusplot está implementado en el software S-PLUS (Pison et al., 1999) y en el software R. Crea un gráfico bivariante que visualiza una partición (clustering) de los datos. Todas las observaciones están representadas por puntos en el gráfico, utilizando componentes principales o escalado multidimensional. Alrededor de cada cluster se dibuja una elipse. Dentro del procedimiento es importante resaltar que la matriz de distancias se obtiene a partir de la función DAISY (Dissimilarity Matrix Calculation) y PAM para adaptar la matriz inicial de los datos.

La librería "cluster" de R tiene una función daisy que calcula la matriz distancia de una matriz de datos usando las distancias euclideana y manhattan y considerando además distintos tipos de variables mediante el uso del coeficiente de Gower. En el algoritmo PAM los datos deben ser de tipo numérico o matrices de disimilaridades. Si la matriz original contiene datos de tipo mixto se aplica el algoritmo DAISY para obtener la matriz de disimilaridades. Consta de dos fases:

1. Fase BUILT: se eligen los k medoides iniciales, siendo la observación cuya suma de disimilarides a las otras observaciones es la más pequeña de todas, el primer medoide elegido. Los restantes mediodes son elegidos de tal manera que la ganancia total sea la más alta. Si los medoides son elegidos al azar esta fase puede ser suprimida.

2. Fase SWAP: intercambia los medoides iniciales por otras observaciones que no han sido elegidas, tratando de mejorar el rendimiento del algoritmo.

El Clusplot para obtener las agrupaciones de las observaciones utiliza el algoritmo PAM o CLARA cuando el número de datos es muy grande.

En resumen, la ejecución en el programa R sería la siguiente:

1.-Se carga en el software la matriz de datos original.

2.-Se ejecuta el algoritmo DAISY, lo que permite el cálculo de la matriz de disimilaridades.

3.-Se lleva a cabo el algoritmo PAM o CLARA indicando el número de cluster elegidos.

4.-Se realiza el algoritmo Clusplot para visualizar la representación gráfica de los puntos de la matriz.

El método de cluster a partir de las coordenadas del HJ-Biplot consta de los siguientes pasos:

1.- Se realiza un HJ-Biplot y se obtienen así las coordenadas mediante el método de descomposición en valores y vectores singulares.

2.-Se normalizan las variables y se obtiene la matriz de distancias al cuadrado entre individuos en base a los criterios de inercia.

La agrupación el método la realiza a partir del criterio jerárquico aglomerativo.

El desarrollo del algoritmo y visualización del mismo se puede llevar a cabo con el software MULTBiplot (Vicente-Villardón, 2017).

En la tabla 4 se presenta un estudio comparativo de ambos métodos.

Tabla 4. Estudio Comparativo Clusplot vs Clustering HJ-Biplot

Similitudes	Diferencias	
El investigador elige el número de cluster		**Tipo de variables:**
	ClusPlot:	Mixto
	Clustering HJ-Biplot:	Numérica
Representación 2D de la solución de clusters		**Software:**
	ClusPlot:	R-projec
	Clustering HJ-Biplot:	MultBiplot y R-project

2.6.- APLICACIÓN DEL CLUSPLOT Y CLUSTERING HJ-BIPLOT A UN CONJUNTO DE DATOS REALES

2.6.1.-INTRODUCCIÓN

La calidad del agua es un tema delicado de interés mundial que se define por una serie de características físicas, químicas y biológicas. La complejidad del estudio de la calidad del agua requiere encontrar modelos simples para identificar las variables que más influyen en ella. Por lo tanto, el uso de técnicas de análisis multivariantes será de inmensa ayuda para encontrar relaciones y conclusiones que nos ayuden a determinar el estado de la calidad del agua a través de indicadores biológicos, físicos y químicos.

La Cuenca del Canal de Panamá es uno de los sistemas hidrológico-hidráulicos más importantes de la República de Panamá y del mundo. La cantidad y la calidad del agua dependen en gran medida del estado de los bosques y del hábitat natural. Por lo tanto, su manejo requiere el conocimiento de la "calidad natural del agua" y la dinámica del agua

a través del ciclo hidrológico, que definen las características que la hacen apropiada o no para su uso (ACP, 2010b).

Dada la dificultad de analizar la gran cantidad de datos disponibles, el propósito es sintetizar dicha información, lo que equivale a reducir la cantidad de estos datos, con la mínima pérdida de información posible. Un modo de conseguir este objetivo consiste en analizar la estructura inicial de la nube de puntos del hiperespacio mediante una configuración simplificada en un espacio de menor dimensión.

Los objetivos de esta investigación son:

-Conocer desde un punto de vista multivariante las posibles relaciones existentes entre los distintos parámetros de calidad del agua establecidas según la Normativa de la Autoridad del Canal de Panamá.

-Estudiar las diferencias de los diferentes puntos de muestreo donde se realizaron las colectas de las muestras de acuerdo a la calidad del agua.

-Encontrar las agrupaciones de las variables físico-químicas y biológicas.

2.6.2.-METODOLOGÍA

Se realizó un estudio de la calidad del agua en el embalse de Gatún en las áreas de Gamboa y Paraíso en los meses de febrero a diciembre del año 2009. Se contó con ciento veinte observaciones como resultado de tres muestras (puntos de muestreo) colectadas en Paraíso y Gamboa dos veces al mes durante febrero (feb), marzo (mar), abril (abr), mayo

(may), julio (jul), agosto (ago), septiembre (sep), octubre (oct), noviembre (nov) y diciembre (dic) del año 2009.

En esta investigación se evaluaron las siguientes variables: temperatura, pH, transparencia, turbidez, nitratos, ortofosfatos, fósforo, nitrógeno total, clorofila a, radiación solar, oxígeno disuelto y microcistinas. Con ellas se determinó la calidad del agua en los lugares de muestreo, Gamboa y Paraíso.

La inspección de matrices de datos multivariantes se ha llevado utilizando el HJ-Biplot (Galindo, 1986), estandarizando por columnas (variables). El método nos permite una representación conjunta de las variables físico-químicas y biológicas (columnas) y de los meses de muestreo (filas) en el mismo sistema de coordenadas con una alta calidad de representación tanto para los puntos de muestreo como para las variables fisicoquímicas y biológicas.

El análisis de cluster se realizó a partir de las coordenadas que se obtuvieron del HJ-biplot (método K-means). El análisis nos permitió identificar las variables físico-químicas y biológicas que incidieron para las agrupaciones entre los distintos puntos de muestreo. Para llevar a cabo el análisis se utilizó el software Multbiplot (Vicente-Villardón, 2017).

La metodología utilizada para realizar el clustering HJ-Biplot se describe en la figura 6.

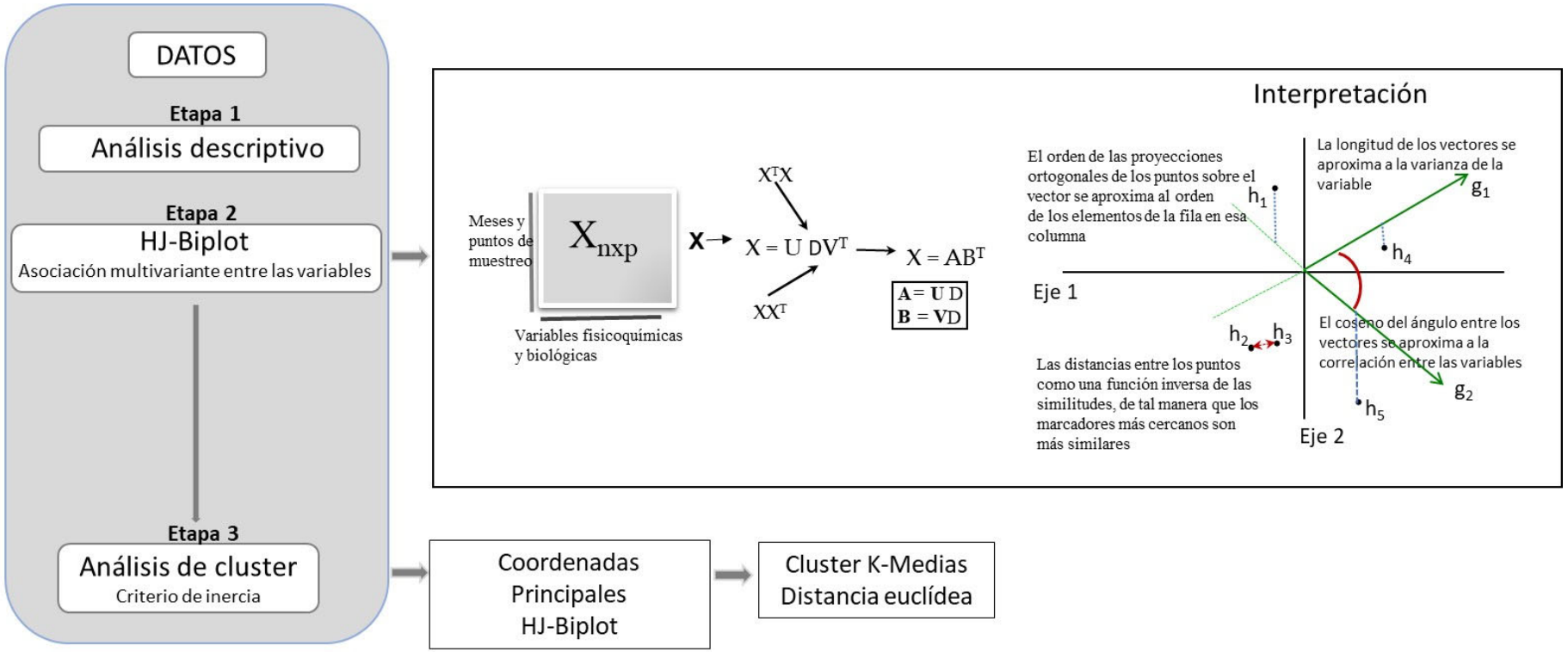

Figura 6. Diagrama de flujo del análisis estadístico

2.6.3.-RESULTADOS

Se retuvieron dos ejes, con ello se consiguió una inercia acumulada del 48,23%, siendo la tasa de inercia en el primer eje del 33,85%, suficiente para caracterizar las variables físico-químicas y biológicas en las áreas de Gamboa y Paraíso.

El eje 1, quedó determinado por las variables pH, transparencia, turbiedad, nitrato, ortofosfato, clorofila a y oxígeno, y el eje 2 por las variables radiación solar y toxina.

Se observó una relación fuerte y directa entre el oxígeno, transparencia y clorofila a, siendo esta última variable la más importante para discriminar entre los puntos de muestreo en el eje 1. Por otro lado, se dio una correlación positiva entre las variables nitrato, turbiedad y ortofosfatos.

Se formaron dos clusters con los distintos puntos de muestreo. En la representación gráfica de los mismos pueden observarse los clusters identificados con las líneas Convex-Hulls (figura 7). El análisis nos permitió identificar las variables físico-químicas y biológicas que incidieron para las agrupaciones entre los distintos puntos de muestreo.

El cluster 1 está caracterizado por la presencia de las siguientes variables: pH, transparencia, clorofila a, oxígeno y temperatura y por los puntos de muestreo de los meses de febrero, marzo, abril y mayo de Gamboa y Paraíso correspondientes a los meses de la estación seca.

-El cluster 2 comprende las variables: nitrato, ortofosfatos, turbidez y P-total. Todos ellos son parámetros que sufren variaciones en la estación lluviosa y que, a su vez, pueden influir en la presencia de ciabobacterias con potencial toxigénico.

La agrupación de los meses de muestreo y las variables nos permite comprender la naturaleza compleja de los problemas de calidad del agua, en el tiempo y el espacio, y simultáneamente interpretar el comportamiento del embalse en un momento dado. Por tanto, estos métodos son útiles para la evaluación de la calidad del agua en un medio acuático y para detectar patrones multivariantes en un grupo complejo de datos con diferentes variables. Con base en la información obtenida es posible comprender mejor los complejos problemas de la calidad y el monitoreo del agua en la región en espacio y tiempo. Además, este enfoque multivariante permite una clasificación confiable de los puntos de muestreo del área de estudio, así como una mejor estrategia para el monitoreo futuro. Finalmente, se espera que esta investigación y enfoque sea útil para mejorar el conocimiento sobre el comportamiento hidrodinámico general del lago Gatún.

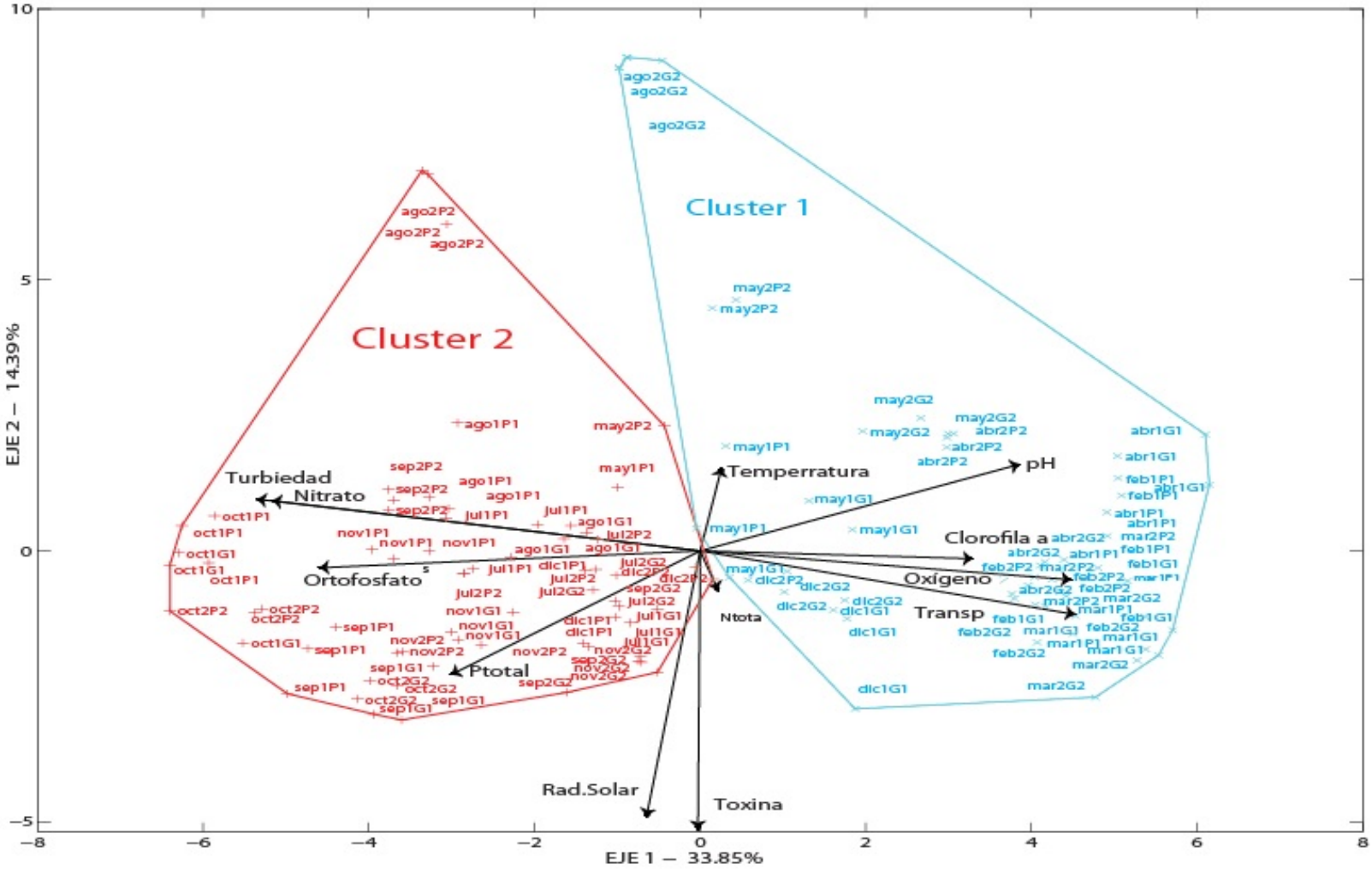

Figura 7. Representación factorial del HJ-Biplot por clúster, plano 1-2

Este trabajo fue publicado en la revista Journal of Hidrology con el título: "**Water quality evaluation through a multivariate statistical HJ-Biplot"** (Carrasco et al., 2019).

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Water quality evaluation through a multivariate statistical HJ-Biplot approach

Gonzalo Carrasco[a], Jose-Luis Molina[b,*], María-Carmen Patino-Alonso[c], Marisela Del C. Castillo[d], María-Purificación Vicente-Galindo[c], María-Purificación Galindo-Villardón[c]

[a] Department of Statistics, Faculty of Natural Sciences, Exact and Technology, University of Panama, Panama
[b] High Polytechnic School of Engineering, Area of Hydraulic Engineering, University of Salamanca, Av. de los Hornos Caleros, 50, 05003 Ávila, Spain
[c] Department of Statistics, University of Salamanca, Campus Miguel de Unamuno, C/Alfonso X El Sabio s/n, 37007 Salamanca, Spain
[d] Faculty of Natural Sciences, Exact and Technology, University of Panama, Panama

ARTICLE INFO

This manuscript was handled by G. Syme, Editor-in-Chief

Keywords:
Cluster
HJ-biplot
Lakes
Multivariate statistical analysis
Water quality

ABSTRACT

Water quality is a sensitive topic of worldwide concern that is defined by a series of physical, chemical and biological characteristics. The complex nature for studying water quality requires finding simple models to identify the variables that influence it the most. Therefore, the use of multivariate analysis techniques will be of immense help to find relationships and conclusions which aid us to determine the state of water quality through biological, physical and chemical indicators. In this investigation, the following variables were evaluated: Temperature, pH, transparency, turbidity, nitrates, orthophosphates, phosphorus, total nitrogen, chlorophyll a, solar radiation, dissolved oxygen and microcystins. These determined water quality at the sample sites, Gamboa and Paraiso. The results obtained through the application of a two-way multivariate analysis method called HJ-Biplot reflect variables relationships of chemical, physical and biological compositions. Furthermore, results conform two clusters of sample points that satisfactorily match to the region seasons. Cluster 1 is characterized by the presence of the following variables: pH, transparency, chlorophyll a, oxygen and temperature. On the other hand, cluster 2 comprises the following variables: nitrate, orthophosphates, turbidity and P-total. They are all parameters that suffer variations in the rainy season, and that, in turn, can influence the presence of cianobacteria with toxigenic potential. Furthermore, this study demonstrates that the multivariate statistical methods are valuable for interpreting complex data sets, specifically, for water quality monitoring network.

Figura 8. Publicación Journal of Hydrology

Hasta el momento, esta investigación comprende el único estudio sobre la calidad del agua en la Cuenca del Canal de Panamá donde se aplicó el HJ-Biplot y Análisis de Cluster que ha permitido evaluar los procesos fisicoquímicos y biológicos relacionados con los cambios estacionales (secos y lluviosos) en Gamboa y Paraíso del embalse de Gatún durante el período analizado.

CAPÍTULO 3: ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN

3.1.-INTRODUCCIÓN

En la investigación científica un problema común es clasificar a los individuos en grupos homogéneos a partir de la información de ciertas variables que pueden ser cuantitativas o categóricas. El objetivo del estudio no depende de si las variables son de una naturaleza u otra, sin embargo, las técnicas estadísticas son distintas en uno y otro caso.

Como se ha visto en los capítulos anteriores cuando todas las variables son descritas al mismo nivel para clasificar los individuos en grupos homogéneos se lleva a cabo un Análisis de Clusters. Sin embargo, si el objetivo es formar grupos de individuos homogéneos pero que sean distintos entre ellos según ciertos predictores con respecto a una variable respuesta, la técnica estadística adecuada es el Análisis de Segmentación.

Los métodos de segmentación son muy requeridos en diversas áreas, como por ejemplo la Entomología, Demografía, Sociología, Marketing o la Medicina. Por ejemplo, en el Marketing cuando una compañía va a incorporar al mercado, un nuevo producto, debe considerar en su planeación, a que perfil de posibles consumidores va a dirigir su propuesta. Es necesario que desarrolle un adecuado, modelo de segmentación, que permita dividir una población en segmentos o partes, que difieren con respecto a un criterio diseñado.

Entre los beneficios de **segmentar** una población o muestra podemos mencionar los siguientes:

a) Permite construir un perfil más preciso de los individuos, que componen un colectivo bajo estudio.

b) Obtener mejores pronósticos, sobre el comportamiento de grandes grupos de datos.

c) Agrupar para conocer mejor un subgrupo poblacional.

El Análisis de Segmentación debe ser utilizado con una finalidad exploratoria ya que es una técnica que a partir de la información que suministran ciertas variables independientes o predictoras, clasifica un conjunto de objetos en grupos, capaces de describir de la mejor manera posible la variable respuesta. Es una técnica estadística que trabaja sobre datos tipo regresión, donde uno de los objetivos primordiales es el de explicar o predecir, la(s) variable(es) respuesta(s) a partir del conjunto de variables independientes. Tienen como característica el que su aplicación está regularmente libre de los supuestos requeridos por los métodos basados en el modelo lineal.

La realización de un estudio de segmentación exige seleccionar la *variable dependiente* (cuantitativa o categórica), recoger información sobre las *variables independientes o predictores* (cuantitativas o categóricas), seleccionar la técnica de segmentación y, por último, interpretar los datos o resultados y proponer una estrategia a seguir.

Kottler (1988) describe las fases de un estudio de Segmentación de la siguiente forma:

1- Identificación de las variables para realizar la segmentación.

2- Desarrollo de los perfiles de cada segmento obtenido.

3- Evaluación del atractivo de cada segmento.

4- Selección del segmento o segmentos objetivos.

5- Identificación de posibles motivos para posicionarse en los segmentos seleccionados.

6- Seleccionar, desarrollar y crear estrategias de actuación para cada segmento objetivo.

En los estudios de mercado, o de actividades relacionadas con el Marketing, los Métodos de Segmentación juegan un rol destacado, a partir de la investigación realizada por Smith (1956), titulada: "Product Differentiation and Market Segmentation as Alternative Marketing Strategies".

Frank et al. (1972) presentan que el primer paso consiste en seleccionar como base para la segmentación, una variable dependiente, cuya conducta se pretende explicar, y un conjunto de variables explicativas o descriptoras de cada segmento. Éstas pueden ser de diferentes tipos: "características del individuo en general" (demográficas, socioeconómicas, de personalidad, de estilo de vida, de comportamiento,), "características en situaciones específicas" (beneficio esperado, necesidades, actitudes,), etc. Algunas de estas variables pueden ser medidas objetivamente (como la edad, renta, etc.) mientras que otras han de ser inferidas mediante valoraciones subjetivas (como actitudes, preferencias, etc.).

La recogida de información sobre las variables explicativas es un proceso complejo y costoso. La información puede provenir de fuentes existentes, o más frecuentemente mediante entrevistas, cuestionarios, etc.

El análisis de segmentación es una técnica de análisis multivariante y se ubica en los denominados métodos de dependencia. En este grupo de técnicas algunos de los métodos más destacados son:

-El método CHAID (*Chi-square AID*) propuesto por Kass (1980). Este método, es una de las versiones más interesantes de los conocidos métodos AID (*Automatic Interaction Detection)*.

-El método CART propuesto por Breiman et al. (1984). Difiere de los métdos AID en el modo de construcción del árbol de segmentación.

En el Departamento de Estadística de la Universidad de Salamanca se ha desarrollado una línea de investigación de algoritmos de segmentación:

-Desarrollos metodológicos y computacionales respecto a la fase de elección del mejor predictor (Ramírez, 1995).

-Los Algoritmos de Segmentación Descendentes Basados en Contrastes de Hipótesis de Independencia Condicionada (DÁVILA 1 y DÁVILA 2) presentados por Avila (1996).

-El Algoritmo Ascendente Basado en Criterios de Entropía (*ADORADO*), o el Algoritmo Descendente Basado en Criterios de Entropía (*DDORADO*), presentados por Dorado (1998).

-El Algoritmo TAID propuesto por Castro (2005).

3.2.-MÉTODOS DE DETECCIÓN AUTOMÁTICA DE LA INTERACCIÓN (AID)

El análisis de segmentación está basado en los métodos de detección automática de la interacción "Automatic Interaction Detection" (AID) propuestos por Morgan & Sonquist, (1963).

Los métodos AID utilizan un procedimiento de tipo iterativo generando en el diagrama arborescente (dendograma) divisiones sucesivas hasta un nodo no significativo o con un tamaño mínimo del grupo. Es un método predictivo de segmentación jerárquica descendente. Para llegar a los grupos finales el procedimiento se desarrolla por etapas

mediante la subdivisión sucesiva de una muestra en una serie de grupos excluyentes. La principal ventaja radica en la simplicidad, el resultado es un diagrama de árbol.

El objetivo de estos métodos es detectar la existencia de interacción en un modelo de predicción, además de ser usados con fines exploratorios y descriptivos.

3.3.-ALGORITMO CHAID

El algoritmo CHAID propuesto por Kass (1980) está pensado para una variable respuesta de tipo cualitativo y predictores cualitativos y se basa en el test chi-cuadrado para contrastar independencia en las distintas etapas del proceso. Es considerado un algoritmo general de segmentación, se utiliza con fines exploratorios y descriptivos, con el objetivo fundamental de encontrar la partición de una muestra de objetos en grupos, capaces de describir de la mejor manera posible la variable dependiente. Una parte importante de la segmentación se basa en contrastes sobre tablas marginales.

El algoritmo CHAID divide la muestra en dos o más grupos distintos, en base al predictor o variable independiente más significativa, con respecto de la variable dependiente, en donde se busca que los objetos que pertenecen al mismo grupo sean lo más homogéneos posibles y objetos de diferente grupo sean heterogéneos. Continúa dividiendo recursivamente, en base a las categorías definidas por la variable independientes o predictor más significativo. Cada uno de estos grupos se dividen en subgrupos más pequeños, en base a otras variables independientes. Este proceso iterativo de partición continúa, hasta no encontrar ninguna variable independiente estadísticamente significativa. CHAID muestra los resultados de la segmentación en forma de un diagrama

de árbol, cuyos nodos o ramas corresponden a los grupos. Los segmentos que CHAID construye, son mutuamente exclusivos y exhaustivos, es decir los segmentos no se sobreponen, y cada objeto de la muestra esta contendida exactamente en un segmento.

El algoritmo CHAID es un proceso secuencial multietápico que presenta cuatro etapas:

Etapa I

En esta etapa Kass propone realizar el test chi-cuadrado, cruzando la variable respuesta con cada predictor, y ver qué categorías tienen un perfil similar con respecto a la variable dependiente o respuesta (que no son significativas). En ese caso, dichas categorías se agrupan o se colapsan. Propone cruzar cada par de categorías y fusionar el par con mayor p-valor no significativo.

El proceso se repite con las categorías agrupadas o ***colapsadas*** para valorar si las nuevas categorías producen nuevas fusiones, hasta que no existan valores no significativos. Es decir, el proceso termina, cuando todas las categorías son significativamente diferentes, o bien cuando se han ***colapsado*** todas.

El agrupamiento o ***colapsamiento*** de las categorías de las variables, está en relación con el tipo de predictor o variable independiente de que se trate. Pueden considerarse los siguientes tipos de predictores:

- Predictor Monótono

Es de tipo ordinal, es decir, las categorías siguen un orden establecido, de modo que sólo se podrán agrupar dos categorías contiguas. Por ejemplo, la variable nivel de estudios. Si esta variable tuviera como valores: "primarios", "secundarios" y "universitarios", el procedimiento permitiría la fusión de las categorías primera y segunda o segunda y tercera, y descartaría la posibilidad de formar un grupo compuesto por sujetos con estudios primarios y universitarios.

Cuando la primera y última categoría se consideran adyacentes se denominan predictores monótonos cíclicos.

- Predictor Libre

Es aquel que sus categorías son de tipo nominal, o sea, no se puede establecer un criterio de orden en sus categorías. En este caso, sí tendrá sentido agrupar dos categorías de forma aleatoria, sin necesidad de que sean contiguas, es decir, cualquier par de categorías puede ser agrupada. Una variable de este tipo es por ejemplo la situación laboral con los valores: "activo", "parado", e "inactivo". La categoría "activo" podría formar grupo con "parados" y/o "inactivo".

- Predictor Flotante

Es aquel que tiene todas sus categorías, es una escala ordinal menos una que es desconocida su lugar en la escala ordinal y se denomina categoría flotante. En este caso al igual que con un predictor monótono, solo se pueden agrupar dos categorías contiguas con excepción de la flotante que podrá quedar sola o unirse a cualquiera de los grupos ya

formados. Por ejemplo, consideraremos que la variable nivel de estudios, tuviera el valor "Ns/Nc". generalmente el "no sabe, no contesta", puede agregarse libremente a cualquiera de las categorías establecidas.

Etapa II

Finalizada la fase anterior, se selecciona el mejor predictor.

El mejor predictor será aquel que presente una mayor asociación, con la variable dependiente, es decir, aquel que tenga el *p*-valor más pequeño o el mayor valor para el coeficiente de asociación elegido. Entonces, el mejor predictor será aquel que discrimine mejor a los individuos según la variable respuesta.

Etapa III

Fijado el p-valor si el predictor seleccionado es significativo, se realiza la segmentación del grupo, considerado en tantos grupos como categorías o niveles tenga el predictor.

Etapa IV

Finalizada la segmentación en la etapa III, se repite el proceso desde la primera etapa y se realizan sucesivas segmentaciones hasta que no haya predictores significativos, en ninguno de los grupos restantes.

En este proceso de finalizar el proceso de segmentación fijando sólo la ausencia de predictores significativos como criterio de parada nos lleva a tener árboles muy poco ocupados, por lo que se cometería un error importante, ya que el estadístico chi-cuadrado se obtendría probablemente a partir de tablas poco ocupadas. Del mismo modo que

seleccionar un tipo de predictor es importante, también lo es limitar el proceso de segmentación mediante controles o filtros. Los filtros de proceso más utilizados son:

- *Significación de Categoría* (SC): Se refiere al nivel de significación utilizado en la fase de agrupación de categorías. Para verificar si dos categorías tienen un perfil semejante, esto es, no son significativamente diferentes, se compara su nivel de significación, con la SC. En el algoritmo CHAID se lleva a cabo cruzando la variable dependiente con las dos categorías del predictor, se calcula el chi-cuadrado, y se compara su valor *p* correspondiente con la SC.

- *Significación del Predictor* (SP): Es el nivel de significación utilizado en la fase de selección del mejor predictor, es decir, para verificar que un predictor es significativo, se compara su significación con SP. En CHAID se lleva a cabo cruzando la variable dependiente con el predictor ya agrupado, se calcula el chi-cuadrado y se compara el valor *p* correspondiente con SP.

- *Filtros de Asociación* (FA): Se trata de fijar una asociación mínima entre la variable dependiente y el predictor, para considerarlo como un potencial candidato para realizar la segmentación. Esto plantea el problema de escoger un coeficiente de asociación entre una gran cantidad de ellos. Si el coeficiente de asociación elegido entre la variable dependiente y el predictor es menor que FA, éste es descartado. Entre algunos de los indicadores más utilizados están el Coeficiente de Contingencia (CC):

$$CC = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

del cual se dice que es igual a cero, cuando hay independencia absoluta entre las variables, pero no es igual a 1 cuando hay dependencia total (su valor máximo depende del número de categorías de las variables).

$$CCmax = \sqrt{\frac{r-1}{r}}$$

dónde:

r es mínimo entre el número de filas y de columnas.

- *Tamaño Antes* (TA): Se establece un tamaño mínimo para que un grupo pueda segmentarse, es decir, si un grupo *G* cualquiera, es menor que TA en individuos, el grupo no se segmenta y se declara terminal.

- *Tamaño Después* (TD): Se establece un tamaño mínimo para que un subgrupo pueda formarse, es decir, si algún grupo formado en la segmentación de G, digamos Gj, tiene menos de TD individuos, la segmentación es descartada.

- *Filtro de Nivel* (FN): Se establece un máximo de número de nivel de segmentación, también denominados niveles de profundidad. Una segmentación con un solo nivel resulta demasiado simple, pero por otro lado una segmentación de muchos niveles puede resultar compleja manejar.

3.3.1.- LIMITACIONES DEL ALGORITMO CHAID

Dentro de la problemática que presenta el algoritmo CHAID cabe mencionar:

1) La Etapa I del algoritmo no tiene en cuenta el carácter asimétrico de las variables, presupone un papel simétrico entre la variable dependiente e independientes.
2) Trabaja sobre información marginal, no garantizando la colapsabilidad de variables.
3) No recoge información ordinal.
4) Elevado número de contrastes, lo que conlleva aun alto riesgo de error tipo I.
5) No considera respuesta multivariante.

El algoritmo CHAID supone que la variable dependiente es categórica y utiliza la prueba Chi-cuadrado, para contrastar independencia, en distintas fases del proceso. Una parte importante de la segmentación se basa en contrastes sobre tablas marginales. Esto lleva a que ocurra el problema 2, ya que nos podemos encontrar con casos en los que dos variables sean independientes de manera marginal pero no lo sean cuando se estudien junto a otras. Esto se conoce como Paradoja de Simpson, en honor de Edward Simpson, quien la describió en 1951 (Simpson, 1951), aunque fue previamente descrita por el estadístico británico G. Udny Yule a inicios de 1900.

La Paradoja de Simpson describe la desaparición de una asociación o comparación significativa de dos variables cuando los datos son desagregados por grupos.

3.4.-ALGORITMO CART

Los árboles de clasificación y regresión (CART) fueron desarrollados por Breiman, Freidman, Olshen y Stone en el libro *Classification and Regression Trees* publicado en 1984.

Desde el planteamiento de los árboles de clasificación y regresión CART por Breiman et al. (1984):

(a) Se presentó gran interés en la utilización de esta metodología por parte de la comunidad científica en distintos campos como la medicina, la biología, debido a su fácil implementación en todo tipo de problemas y su clara interpretación de los resultados.

(b) Aunque algunos autores posteriormente han planteado variaciones del método en sus distintas etapas, la idea inicial del particionamiento recursivo es la misma.

(c) Son una alternativa al análisis de clasificación o predicción tradicional.

Los árboles CART permiten utilizar variables numéricas y/o categóricas. Entre las principales ventajas cabe destacar su robustez a outliers, la invarianza en la estructura de sus árboles de clasificación o de regresión a transformaciones monótonas de las variables independientes y sobre todo, su interpretabilidad.

Un árbol de clasificación consta de tres tipos de nodos: raíz, interno y terminal. Existe un único nodo raíz que contiene a todas las observaciones. A partir de él se dividen en dos ramas, cada una de las cuales da lugar a un nodo que puede ser interno o terminal. El nodo se dirá interno cuando, a su vez, se divide en dos ramas, en caso en que no se divida se dirá terminal. La metodología para construir árboles de regresión es la misma, la diferencia radica en la elección de la función de impureza para dividir un nodo.

3.4.1.-FUNCIÓN IMPUREZA

Árboles de clasificación

La construcción del árbol se hace a partir del algoritmo denominado particionamiento recursivo (Izenman, 2008) y es la clave en el método estadístico no paramétrico CART. El planteamiento estadístico consiste en establecer una relación entre una variable respuesta Y, y las p variables predictoras $X_1, X_2, \ldots, X_p$, donde las X´s son tomadas fijas de tal forma que sea posible predecir Y basado en los valores de las X´s. Se quiere estimar la probabilidad condicional de la variable aleatoria Y:

$$P\ (Y = y/\ X = x_1, x_{2,\ldots,} x_p)$$

El algoritmo CART, por tanto, es un método que genera un árbol binario a través de particiones binarias recursivas. Divide los datos en subconjuntos, donde cada división se basa en una sola variable, pudiendo ser usada varias veces varias variables mientras que otras pueden no resultar significativas. Para elegir la mejor variable debe utilizarse una medida de pureza (purity) en la valoración de los 2 nodos hijos posibles (la variable que consigue una mayor pureza se convierte en la utilizada en primer lugar, y así sucesivamente). Debe buscarse una función de partición (splitting function) que asegure que la pureza en los nodos hijos sea la máxima. El objetivo es acabar en nodos terminales que sean homogéneos.

Sea Y una variable dicotómica con valores 0 y 1. Para construir el árbol saturado, en el proceso de particionamiento recursivo se tiene que para el nodo menos impuro la impureza es 0 y debe tener como resultado $P(Y = 1/\delta) = 0$ o $P(Y = 1/\delta) = 1$. El nodo δ es más impuro cuando su impureza es 1 con $P(Y = 1/\delta) = \frac{1}{2}$. Por lo cual la función de impureza se puede definir como $i(\delta) = \theta(\{Y = 1/\delta\})$.

Donde θ tiene las siguientes propiedades:

- $\theta \geq 0$
- Para cualquier $p \in (0,1), \theta(p) = \theta(1-p)\ y\ \theta(0) = \theta(1) < \theta(p)$

Las funciones de impureza más comunes para la construcción de árboles de clasificación son:

- $\theta(p) = \min(p, 1-p)$, (mínimo error o error de Bayes)
- $\theta(p) = -plog(p) - (1-p)\log(1-p)$, (entropía)
- $\theta(p) = p(1-p)$ (índice Gini)

Árboles de regresión

Para una respuesta continua, la elección de la impureza para un nodo δ es la varianza de la respuesta dentro del nodo:

$$i(\delta) = \sum_{sujeto\ i \in \delta} \left(Y_i - \bar{Y}(\delta)\right)^2$$

Donde $\bar{Y}(\delta)$ es el promedio de las $Y_i^{'}$s dentro del nodo δ. Para dividir un nodo δ en dos nodos hijos, $\delta L\ y\ \delta R$ se define la función de división

$$\varphi(s, \delta) = i(\delta) - i(\delta L) - i(\delta R),$$

donde s es la división permitida para el nodo δ.

3.4.2.-CRITERIO DE LA PODA

El árbol obtenido es generalmente sobre ajustado o saturado por lo tanto es podado. La poda consiste en encontrar el subárbol del árbol saturado con la mejor calidad en cuanto a que sea el más predictivo de los resultados y menos vulnerable al ruido en los datos. Es decir, la calidad de un árbol es estrictamente la calidad de sus nodos terminales. Para esto, se debe definir una medida de calidad de un árbol.

Por lo tanto, para un árbol S se define:

$$R(S) = \sum_{\tau \in \tilde{S}} P(\delta) r(\delta)$$

Dónde:

$\tilde{S}$ es el conjunto de nodos terminales de S

$r(\delta)$ es una medida de calidad del nodo δ la cual es similar a la suma de cuadrados de los residuales en regresión lineal.

La intención de la poda es seleccionar el mejor subárbol, $\tilde{S}$, de un árbol saturado inicialmente S_0, tal que R (S) sea mínimo.

3.5.-ALGORITMO DÁVILA

Avila (1996) propone algoritmos basados en hipótesis de independencia condicionada con la forma **i ⫫ j / V \ {j},** ya que implican que para estudiar la relación entre **i y V \ {j},** es posible colapsar J. Los métodos de segmentación que propone son descendentes, partiendo de todas las variables explicativas originales concatenadas o el árbol completo. Los algoritmos de Avila (1996) usando hipótesis de independencia condicionada analizan las tablas trifactoriales sin necesidad de reducirlas a tablas bifactoriales.

En una tabla bidimensional, solo una hipótesis de interés debe ser contrastada. Cuando en un análisis intervienen tres variables, van existir varias hipótesis de interés, las cuales puede ser expresadas de acuerdo a Dawid(1979) y posteriormente utilizadas por Whittaker (1990) o Cox & Wermuth (1996).

En primer lugar tenemos la independencia completa:

i ⫫ j ⫫ k, donde las tres variables (i,j,k) son mutuamente independientes.

En segundo lugar, tenemos la independencia múltiple:

i ⫫ (j, k), j ⫫ (i, k) o k ⫫ (i, j), una de las variables es independiente de las otras dos, pero estas dos pueden estar relacionadas.

Y en tercer lugar, tenemos la independencia condicionada:

i ⫫ j / k, i ⫫ k / j o k ⫫ j / i, donde las dos primeras variables de cada caso son independientes entre sí para cada nivel de la otra, pero cualquiera de ella puede estar relacionada con la tercera e, incluso, ambas pueden estarlo.

Matemáticamente, un grafo G es un conjunto de vértices V y de bordes B que son pares de elementos tomados de V.

$$G = \{V, B\}$$

Una tabla trifactorial *i, j, k* será colapsable sobre el factor *i* si los odds ratio en la tabla marginal $p_{.jk}$ son idénticos a los odds ratio para cada fila de la tabla trifactorial de partida.

En términos de odds ratio diremos que una tabla trifactorial es colapsable sobre *i,* si para todo *i, j, j', k, k'* se verifica:

$$\frac{p_{.jk}p_{.j'k'}}{p_{.j'k}p_{jk'}} = \frac{p_{ijk}p_{ij'k'}}{p_{ij'k}p_{ijk'}}$$

3.5.1.-ALGORITMO 1 (DÁVILA 1)

El algoritmo 1 llamado DÁVILA1 de segmentación descendente basado en hipótesis independiente condicionada, sigue el siguiente procedimiento.

Etapa 1

En esta etapa se crea el árbol completo en el que se consideran todas las variables explicativas y se colapsan o eliminan las variables sin información significativa.

Etapa 2

En esta etapa se buscan aquellos predictores que sean independientes de la variable respuesta.

Sea i la variable respuesta y V representa los posibles predictores. Se buscan todas las variables $k \in V$ tales que $i \coprod k / V \setminus \{k\}$.

Todas las variables para las cuales $k \in V$ esta hipótesis se acepta son eliminadas y el árbol es podado en relación a todas esas variables.

Etapa 3

En esta etapa se segmenta la población. El objetivo es colapsarla en segmentos. Se elige como variable para segmentar aquella que tenga el ***p-valor más pequeño*** en la hipótesis de independencia contrastada, sobre la tabla colapsada.

Supongamos que el mejor predictor es la variable j. El árbol tiene J ramas, que representan J subgrupos que serán analizados por separado.

Etapa 4

En esta etapa se agrupan las categorías del predictor. La idea es unir las ramas del árbol que tengan un perfil de respuesta similar.

Por lo cual constatamos la hipótesis de independencia condicionada:

i ⫫ j (j_1,j_2) / V \ {j (j_1,j_2)} donde la notación utilizada significa que se restringe la variable j a sus categorías (j_1,j_2). y se contrasta la hipótesis para cada pareja de categorías de la variable.

Etapa 5

Esta etapa es el de la interacción. Se repiten los pasos 1 a 4 en cada segmento. Se repite el algoritmo hasta que todos los nodos sean terminales, no se pueda seguir colapsando y no hay variables para separar.

3.5.2.-ALGORITMO 2 (DÁVILA 2)

Este algoritmo coincide con el algoritmo1 en las etapas 1 y 2, y también está basado en contrastes de hipótesis de independencia condicionada.

Etapa 3

En esta se efectúa la segmentación de la población, En este caso se realiza el contraste de hipótesis condicionada, pero, se escogerá como variable para colapsar aquella que, siendo significativa, presenta el ***p-valor más grande***, ya que es más probable encontrar subgrupos en los cuales se pueda colapsar.

La búsqueda de esos subgrupos se lleva a cabo realizando el contraste de hipótesis **i ⫫ m / (k, V \ {m, k}) con** k=1,…,K ∀ **k ∈ V \ {m},** y se colapsa en aquellos grupos en los que el contraste sea no significativo.

De todas las posibles variables se selecciona aquella con menor p-valor. Las categorías de esta variable serán las que nos den las distintas ramificaciones en las que segmentaremos a la población.

En el caso de no poder colapsar dentro de las categorías de ninguna de las variables se concatenan dos, tres, etc. y se vuelve a repetir todo el proceso.

El proceso continúa hasta no encontrar ninguna rama en la que se pueda colapsar.

Etapa 4

Se repiten los pasos anteriores hasta cuando todos los nodos sean terminales por no poder seguir colapsando y por no tener variables para separar.

3.6.-ALGORITMO DORADO

Las alternativas que propone Dorado (1998) están basada en la utilización de algoritmos ascendentes y descendentes basados en criterios de entropía y contrastes de independencia condicionada.

3.6.1.-ALGORITMO (ADORADO)

Es un algoritmo divisivo, ascendente y está basado en criterios de entropía. Los pasos del algoritmo son los siguientes:

Etapa 1

Para valorar el grado de heterogeneidad ("desorden") existente en el conjunto de partida, calculamos la entropía correspondiente a la variable respuesta **i**.

$$H(i) = -\sum_{i} P_i log P_i$$

Pretendemos encontrar una segmentación de la población considerada, que proporcione los segmentos más homogéneos posibles, en relación a la respuesta. Esa división podemos hacerla en relación a cualquiera de las variables explicativas. De entre ellas, elegiremos aquella que produzca un mayor descenso en la entropía; por tanto, calcularemos:

$$H\left(i/_j\right) = -\frac{i}{j}\sum_{i}^{J}\sum_{j}^{J} P\left(i = i/_j\right) log P\left(i = i/_j\right) \quad \forall j \in V$$

Dónde V = {j, k, l, ...} son las variables explicativas:

$$I\left(i/_j\right) = \hat{H}(i) - \hat{H}\left(i/_j\right)$$

Teniendo en cuenta que $T = 2N\hat{I}$ sigue una ji-cuadrado con (I-1)(J-1) grados de libertad, podemos saber si el cambio en la entropía media condicionada es significativo.

De entre todas las variables que produzcan un descenso significativo en la entropía, elegiremos aquella para la cual $\hat{H}(i) - \hat{H}\left(i/_j\right)$ sea mayor, y segmentaremos en tantos segmentos como categorías presenta la variable elegida.

Etapa 2

En esta etapa se repite el proceso de la primera etapa en cada segmento hasta que no encontremos ninguna variable predictora que produzca un descenso de la entropía significativo.

3.6.2.-ALGORITMO (DDORADO)

Este algoritmo recoge las ventajas de los algoritmos anteriormente descritos, denotados como **DÁVILA 2** Y **ADORADO**. Está basado en criterios de entropía asociados a contrastes de independencia condicionada.

Etapa 1

En esta etapa se crea el árbol completo en el que se consideran todas las variables explicativas. Sin importar el orden en que lo hagamos.

Etapa 2

En esta etapa se *establece la búsqueda de las variables independientes de la respuesta.* Se pretende saber si es posible eliminar variables que no aporten información al comportamiento de la variable respuesta. Para esto se calculará la entropía de cada variable y se eliminarán aquellas que no produzcan un incremento significativo de la entropía.

Todas las variables que verifiquen esa condición se eliminan del estudio y el árbol pierde automáticamente las ramas relacionadas con esa(s) variable(s). En el caso de que todas las variables fuesen no significativas, solo se eliminaría aquella con mayor p-valor y se repite el proceso.

Etapa 3

Una vez que ya hemos descartado las variables para las cuales no se produce un incremento en la entropía significativo a nivel global, trataremos de analizar variable a variable la diferencia de la entropía en cada una de sus categorías ya que el hecho de que

el incremento de esta sea significativo puede deberse solo a una de ellas, es decir, si en una categoría el incremento de la entropía es significativo, bastaría para que fuera significativo el contraste.

Etapa 4

Una vez que los pasos 1, 2 y 3 se han repetido las veces necesarias para llegar al árbol definitivo, y descartada la posibilidad de seguir colapsando, puede ocurrir que encontremos ciertas ramas para las cuales aparezca la misma estructura. Si eso ocurre, se pasaría a realizar el análisis de ramas simétricas en el árbol simétrico abordado en Dorado, (1998).

3.7.-ALGORITMO TAID

El algoritmo TAID fue propuesto por Castro (2005). Este algoritmo difiere de los anteriormente descritos en que para segmentar utiliza el análisis no simétrico de correspondencias y se basa en las ideas de Siciliano & Mola (1997) que apuntan que no solo se debe prestar atención a la identificación de variables para segmentar, sino también tratar de conocer qué categorías de cada variable explicativa son las que mejor predicen cada categoría de la respuesta y analizar si el poder predictivo de la categoría es positivo o negativo. Para describir el algoritmo TAID se van a describir previamente algunos análisis estadísticos como el análisis de clases latentes, el coeficiente de predictividad y el índice de Catanova.

3.7.1.-ANÁLISIS DE LAS CLASES LATENTES

Los algoritmos utilizados tradicionalmente solo contemplan una variable respuesta. En aquellos casos en los que existan varias variables respuesta, el primer paso será definir una variable latente que recoja el carácter multivariante de la respuesta.

Existen dos supuestos a considerar en un análisis de clases latentes, uno de ellos es el supuesto de *independencia local*, el cual considera que las variables predictoras llamadas variables manifiestas son estadísticamente independientes dentro de cada clase latente, por lo que la relación entre estas variables viene dada exclusivamente por la pertenencia de un individuo a una clase en particular, ya que se espera que si la variable latente permanece constante, cualquier relación existente entre las variable manifiestas desaparece. Clogg, (1988) señala que un aspecto importante y que tiene que ver con la ***colapsabilidad de categorías***, es que, si para un conjunto de variables manifiestas el supuesto de independencia local se verifica, también se verificará para un subconjunto de estas variables.

El segundo supuesto es que las clases latentes son *internamente homogéneas*, es decir, que los individuos que pertenecen a la misma clase latente tendrán la misma distribución de probabilidad, con respecto a la variable latente, y ésta será distinta a la distribución de probabilidades para los individuos pertenecientes a otra clase, por lo que individuos de diferentes clases presentarán características diferentes. Este hecho sirve para diferenciar a los individuos pertenecientes a diferentes clases y poder caracterizar tanto la variable latente como las clases latentes.

La variable latente se representará como *Y* con *T* categorías latentes. *donde cada categoría de Y es una clase latente y p* variables manifiestas l_1, l_2,…, l_p se consideran como indicadoras de la variable latente *Y*.

$$P(L = l) = \sum_t^T P(Y = t, L = l)$$

$$= \sum_t^T P(Y = t) P(^{L = l}/_{Y = t})$$

$$= \sum_t^T P(Y = t) \prod_{j=1}^p P(^{l_j = l_q}/_{Y = t})$$

Dónde:

$l = (l_1, l_2, \dots, l_p)$ denota un determinado patrón de respuesta en el cual cada una de las l_j toman diferentes valores dependiendo de las categorías de la correspondiente variable manifiesta.

P(L=l) es la probabilidad conjunta de las variables manifiestas.

P (Y=t, L=l) es la probabilidad conjunta de tener un patrón de respuesta l y pertenecer a la clase latente *t*.

P(Y = *t*), es la probabilidad de pertenecer a la clase latente *t*, conocida como probabilidad *a priori,* $P(^{L = l}/_{Y = t})$ es la probabilidad condicional de obtener un determinado patrón de respuesta para un individuo de la clase latente *t*.

$P(^{l_j = l_s}/_{Y = t})$es la probabilidad de obtener un determinado valor en la variable *ls*, dado que se está en la clase latente *t*.

Los individuos se asignan a la clase latente para la cual su probabilidad *a posteriori*

p (Y = *t* / L = l) es mayor.

El teorema de Bayes se utiliza para estimar las probabilidades *a posteriori:*

$$P(Y = t/L = l) = \frac{P(Y = t, L = l)}{P(L = l)}$$

Una vez identificada las clases latentes, el problema de la segmentación se aborda como en el caso en el que teníamos una sola variable respuesta ya que aquí cada clase latente juega un papel similar al que jugaría cada categoría de la variable respuesta cuando ésta tenía carácter univariante.

3.7.2.-COEFICIENTES DE PREDICTIVIDAD E ÍNDICE DE CATANOVA

El coeficiente de predictividad φ fue introducido originariamente por Goodman & Kruskal (1954) para una matriz de probabilidad para medir el incremento relativo de la probabilidad de predecir correctamente la variable fila conociendo el nivel de la variable columna. Posteriormente Light & Margolin (1971), lo utilizaron para una muestra con el fin de analizar la heterogeneidad o la variabilidad de datos categóricos.

Se considera una tabla de contingencia de dos vías I×K donde I y K son el número de categorías de las variables Y y X respectivamente. Se denota por:

f_{ik} a la frecuencia observada (es decir, número de los sujetos que pertenecen conjuntamente a la categoría i-ésima de la variable I y a la categoría k-ésima de la variable X).

$f_{i.}$ el total marginal de cada fila

$f_{.k}$ el total marginal de cada columna

$f_{..}$ el total general

$p_{i.} = \frac{f_{i.}}{n}$ son las probabilidades totales por filas

$p_{.k} = \frac{f_{.k}}{n}$ son las probabilidades totales por columnas

$p_{ik} = \frac{f_{ik}}{n}$ son las probabilidades relativas de cada celda

El índice de predictividad se define de la siguiente manera:

$$\varphi = \frac{\left(\sum_{ik} \frac{f_{ik}^2}{f_{..} f_{,k}} - \sum_i \left({}^{f_{i.}}/_{f_{..}}\right)^2\right)}{1 - \sum_i \left({}^{f_{i.}}/_{f_{..}}\right)^2}$$

dónde el denominador corresponde con el coeficiente de heterogeneidad de GINI y explica la medida de heterogeneidad de las categorías de la variable respuesta. Por otro lado, el numerador, que corresponde a la heterogeneidad explicada, es la heterogeneidad provocada por el poder predictivo de las categorías del predictor. Varía entre 0 (sin poder predictivo) y 1 (predicción perfecta), esto es:

- Si $\varphi = 0$ existe independencia para cada celda (i,k), es decir $p_{i.} = \frac{p_{ik}}{p_{.k}} = p_{i.}$
- Si $\varphi = 1$ si para cada categoría columna *k* existe una categoría fila *i* tal que $p_{ik} = p_{.k}$

Procede hacer la primera segmentación siempre que algunos de los índices de predictividad sean significativos.

Para estudiar la significatividad del índice de predictividad se utiliza el índice de Catanova propuesto por Ligth & Margolin (1974) y extendido al análisis de categorías ordinales por Anderson & Landis (1982) con una variable respuesta ordinal. El test chi-cuadrado requiere una restricción en la frecuencia esperada que en el caso del método de Catanova no existe.

Este índice nos permite probar si la predictividad es significativa, ya que:

$$C = (I-1)\,(J-1)\,\varphi$$

sigue una distribución chi-cuadrado con (I-1) (J-1) grados de libertad

3.7.3.-ANÁLISIS NO SIMÉTRICO DE CORRESPONDENCIAS

El análisis de correspondencias es utilizado para detectar la relación entre variables categóricas (no necesariamente dicotómicas). Generalmente estas técnicas utilizan como gráficos el Biplot, mapas factoriales simétricos y asimétricos con coordenadas estándar y coordenadas principales.

En el análisis de correspondencias cuando se estudian dos variables categóricas sin necesidad de que quede impuesto cuál de ellas es la variable dependiente y cuál es la independiente decimos que estamos ante un análisis de correspondencias simétrico.
En el caso, cuando exista una relación de dependencia entre las variables, entonces decimos que estamos ante un análisis de correspondencias no simétrico según Beh (2008).

El análisis de correspondencias no simétrico fue propuesto por Lauro & D'Ambra (1984) como una variación del análisis de correspondencias simétrico propuesto por Benzécri (1973).

El análisis de correspondencias simétrico utiliza el índice de asociación Φ^2, que está basado en el test chi-cuadrado y supone una relación simétrica de las variables. Lauro & D'Ambra (1984) propusieron un nuevo estadístico, el coeficiente de predictividad φ descrito previamente.

Características del análisis de correspondencias no simétrico:

1. En este análisis la variable respuesta es considerada la variable dependiente y se expone por filas y el predictor como variable independiente se coloca en columnas.
2. El estadístico usado se basa en el φ de Goodman & Kruskal (1954). Igualmente se confirma lo siguiente:

    - Si se verifica que $\varphi = 0$ entonces la variable respuesta no se presenta alterada por el predictor o la variable independiente.

    - Si se verifica que $\varphi = 1$ podemos decir que la variable independiente del estudio explica completamente a la variable dependiente.

3. En este análisis si alguna categoría de la variable respuesta se representa cerca de una categoría de la variable independiente podemos decir que presentan una fuerte relación. De forma similar, si sendas categorías no se representan de manera cercana podemos decir que la casilla de estudio tiene un porcentaje pequeño de individuos sobre el total.

    La distancia de un punto k al origen será:

$$d^2(k, O) = \sum_{i=1}^{I} \left(\frac{f_{ik}}{f_{.k}} - \frac{f_{i.}}{f_{..}} \right)^2$$

Diremos entonces que un punto cualquiera k estará más alejado del origen cuanto mayor sea la desviación a la hipótesis de independencia.

De forma similar tenemos la distancia de un punto *i* al origen:

$$d^2(i,O) = \sum_{k=1}^{K} \left(\frac{f_{.k}}{f_{..}}\right)\left(\frac{f_{ik}}{f_{.k}} - \frac{f_{i.}}{f_{..}}\right)^2$$

esta distancia se ve afectada por el peso de las columnas, por ello, las columnas con mayor peso son las que hacen que los puntos fila se alejen más del origen.

4. Las ventajas del análisis de correspondencias no simétrico con respecto al simétrico son que:
    - La inercia total en el simétrico es sensible a proporciones marginales pequeñas de la variable respuesta. Contrariamente, esto no ocurre en el caso del análisis no simétrico.
    - Las columnas con mayor peso serán las que más contribuyan a la inercia y aquellas con pesos pequeños podrían eliminarse.
    - El análisis de correspondencias no simétrico se basa en la métrica euclídea mientras que el simétrico se basa en la métrica chi-cuadrado.

3.7.4.-DESARROLLO DEL ALGORITMO TAID

El Algoritmo TAID se desarrolla en las siguientes etapas:

Etapa I

Este algoritmo permite presentar varias variables respuestas. En este caso se definirá una variable latente capaz de recoger el carácter multivariante del conjunto de variables respuesta, es decir, se realiza un modelo de clases latentes con todas las variables dependientes.

Etapa II

En esta etapa hemos de buscar el mejor predictor para segmentar. Este algoritmo utiliza para segmentar el coeficiente de predictividad y el cálculo del índice de Catanova, De entre todos los predictores significativos se escogerá para segmentar aquel con un mayor coeficiente de predictividad.

Etapa III

El algoritmo se base en las ideas de Siciliano & Mola (1997) de construir árboles ternarios. Esta construcción se base a la representación de la inercia recogida en el índice de predictividad en el plano factorial del análisis de correspondencias no simétrico.

Siguiendo las ideas de Siciliano & Mola (1997), se clasificaran las categorías de la siguiente manera:

$\theta_{j1} > 1$ categorías fuertes por la derecha

$\theta_{j1} < 1$ categorías débiles

$\theta_{j1} \leq -1$ categorías fuertes por la izquierda

donde θ_{j1} representa la correspondiente coordenada sobre el primer eje factorial del análisis de correspondencia no simétrico. Se crearán tres segmentos en la población objeto de estudio: uno con las categorías con alto valor predictivo positivo, otro con alto poder predictivo negativo y otro con las categorías débiles; son categorías sin un claro poder predictivo, para ninguna de las categorías de la respuesta, es decir, con aquéllas que por sí solas no pueden explicar la respuesta y precisan la ayuda de otros predictores.

Etapa IV

En esta etapa se recogen los resultados obtenidos anteriormente, el mejor predictor y la clasificación de las categorías y se realiza la segmentación.

Etapa V

En esta fase se buscan los segmentos terminales. Se aplica unos criterios de parada. Se consideró terminal aquel nodo que no tuviera más predictores significativos, o sea que el p-valor asociado al índice de Catanova de los predictores restantes fuesen mayores que 0,05 o aquel nodo cuyo tamaño fuese inferior al 10% de la muestra total.

El procedimiento en esta etapa es, repetir las etapas dos, tres y cuatro en cada uno de los nodos que no sean terminales.

3.8.- PAQUETES ESTADÍSTICOS PARA EL DESARROLLO DE ALGORITMO DE SEGMENTACIÓN

En la siguiente tabla se presenta información sobre paquetes en el entorno estadístico de tipo comercial o código abierto para el desarrollo de algoritmo de segmentación.

Tabla 5. Paquetes que desarrollan algoritmos de segmentación

Método / Año	Paquete
CHAID **G. Kass (1980)**	• XLSTAT / EXCEL • STATA • SPSS
CART **Breinam et al (1980)**	• XLSTAT / EXCEL • STATA • SPSS • RPART /R
DAVILA1 **y DAVILA2** **C. Avila (1996)**	• No están programado en software especifico
DORADO **y ADORADO** **A.Dorado (1998)**	• No están programado en software especifico
TAID **C. Castro (2005)**	• Winmira

- **Paquete rpart**

El paquete rpart de R versión 4.1-15, realiza partición recursiva para clasificación, árboles de regresión y supervivencia. El argumento que se utiliza es:

rpart(formula, data, weights, subset, na.action = na.rpart, method, model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...).

dónde:

Formula	Una fórmula, con una respuesta, pero sin términos de interacción.
data	Un marco de datos opcional en donde hay que interpretar las variables nombradas en la fórmula.
weigths	Ponderaciones de los casos opcionales.
subset	Expresión opcional que expresa que solo un subconjunto de las filas de los datos debe usarse en el ajuste.
na.action	La acción predeterminada elimina todas las observaciones para las que falta y, pero mantiene aquellas en las que faltan uno o más predictores
method	Uno de "anova", "poisson", "clase" o "exp". Alternativamente, el método puede ser una lista de funciones llamadas init, split y eval.
model	Si es lógico: ¿guardar una copia del marco del modelo en el resultado? Si el valor de entrada para el modelo es un marco de modelo (probablemente de una llamada anterior a la función rpart), entonces este marco se usa en lugar de construir nuevos datos.
x	Mantener una copia de la matriz x en el resultado.
y	Mantenga una copia de la variable dependiente en el resultado. Si falta y se proporciona el modelo, el valor predeterminado es FALSO.
parms	Parámetros opcionales para la función de división.
control	Una lista de opciones que controlan los detalles del algoritmo rpart.
cost	Un vector de costos no negativos, uno para cada variable del modelo.
. . .	Los argumentos de rpart.control también se pueden especificar en la llamada a rpart. Se cotejan con la lista de argumentos válidos.

3.9.-CONTRIBUCIÓN AL ESTUDIO DE LOS ALGORITMOS DE SEGMENTACIÓN

El análisis de segmentación es una técnica de análisis multivariante y se ubica en los denominados métodos de dependencia, donde uno de los objetivos primordiales, es el de explicar o predecir, la variable(es) respuesta(s) a partir del conjunto de variables independientes o explicativas. El método CHAID (*Chi-square AID*), propuesto por Kass (1980) es una de las versiones más interesantes de los conocidos métodos AID (*Automatic Interaction Detection*. Se propone para una variable respuesta de tipo cualitativo y predictores cualitativos. Está basado en contrastes de asociación en tablas marginales, pero no garantiza la colapsabilidad sobre la que se basa y en caso de la paradoja de Simpson no es capaz de detectarla.

Seguidamente fue propuesto el método CART de Breiman et al. (1984). Se trata de una familia de métodos que sirven tanto para predecir una variable cuantitativa para hacer regresión como para predecir una variable cualitativa para hacer discriminación. Difiere de los métodos AID en el modo de construcción del árbol de segmentación Entre las ventajas de los árboles CART cabe destacar la robustez a outliers y su interpretabilidad.

En el Departamento de Estadística de la Universidad de Salamanca, existe una línea de investigación, respecto del desarrollo de métodos para la obtención de algoritmos de segmentación y propuestas de mejora en éstos. En este contexto, podemos mencionar: Los algoritmos de Segmentación Descendentes Basados en Contrastes de Hipótesis de Independencia Condicionada (DÁVILA 1 y DÁVILA 2), presentados en Avila (1996). Están basados en la utilización de hipótesis de independencia condicionada. Los métodos de segmentación que proponen son descendentes. Permiten estudiar las tablas

trifactoriales sin necesidad de reducirlas a tablas bifactoriales y que es posible colapsar una tabla multidimensional, corrigiendo la problemática de la paradoja de Simpson que se da en el algoritmo CHAID.

Posteriormente, se desarrollaron los Algoritmos (*ADORADO*), o (*DDORADO*), presentados por Dorado (1998). Estos algoritmos ascendentes (ADORADO) y descendente (DDORADO) están basados en criterios de entropía y contrastes de independencia condicionada. El algoritmo descendente basado en contrastes de entropía resuelve la limitación fundamental del algoritmo descendente basado en contrastes de hipótesis de independencia condicionada propuesto por ÁVILA en 1996, ya que las frecuencias bajas en la tabla de datos, que podían producir importantes incrementos en el riesgo tipo I.

La limitación de los algoritmos de Avila y Dorado es que no consideran varias variables respuesta y no están incorporados en software de tipo comercial o de código abierto, lo que dificulta su aplicación.

Por ultimo Castro (2005) propone el algoritmo TAID proponiendo como alternativa al estadístico Chi-cuadrado el índice de predictividad y el coeficiente de Catanova. Además, permite en aquellos casos en los que existan varias variables respuesta, definir una variable latente que recoja el carácter multivariante de la respuesta.

En la tabla 6 se presenta un resumen de los aspectos algebraico y computacionales de algoritmos de segmentación anteriormente descritos.

Tabla 6. Aspecto algebraico y computacionales de algoritmos de segmentación

Método / Autor	Medida	Ventajas	Desventajas
CHAID **G. Kass (1980)**	Test chi-cuadrado simétrico	Técnica fácil de aplicar	• Reducción de las tablas trifactoriales en bifactoriales. La Paradoja de Simpson • Utilizado el test chi-cuadrado simétrico • Se desarrollan cuando tiene un conjunto de datos en el cual sólo se considere una única variable respuesta
CART **Breinam et al (1980)**	$R(S) = \sum_{\tau \in \tilde{\tilde{S}}} P(\delta) r(\delta)$	• Robustez a outliers • Fácil interpretación	Se desarrollan cuando tiene un conjunto de datos en el cual sólo se considere una única variable respuesta
DAVILA1 **y** **DAVILA2** **C. Avila (1996)**	Hipótesis de independencia condicionada **i ∐ j / V \ {j}**	• Utilizan hipótesis de independencia condicionada • Estudia las tablas trifactoriales	• Se desarrollan cuando tiene un conjunto de datos en el cual sólo se considere una única variable respuesta • Falta de software especifico
DORADO **y** **ADORADO** **A.Dorado (1998)**	l $H(i) = -\sum_i P_i log P_i$	• Utiliza el criterio de entropía • Contrastaste de hipótesis condicionada • Estudia tablas trifactoriales	• Se desarrollan cuando tiene un conjunto de datos en el cual sólo se considere una única variable respuesta • Falta de software especifico
TAID **C. Castro (2005)**	$\varphi = \frac{\left(\sum_{ik} \frac{f_{ik}^2}{f_{..} f_{.k}} - \sum_i \left(f_{i.}/f_{..}\right)^2\right)}{1 - \sum_i \left(f_{i.}/f_{..}\right)^2}$	• Utilización del índice de predictividad y el método Catanova • Permite trabajar con más de una variable respuesta	• Falta de implementación en software de uso general

CAPÍTULO 4: AGRUPAMIENTO Y ANÁLISIS DE COMPONENTES PRINCIPALES DISJUNTOS

4.1.-INTRODUCCIÓN

Actualmente, vivimos en un mundo de grandes volúmenes de datos. Debido al gran avance y uso de las tecnologías de información los datos crecen en volumen, veracidad y velocidad. En la actualidad el gran volumen de información se debe a los avances electrónicos e informáticos, como satélites, bandas magnéticas, GPS, tecnologías web, y redes sociales.

En los años 90 surge el término de Big Data como consecuencia de la creación de internet en 1989 que abre el camino a la generación masiva de datos. Big Data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados. Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales.

El análisis de Big Data ayuda al mundo empresarial a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades, conseguir valor reduciendo costes, mejorar la toma de decisiones y ofertar nuevos productos y servicios. Esto, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más satisfechos.

Sin embargo, la calidad de datos de Big Data también se enfrenta algunos desafíos como son el manejo de muchas fuentes y tipos de datos y el enorme volumen de los mismos. Lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que

tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible. La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los datos permiten que las empresas se muevan mucho más rápidamente, sin problemas y de manera eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

La alta dimensionalidad puede significar cientos o miles de variables de entrada. Cuando se trata de datos de gran dimensión, suele ser útil reducir la dimensionalidad proyectando los datos a un subespacio de menor dimensión con la mínima pérdida de información posible. La reducción de la dimensionalidad produce una representación más compacta y más fácilmente interpretable, centrando la atención del usuario en las variables más relevantes.

La reducción y síntesis de objetos y variables es uno de los análisis más utilizados para la exploración de los datos. El objetivo es detectar la información más relevante que permita una interpretación adecuada de los datos. Para la reducción de la dimensionalidad de los individuos generalmente se utiliza el análisis de clúster utilizando un algoritmo de partición. Para la reducción de la dimensionalidad de las variables se suele utilizar el análisis de componentes principales (PCA) o el análisis factorial. El PCA es un método de reducción de la dimensionalidad que es sin duda el método clásico más conocido. Dado un conjunto de datos en un espacio multidimensional, PCA realiza un cambio del sistema

de coordenadas de tal manera que las primeras dimensiones recojan la mayor variabilidad de los datos en dicho sistema.

La reducción de objetos y variables se suele obtener aplicando las dos técnicas de forma secuencial. A menudo, esto se hace llevando a cabo primero un PCA y luego aplicando un algoritmo de agrupamiento en las puntuaciones de las primeras componentes de los individuos. Sin embargo, DeSarbo et al. (1990) (1990), De Soete y Carroll (1994) y Vichi y Kiers (2001) desaconsejan este enfoque, llamado "*análisis en tándem*", porque el PCA o FA podrían identificar dimensiones que no necesariamente contribuyen mucho a percibir la estructura del agrupamiento en los datos y que, por el contrario, podrían ocultar o enmascarar la información taxonómica.

Vichi & Saporta (2009) proponen una nueva metodología llamada Agrupación en Conglomerados y Análisis de Componentes Principales Disjunto (CDPCA) que busca agrupar simultáneamente individuos y variables.

4.2.-MODELO DE AGRUPACIÓN EN CLÚSTERES Y PCA DISJUNTO

El modelo propuesto por Vichi & Saporta (2009) detecta una partición óptima de variables en Q clases. Para cada clase de la partición encuentra una componente que es la combinación lineal de las variables en la clase con varianza máxima.

A continuación, se enumera la notación y la terminología comunes utilizada en el desarrollo del capítulo.

$X_{(IxJ)} = [x_{ij}]$, matriz de datos de dos vías (objetos o individuos y variables) que describe los perfiles J-variable de I objetos. Las variables a analizar se suponen adecuadas, y por lo tanto si se expresan en diferentes unidades de medida se estandarizan para tener media cero y varianza unitaria.

$E_{(IxJ)} = [e_{ij}]$ (IxJ) matriz error.

$U_{(IxP)} = [u_{ip}]$ matriz de objeto binario que define una partición de los objetos en P grupos, donde $u_{ip} = 1$ si el objeto i pertenece al p-esimo grupo; $u_{ip} = 0$, de lo contrario. La matriz U es estocástica de filas, es decir, no tiene elementos negativos que sumen uno por fila y, por lo tanto, solo tiene un elemento distinto de cero por fila.

$V_{(JxQ)} = [v_{jq}] = [\boldsymbol{v_q}]$ matriz binaria que define una partición de variables en Q grupos, donde $v_{jq} = 1$ si la j-ésima variable pertenece al q-ésimo grupo, $v_{jq} = 0$ en caso contrario. La matriz V tiene solo un elemento distinto de cero por fila.

$\bar{X}_{(PxJ)}$ matriz de centroide de objeto $[\bar{x}_1, \dots, \bar{x}_j, \dots, \bar{x}_p]'$, donde $\bar{x}_p$ representa el centroide en el espacio de las variables observadas. En el caso de una estimación por mínimos cuadrados, la matriz de centroide tiene la forma $\bar{X} = (U'U)^{-1}U'X$ y $\bar{x}_p = (u_p'u_p)u_p'X$ donde u_p es la p-esima columna de U.

$\bar{Y}_{(PxQ)}$ matriz de centroide de objeto $[\bar{y}_1, \dots, \bar{y}_j, \dots, \bar{y}_p]'$, donde $\bar{y}_p$ representa el centroide en el espacio reducido. En el caso de una estimación por mínimos cuadrados, la matriz de centroide tiene la forma $\bar{Y} = (U'U)^{-1}U'XA$; y $\bar{y}_p = (u_p'u_p)u_p'XA$ donde u_p es la p-esima columna de U.

$A_{(JxQ)} = [a_{jq}] = [\boldsymbol{a_q}]$ matriz de los coeficientes de la combinación lineal con $\sum_{j=1}^{J}(a_{jq}a_{jr})^2 = 0$, para cualquier q y r (q≠r) $\sum_{j=1}^{J} a_{jq}^2 = 1$. La matriz A es ortonormal en columnas.

$C_{(JxQ)} = [c_{jq}] = [\boldsymbol{c_q}]$ matriz ortonormal en columna.

$Y_{(IxQ)} = [Y_{iq} = \sum_{j=1}^{J} a_{jq}\, X_{ij}]$ matriz de puntuación de componentes, donde y_{iq} es el valor del i-ésimo objeto para el q-ésimo componente y_q sintetizar la información común de un subconjunto de variables.

El modelo asociado a la agrupación en conglomerados y el análisis de componentes principales disjuntos se puede escribir formalmente de la siguiente manera

$$X = U\bar{Y}A' + E \tag{1}$$

Donde $\bar{Y}$ es una matriz de orden (PxQ) de centroides en el espacio reducido, la matriz U es estocástica binaria y por fila, es decir:

$$u_{ip} \in \{0,1\},\ (i=1,2,\ldots P\ ;\ p=1,2,\ldots,P); \tag{2}$$

$$\sum_{p=1}^{P} u_{ip} = 1,\ (i=1,\ 2,\ldots,I); \tag{3}$$

y A, con rango (A) = Q ≤ J, satisface las restricciones:

$$\sum_{j=1}^{J} a_{jq} = 1\ \ q = 1,..,Q; \tag{4}$$

$$\sum_{J=}^{J}(a_{jq}a_{jr})^2 = 0\ \ q=1,\ldots,Q\text{-}1\ ;\ \ r=q+1,..,Q \tag{5}$$

y por tanto, es ortonormal.

El modelo (1) especifica una partición de los objetos a través de la matriz de pertenencia U y la matriz de centroide $\bar{Y}$, y simultáneamente una reducción de dimensionalidad a través de la matriz de carga de componentes A, que permite la partición de las variables

en clases, cada una resumida por una combinación lineal ortonormal con restricciones (5).

Si se quiere usar todas las variables observadas, el CDPCA tiene que satisfacer la siguiente restricción adicional $\sum_{q=1}^{Q} a_{jq} > 0, \;\; q = 1, \dots Q$

4.3.-MINIMIZACIÓN EN CDPCA

A partir del modelo (1), los estimadores de mínimos cuadrados del CDPCA son las soluciones óptimas del siguiente problema cuadrático [PC1] con respecto a la incógnita A, U y $\bar{Y}$

$$F(U, \bar{Y}, A) = \left\| X - U\bar{Y}A' \right\|^2 = \sum_{i=1}^{I} \sum_{j=1}^{j} \left(x_{ij} - \sum_{p=1}^{P} \sum_{q=1}^{Q} u_{ip}\, \bar{y}_{pq} a_{jq} \right)^2 \rightarrow min_{A,U,\bar{Y}} \quad (7)$$

sujeto a:

U estocástico binario y de filas, es decir, que satisfaga las ecuaciones. (2) y (3); y A satisface las ecuaciones (4) y (5). ([PC1])

Se puede observar que la siguiente descomposición se cumple

$$\|X\|^2 = \left\| X - U\bar{Y}A' \right\|^2 + \left\| U\bar{Y}A' \right\|^2 \quad (8)$$

donde el primer término del lado derecho de la Ecuación. (8) es la desviación "dentro" reconstruida (por Y = XA) de la partición dada por U de los datos observados X, y es, también, la función objetivo del CDPCA.

La descomposición se puede probar, recordando que $\bar{Y} = \bar{X}A$ mostrando que:

$$\left\| X - U\bar{Y}A' \right\|^2 + \left\| U\bar{Y}A' \right\|^2 = tr\left\{ \left[X - U\bar{X}AA' \right] \left[X - U\bar{X}AA' \right]' \right\} + tr\left\{ \left[U\bar{X}AA' \right] \left[U\bar{X}AA' \right]' \right\}$$

$$= tr\{XX'\} - 2tr\{U\bar{X}AA'X'\} + 2tr\{U'U\bar{X}AA'\bar{X}'\}$$

$$= tr\{XX'\} - 2tr\{U\bar{X}AA'X'\} +$$

$$2tr\left\{U'U(U'U)^{-1}U'XAA'\bar{X}'\right\}$$

$$= tr\{XX'\}$$

De la descomposición de (8), se maximiza el segundo término del lado derecho de (8).

$$\left\|U\bar{Y}A'\right\|^2 \tag{9}$$

que corresponde a la desviación reconstruida (por $Y = XA$) de la clase "entre", de la partición dada por U de X. Se obtiene que:

$$\left\|U\bar{Y}A'\right\|^2 = \left\|U\bar{X}AA'\right\|^2 = tr\left\{\left[U\bar{X}AA'\right]\left[U\bar{X}AA'\right]'\right\} = tr\{[U\bar{X}A][U\bar{X}A]'\} = \|U\bar{X}A\|^2$$

Por lo tanto, el problema [PC1] es equivalente a la maximización de la desviación de la clase "entre" $\|U\bar{X}A\|^2$del espacio reducido, sujeto a las restricciones (2) - (5).

Observación 1. El modelo (1) es el modelo conjunto asociado a las k-medias aplicadas a X y el análisis de componentes principales aplicado a la matriz de centroides.

Las k-medias aplicada a X corresponden al ajuste del modelo

$$X = U\bar{X} + E^{(1)} \tag{10}$$

por

$$\|X - U\bar{X}\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \sum_{p=1}^{P} u_{ip}\,\bar{x}_{pj}\right)^2 \rightarrow min_{U,\bar{X}} \tag{11}$$

sujeto a que: U sea una matriz estocástica binaria y por filas, donde $E^{(1)}$ es la matriz de términos de error asociada a la conglomeración de k-medias.

El análisis de componentes principales aplicado en la matriz de centroide $\bar{X}$ corresponde al ajuste del modelo

$$U\bar{X} = U\bar{Y}A' + E^{(2)} \tag{12}$$

por

$$\left\|U\bar{X} - U\bar{Y}A'\right\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{p=1}^{P} u_{ip}\,\bar{x}_{pj}\sum_{q=1}^{Q}\sum_{p=1}^{P} u_{ip}\bar{y}_{pq}\,a_{jq}\right)^2 \rightarrow min_{A,\bar{Y}} \tag{13}$$

Sujeto a:

$$A'A = I_Q \tag{14}$$

donde $E^{(2)}$ es la matriz de términos de error del PCA aplicada a $U\bar{X}$.

La ecuación (1) es el modelo especificado por el k-medias reducido, sujeto a las restricciones de U binaria, estocástica por filas y ortonormalidad (14) (G. De Soete & Carroll, 1994), es decir:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \sum_{p=1}^{P}\sum_{q=1}^{Q} u_{ip}\,\bar{y}_{pq}a_{jq}\right)^2 \rightarrow min_{A,U,\bar{Y}}$$

sujeto a que:

U sea una matriz binaria y estocástica de filas.

$$A'A = I_Q$$

Observación 2. El modelo factorial de k-medias para la agrupación simultánea y PCA también está vinculado al CDPCA. Vichi & Kiers (2001) señalan que las k-medias se definen matemáticamente como:

$$XAA' = U\bar{X}AA' + \mathrm{E} \tag{15}$$

donde U es la matriz de pertenencia al objeto, $\bar{X}$ es la matriz del centroide y A es la matriz de carga de componentes de rango (A) ≤ J y ortonormal, es decir, $A'A = I_Q$. En el k-medias factorial, el óptimo U, $\bar{X}$ y A se obtienen mediante:

$$\left\|YA' - U\bar{Y}A'\right\|^2 = \|XA - U\bar{X}A\|^2 = \sum_{l=1}^{I}\sum_{q=1}^{Q}\left(y_{iq} - \sum_{p=1}^{P}\sum_{j=}^{J} u_{ip}\bar{x}_{pj}\, a_{jp}\right)^2 \rightarrow min_{U,\bar{Y}} \quad (16)$$

sujeto a que:

U sea una matriz binaria y estocástica de filas;

$$A'A = I_Q$$

La partición de los objetos en k-medias factorial lleva a la siguiente descomposición de la desviación total de la matriz de puntuación de componentes Y,

$$\|XA\|^2 = \|XA - U\bar{X}A\|^2 + \|U\bar{X}A\|^2 \quad (17)$$

donde el primer término en el lado derecho es la desviación intra clase (*within-class deviance*) de la partición dada por U en el espacio reducido - minimizado por k-medias factorial - y el segundo término es la desviación inter clases (*between-class deviance*) de la partición en el espacio reducido.

Por lo tanto de las observaciones 1 y 2 propuestas por Vichi & Saporta (2009) concluyen que el k-medias factorial minimiza la desviación intra clase (*within-class deviance*) $\|Y - U\bar{X}A\|^2$ de Y, producida por la partición de los objetos; mientras que la solución de CDPCA maximiza la desviación inter clases (*between-class deviance*) $\|U\bar{X}A\|^2$ de Y, producida por la partición de los objetos. En CDPCA, Vichi & Saporta (2009) prefieren maximizar la desviación inter clases (*between-class deviance*) del espacio reducido, porque están particularmente interesados en definir factores de varianza máxima que se utilizarán para especificar la clasificación de las variables, y esto está garantizado solo si

la desviación inter clases (*between-class deviance*) se maximiza como en las k-medias reducidas.

4.4.-ESTIMACIÓN POR MÍNIMOS CUADRADOS DE LA AGRUPACIÓN EN CONGLOMERADOS Y DEL PCA DISJUNTO Y ALGORITMO DE MINIMO CUADRADO ALTERNO (ALS)

Los tres pasos básicos del algoritmo se pueden describir de la siguiente manera: (i) actualizar U, dada la estimación actual de A y $\bar{X}$, sujeto a las restricciones estocásticas binarias y por filas en U; (ii) actualizar $\bar{X}$, dados A y U actuales; y finalmente (iii) actualizar A, dados U y $\bar{X}$, sujeto a las restricciones (4) y (5) en A.

La matriz A que satisface las restricciones (4) y (5) se reescribe en el producto de dos matrices A = BV, donde V es la matriz de pertenencia de variable que especifica la partición de variables y, por lo tanto, la parte combinatoria de nuestro PCA restringido, mientras que la matriz B es una matriz diagonal (J × J) que ayuda a especificar las cargas de los componentes y representa la parte continua del problema. La matriz diagonal B tiene la forma:

$$B = \left(\sum_{q=1}^{Q} diag(v_q) diag(c_q)\right), \tag{18}$$

donde la notación diag (a) especifica una matriz diagonal con diagonal igual al vector a y

$$c_q = \left[c_{1q}, \ldots, c_{jq}, \ldots, c_{Jq}\right]'$$

es un vector normalizado dimensional J que se usa para encontrar la q-ésima carga del componente.

La ecuación (18) divide la estimación de la matriz A en dos partes: la matriz de pertenencia V para la partición de las variables y las cargas de los componentes c_q para especificar la matriz diagonal B como se indica en (18).

El problema de estimación de suma de cuadrados (9) que se va a maximizar se puede reescribir:

$$F(B,\bar{X},U,V) = \|U\bar{X}BV\|^2$$

$$F = tr\left(V`B`\bar{X}`U`U\bar{X}BV\right)$$

$$= tr[V'\left(\sum_{q=1}^{Q} diag(\boldsymbol{v_q})diag(\boldsymbol{c_q})\right)`\bar{X}`U`U\bar{X}\left(\sum_{q=1}^{Q} diag(\boldsymbol{v_q})diag(\boldsymbol{c_q})V\right)]$$

$$= \sum_{p=1}^{P}\sum_{q=1}^{Q}\left(\sum_{j=1}^{J} v_{jq}\, c_{jq}\bar{x}_{pj}\right)^2 \sum_{i=1}^{I} u_{ip} \quad (19)$$

Sujeto a:

$$u_{ip} \in \{0,1\},\ (i=1,\ldots I\ ;\ p=1,\ldots,P); \quad (20)$$

$$\sum_{p=1}^{P} u_{ip} = 1,\ \ (i=1,\ldots,I)\ ; \quad (21)$$

$$v_{jq} \in \{0,1\},\ (j=1,\ldots,J\ ;\ q=1,\ldots,Q)\ ; \quad (22)$$

$$\sum_{q=1}^{Q} v_{jq} = 1,\ (j=1,\ldots,J) \quad (23)$$

$$\sum_{J=1}^{J} c_{jq}^2 = 1,\ (q=1,\ldots,Q) \quad (24)$$

$$\sum_{j=1}^{J} c_{jq}\, c_{jr} = 0,\ q=1,\ldots,Q\text{-}1\ ;\ r=q+1,\ldots,Q \quad (25)$$

La agrupación en conglomerados y análisis de componentes principales disjuntos - definido por la maximización de (9) sujeto a las restricciones (2) - (5) se ha reformulado en el problema equivalente a maximizar (19) con respecto a variables binarias u_{ip}, las variables v_{jq} y reales c_{jq} , sujeto a (20) - (25).

La ecuación (19) puede simplificarse de la siguiente manera:

$$F(C,\bar{X},U;V)=\sum_{q=1}^{Q} tr[V^{'}diag(c_q)diag(v_q)\,\bar{X}^{`}U^{`}U\bar{X}\,diag(v_q)diag(c_q)V]$$

$$+\sum_{q=1}^{Q}\sum_{\substack{r=1\\ r\neq q}}^{Q} tr[V^{'}\mathrm{diag}(c_q)\mathrm{diag}(v_q)\,\bar{X}^{`}U^{`}U\bar{X}diag(v_r)diag(c_r)V] \quad (26)$$

ACTUALIZACIÓN DE C Y B

La maximización de (26), cuando se estiman $\hat{\bar{X}},\ \hat{U},\hat{V}$, implica la maximización con respecto a c_q, q=1,…Q . De hecho, para cada columna c_q de C, (q=1,…,Q) es necesario resolver:

$$F\left(c_q,\hat{\bar{X}},\hat{U},\hat{V}\right)=tr[\hat{V}^{'}diag(c_q)diag(\hat{v}_q)\hat{\bar{X}}^{'}\hat{U}^{'}\hat{U}\hat{\bar{X}}diag(\hat{v}_q)diag(c_q)\widehat{V}] \quad (27)$$

Sujeto a:

$$\sum_{j=1}^{J} c_{jq}^{2}=1 \quad (q=1,\ldots,Q); \quad (28)$$

$$\sum_{j=1}^{J} c_{jq}\,c_{jr}=0 \quad (q=1,\ldots,Q\text{-}1)\ ;\ r=q+1\ldots.,\ Q\ ) \quad (29)$$

La actualización de B se da por (18), es decir, $B=\left(\sum_{q=1}^{Q} diag(\hat{v}_q)diag(\hat{c}_q)\right.$.

ACTUALIZACIÓN DE V

La maximización de $F\left(\hat{C},\hat{\bar{X}},\hat{U};V\right)$ con respecto a V, cuando $\hat{C},\hat{\bar{X}}\ y\hat{U}$ son fijos, se obtiene para cada j (j=1,…,J) calculando:

$$v_{jq}=1 \text{ si } F\left(\hat{c}_q,\hat{U},\hat{\bar{X}},[v_{jq}]\right)=max\left\{F\left(\hat{c}_r,\hat{U},\hat{\bar{X}},[v_{jr}=1]\right):r=1,..,Q;(r\neq q)\right\} \quad (30)$$

$v_{jq} = 0$ *en cualquier otro caso*

Cuando v_{jr} se fija igual a 1, c_r se actualiza siguiendo el procedimiento descrito anteriormente, para maximizar (26) con respecto a c_r. Por tanto, la actualización de V lleva a la actualización de las columnas de C.

ACTUALIZACIÓN DE $\bar{X}$

La maximización de (19), con respecto a $\bar{X}$ cuando $\hat{B}, \hat{U}, \hat{V}$ son fijos, es equivalente a la

minimización $\left\| X\hat{B}\hat{V} - \hat{U}\bar{X}\,\hat{B}\hat{V} \right\|^2$ (31)

cómo puede verse en la descomposición (17) estableciendo A = BV. La minimización de (31) atañe a la solución del problema de regresión multivariante.

$$\bar{X} = \left(\hat{U}'\hat{U}\right)^{-1}\hat{U}''X \tag{32}$$

ACTUALIZACIÓN DE U

La maximización de $F(\hat{B}, \hat{\bar{X}}, U, \hat{V})$ con respecto a U, cuando $\hat{B}, \hat{\bar{X}}\, y\, \hat{V}$ se estiman considerando (31), es equivalente a la minimización:

$$\left\| X\hat{B}\hat{V} - U\hat{\bar{X}}\hat{B}\hat{V} \right\|^2 = \sum_{i=1}^{I}\sum_{p=1}^{P}\left\| \hat{V}'\hat{B}x_i - \hat{V}'\hat{B}x_p \right\|^2 u_{ip} \tag{33}$$

que implica la minimización de I sub problemas de agrupamiento (i=1,…, I)

$$\sum_{p=1}^{P}\left\| \hat{V}'\hat{B}x_i - \hat{V}'\hat{B}x_p \right\|^2 u_{ip} \tag{34}$$

Sujeto a:

$$u_{ip} \in \{0,1\},\ (i=1,\ldots I\ ;\ p=1,\ldots,P) \tag{35}$$

$$\sum_{p=1}^{P} u_{ip} = 1\ , \qquad \text{(i=1,…,I)} \tag{36}$$

El problema (34), sujeto a (35) y (36), es un problema de asignación que se resuelve en tiempo lineal fijando:

$$u_{ip} = 1 \text{ Si } \left\|\hat{V}`\hat{B}x_i - \hat{V}`\hat{B}x_p\right\|^2 = min\left\{\left\|\hat{V}`\hat{B}x_i - \hat{V}`\hat{B}x_s\right\|^2 \quad s = 1, \dots, P; s \neq p\right\}, \tag{37}$$

$$u_{ip} = 0\ , en\ cualquier\ otro\ caso$$

Observación 3. supongamos que la matriz $B = I_J$, es decir, B es la matriz de identidad de orden J, por lo que el modelo (1) se expresa:

$$X = U\bar{Y}V^{´} + E \tag{38}$$

lo que implica que todas las variables tienen las mismas cargas iguales a 1. El CDPCA degenera en el doble de k-medias (Vichi, 2000), que especifica una partición, tanto para objetos como para variables, en clases P y clases Q, respectivamente. Por lo tanto, Vichi & Saporta (2009) concluyen que las k-medias dobles son un caso relevante del CDPCA más general. En las k-medias dobles, tanto los objetos como las variables se sintetizan mediante perfiles medios de objetos, que pertenecen a la clase de objeto y perfiles medios para las variables que pertenecen a dicha clase de variables. En el CDPCA hay un tratamiento asimétrico de los dos modos de la matriz de datos. Los objetos se sintetizan mediante perfiles medios de conglomerados, mientras que los componentes se sintetizan mediante combinaciones lineales.

4.4.1.-UN ALGORITMO DE MÍNIMOS CUADRADOS ALTERNOS PARA AGRUPACIÓN EN CONGLOMERADOS Y PCA DISJUNTO

El problema restringido de maximizar (9) o (9'') se puede resolver utilizando un algoritmo de mínimos cuadrados alternados (ALS), que comprende cuatro pasos: actualizar V (asignación de variables) y B (el paso de PCA), actualizar la matriz centroide $\overline{X}$ y finalmente actualizar U (la asignación de objetos).

Inicialización. Los valores iniciales se eligen para U y V. Dichos valores se pueden elegir al azar o de una manera racional (por ejemplo, basándose en la solución de agrupamiento de k-medias aplicada en la matriz X y X') y, en ambos casos, deben satisfacer las restricciones en U y V. Para $\overline{X}$, se utiliza la fórmula (32).

Paso 1. B se actualiza, dada la actual $\hat{\bar{X}}, \hat{U}, \hat{V}$ maximizando $F(B, \hat{U}, \hat{\bar{X}}, \hat{V})$ sobre cada columna c_q de C. El vector c_q es el vector propio asociado al valor propio más grande de la matriz $diag(v_q)\hat{\bar{X}}'\hat{U}'\hat{U}\hat{\bar{X}}diag(v_q)$. Además, también se puede optar por una rotación de esta solución, ya que no afecta a los productos escalares. Para actualizar se utiliza la fórmula B (18).

Paso 2. V se actualiza, dada la estimación actual de $\hat{B}, \hat{\bar{X}}\ y\ \hat{U}$. Este problema se resuelve secuencialmente para las diferentes filas de V tomando:

$$v_{jq} = 1, \text{ Si } F\left(\hat{c}_q, \hat{U}, \hat{\bar{X}}, [v_{jq}]\right) = \max\{F\left(\hat{c}_r, \hat{U}, \hat{\bar{X}}, [v_{jr} = 1]\right) : r = 1, \dots, Q\ ; (r \neq q)\}$$

$$v_{jq} = 0, \qquad en\ cualquier\ otro\ caso$$

Paso 3. U se actualiza, con $\hat{B}$, $\hat{\bar{X}}$ y $\hat{V}\ actuales$. Este problema se resuelve para las diferentes filas de U tomando:

$u_{ip} = 1$ Si $\left\|\hat{V}'\hat{B}x_i - \hat{V}'\hat{B}\hat{\bar{x}}_p\right\|^2 = min\left\{\left\|\hat{V}'\hat{B}x_i - \hat{V}'\hat{B}\hat{\bar{x}}_s\right\|^2 : s = 1, \dots P\ ; (s \neq p)\right\}$

$u_{ip} = 0, en\ cualquier\ otro\ caso$

Paso 4. $\bar{X}$ se actualiza, dado que $\widehat{B}, \widehat{U}$ y $\hat{V}$ por $\bar{X} = \left(\widehat{U}'\widehat{U}\right)^{-1}\widehat{U}'X$

Regla de alto. El valor de la función $F(\hat{B}, \hat{\bar{X}}, \widehat{U}, \hat{V})$ se calcula para los valores actuales de $\hat{B}, \hat{\bar{X}}, \widehat{U}$ y $\hat{V}$ .Cuando dichos valores actualizados han aumentado considerablemente (más que un pequeño valor arbitrario de tolerancia de convergencia), el valor de la función B, $\bar{X}$, U y V se actualizan una vez más, según los pasos 1 a 4. De lo contrario, se considera que el proceso ha convergido.

4.5.-ESTUDIO COMPARATIVO CDPCA vs ANÁLISIS DE CLUSTER NO JERARQUICO

El CDPCA es particularmente apropiado cuando el investigador tiene como objetivo reducir la dimensionalidad de individuos y variables por razones de interpretación. Con este método se obtiene una doble ventaja: primero, se identifica la clasificación de variables y una clasificación de los objetos; segundo, se obtiene una reducción de la dimensionalidad de la matriz de datos mediante el conjunto reducido de centroides para objetos y el conjunto reducido de componentes (combinaciones lineales) de variables.

Otra ventaja importante del CDPCA es la facilidad de interpretación de los componentes, ya que cada uno se caracteriza por un conjunto disjunto de variables. Por lo tanto, con el CDPCA no es posible tener una variable observada que caracterice dos componentes.

Generalmente, los componentes del CDPCA no son ortogonales. Sin embargo, para los autores esto no se debe considerar un problema porque, si dos componentes del CDPCA

están altamente correlacionados, esto significa que solo uno de los dos es necesario, y simplemente, es necesario reducir el número de conglomerados para las variables permitiendo la unión de los dos componentes.

En el caso de que todas las variables observadas estén altamente correlacionadas, el CDPCA debe encontrar una componente solamente y, por lo tanto, no hay clasificación de variables.

El análisis de la agrupación en conglomerados y PCA disjunto presenta características semejantes a los algoritmos no jerárquico, entre las cuales se pueden mencionar las siguientes:

- Se fija el número de conglomerados.
- Es sensible a las particiones iniciales aleatorias, por lo cual se repite el análisis varias veces con diferentes particiones iniciales y reteniendo la mejor solución. Esta estrategia de multi inicio es necesaria para evitar que el algoritmo se detenga en un máximo local del problema.

CAPÍTULO 5: APLICACIÓN DEL MÉTODO CART A UN CONJUNTO DE DATOS REALES

5.1.-INTRODUCCIÓN

El deterioro de la calidad del agua es motivo de gran preocupación debido a razones como el crecimiento desproporcionado de la población humana, la expansión de las actividades industriales y agrícolas o las amenazas provocadas por el cambio climático, entre otros. Esto podría generar importantes perturbaciones en el ciclo hidrológico que implicarían serias reducciones en la disponibilidad de agua para cualquier uso (Rose, Winslow, Read, & Hansen, 2016).

El sistema hidrológico del embalse Alhajuela suministra agua potable a las poblaciones de la ciudad de Panamá a través de la toma de agua ubicada en el embalse, el cual abastece de agua cruda a la potabilizadora Federico Guardia Conte, Chilibre. Además, este sistema hídrico produce el 40% del volumen anual de aguas que usa el Canal.

Los ríos que lo conforman corren en forma paralela, sobre formaciones de rocas ígneas, formando amplias secciones que captan agua hasta del tercer orden de ramificación, descargando sus aguas en el embalse.

Los ríos principales que componen la red hidrográfica del sistema del embalse Alhajuela son: Boquerón, Pequení y Chagres. También, recibe las aguas de afluentes como los ríos: Las Cascadas, La Puente y Salamanca y las quebradas: Ancha, La Tranquila, Benítez y Bonita.

El embalse Alhajuela es parte de la Cuenca Hidrográfica del Canal, uno de los elementos hidrológicos más importante de la República de Panamá. La cantidad y calidad del agua está condicionada por el estado de los bosques y del entorno natural; por tanto, su administración requiere del conocimiento de la "calidad natural del agua" y de la dinámica del agua a través del ciclo hidrológico; lo que define las características propias que la hacen apta o no para su uso.

La Unidad de Calidad de Agua, a través del Programa de Vigilancia y Seguimiento de la Calidad de Agua (PVSCA), en la Cuenca Hidrográfica del Canal de Panamá (CHCP), continúa cumpliendo sus funciones dentro de la División de Agua del Departamento de Ambiente, Agua y Energía.

El desmejoramiento de la calidad del agua es un motivo de preocupación a nivel mundial por el crecimiento desproporcionado de la población humana, por la expansión de la actividad industrial y agrícola y por las amenazas del cambio climático que pueden tener importantes alteraciones en el ciclo hidrológico. Por ello, es importante evaluar las relaciones y la forma como se agrupan de las diversas características fisicoquímicas y biológicas del agua del embalse Alhajuela. En este sentido, los métodos fisicoquímicos proporcionan información inmediata del estado del agua (Wolska, Sagajdakow, Kuczynska, & Namiesnik, 2007).

Este trabajo tiene como objetivo investigar los patrones de contaminación fecal, en función de factores físico-químicos y biológicos utilizando el algoritmo de árbol de clasificación CART.

5.2.-METODOLOGÍA

Área de estudio

El embalse Alhajuela, de uso multiple, es creado mediante la construcción del complejo hidrotécnico Madden en 1935, interrumpiendo el cauce natural del río Chagres en su tramo medio. Comprende un área de drenaje de 1.026 km^2, a su nivel de aguas máximas de operación (NAMO = 76,80 m PLD), la superficie del espejo de agua es de 50 km^2. El embalse Alhajuela registra un caudal promedio anual de entrada de 2 310 millones de

metros cúbicos (MMC) y posee un área de aportes de sedimentos de 976 km^2. Su capacidad de almacenamiento de agua a enero de 2008 es de 675,7 MMC (ACP, 2010).

Sitios de muestreo

Las estaciones con sus respectivos códigos son: Boquerón-Pequení (BOP), Estrecho Reporte (ERP), Punta del Ñopo (PNP), Chagres-Alhajuela (DCH) y Toma de Agua IDAAN (TAG). Los datos de los años 2008 al 2015, fueron colectados conforme a los protocolos de la Unidad de Calidad de Agua de la Autoridad del Canal de Panamá (ACP).

Muestreo

En cada una de las estaciones de calidad de agua en el embalse Alhajuela se colectaron muestras de agua a 0.5 m de la superficie con una periodicidad mensual. Todos los muestreos fueron desarrollados en horario diurno abarcando las estaciones seca y lluviosa.

Para el análisis se tomarán los datos resultantes de los parámetros fisicoquímicos y biológicos desarrollados por la Unidad de Calidad de Agua en sus cinco estaciones de muestreo.

Punto de muestreo

Cada punto de muestreo constituye una muestra, que representa el mes, el nivel del muestreo superficie, la estación del muestreo y el año. Por ejemplo, el punto de muestreo febSTAG08 representa al mes de febrero, nivel de muestreo superficie, estación de muestreo toma de agua del IDAAN y el año 2008.

En cada sitio se realizaron mediciones *in situ* y se colectaron muestras para realizar determinaciones analíticas en el laboratorio de la Unidad de Calidad de Agua, División de Agua del Canal de Panamá. Los instrumentos de medición utilizados fueron previamente calibrados en laboratorio. Se utilizaron envases apropiados (vidrio, plástico de polipropileno, etc.). Las muestras colectadas fueron colocadas en neveras con hielo hasta su transporte al laboratorio. El acceso a los sitios se realizó vía acuática.

Procedimiento de análisis físico-químico

Las metodologías de análisis corresponden a las descritas en el Standar Method for the Examination of Water and Wastewater (APHA, AWWA, WEF, 21°Edicion).

Un total de 16 parámetros o variables de calidad de agua a nivel de superficie, los cuales incluyen: temperatura (°C), pH, t, turbidez (NTU), nitratos (mg/l), ortofosfato (mg/l), clorofila (µg/l), oxígeno disuelto (% y mg/l), fueron los medidos en campo y laboratorio.

En la tabla 7 se presentan los métodos analíticos empleados para la determinación de los parámetros de la calidad del agua.

Tabla 7. Métodos analíticos empleados para la determinación de parámetros de calidad de agua

Tipo de medición	No.	Características (parámetro)	Método	Unidades de medida	Límite detección	Decimales a reportar
In situ	1	Temperatura	SM 2550-A	°C	0,1°C	1
	2	pH	SM 4500-H+B	pHunits	0,1 pHunits	2
	3	Conductividad	SM 2510	microh/cm	1microh/cm	0
	4	Oxígeno disuelto	SM 4500-O C	mg/l	0,1mg/l	2
Sólidos	5	Sólidos totales disueltos	SM 2540-C	mg/l	10mg/l	0
	6	Turbiedad	SM 2130-B	NTU	0.05 NTU	1
	7	*E. coli*	SM 9223-B (Colilert)	NMP/100ml	n/a	0
	8	Clorofila a	SM 10200-H, modificación USEPA 445.0	µg/l	0,05 µg/l	1
Aniones mayoritarios	9	Alcalinidad Total (OH-, HCO3, CO3)	SM 2320-B	mg/l	1mg/l	0
	10	Cloruros	SM 4500-Cl - D	mg/l	1,0 mg/l	1
	11	Sulfatos	SM 4500-E-SO4.	mg/l	1mg/l	1
	12	Dureza total	SM 2340-B (calculada)	mg/l	n/a	1
Cationes mayoritarios	13	K+	SM 3111-B	mg/l	0.001mg/l	2
Nutrientes	14	N-Nitratos	SM 4500-NO3- E	mg/l	0,01 mg/l	3
	15	P-Fosfatos	SM 4500-P E	mg/l	0,02 mg/l	3

5.3.-APLICACIÓN DEL MODELO CART PARA EVALUAR LA CONTAMINACIÓN DE LA CALIDAD DE AGUA

La calidad biológica de las aguas es un modo de definir la riqueza biológica y el valor ambiental de las comunidades de seres vivos asociados al ecosistema de un curso fluvial, o de un tramo concreto de él (Martínez, Fonseca, Ortega, & García-Luján, 2009).

Todos los seres vivos necesitan agua para su supervivencia con una adecuada calidad. Entre los contaminantes naturales del agua se encuentran virus, bacterias y otras formas de vida, especies minerales disueltos, productos orgánicos solubles y sólidos orgánicos e inorgánicos suspendidos. La concentración de estos contaminantes naturales puede incrementarse o aún ser suplida por otros materiales producto de la tecnología industrial o agrícola. Con el fin de asegurar y preservar la calidad del agua en los sistemas de abastecimiento hasta la entrega al consumidor, la misma debe ser sometida a tratamientos de potabilización. Un alto riesgo de contaminación la presenta el agua potable que contenga material fecal (Mushi et al., 2012; Vaccari, Collivignarelli, Tharnpoophasiam, & Vitali, 2009).

El control de la calidad sanitaria de los recursos del ambiente puede llevarse a cabo mediante la enumeración de bacterias indicadoras de contaminación fecal. Estas bacterias pueden ser utilizadas para valorar la calidad de los alimentos, sedimentos y aguas destinadas al consumo humano, la agricultura, la industria y la recreación. No existe un indicador universal, por lo que se debe seleccionar el más apropiado para la situación específica en estudio (Bachoon, Markand, Otero, Perry, & Ramsubaugh, 2010; Luby et al., 2008).

Los indicadores de contaminación fecal más utilizados son los coliformes totales y termotolerantes, *Escherichia coli* y enterococos (Rossen et al., 2008).

Escherichia coli es miembro de la familia Enterobacteriaceae. Es una bacteria gram negativa, anaerobia facultativa que forma parte del microbiota normal del intestino del ser humano y los animales homeotermos, siendo la más abundante de las bacterias anaerobias facultativas intestinales. Se excreta diariamente con las heces (entre 108-109 Unidades Formadoras de Colonias (UFC) g-1 de heces) y por sus características es uno de los indicadores de contaminación fecal más utilizados (Larrea et al., 2009).

E.coli tiene como hábitat natural el tracto intestinal de hombre y animales. Es el indicador clásico de la posible presencia de patógenos entéricos en el agua, en los moluscos, en los productos lácteos y en otros alimentos crudos. Una práctica común es utilizar las pruebas para coliformes, que incluyen *E. coli*, en los ensayos de "screening" o preliminares. Si de estas pruebas iniciales se deduce la posibilidad de contaminación fecal, los coliformes y otras Enterobacteriaceae se someten a posteriores estudios para determinar si entre ellos está presente *E. coli*.

El término habitual "coliformes" comprende *E. coli* y diversas especies pertenecientes a otros géneros de la familia Enterobacteriaceae fermentadores de la lactosa con producción de gas a 31-37ºC. Pueden ser o no fecales.

Los "coliformes fecales" incluyen un grupo de microorganismos seleccionados por incubación de los inóculos procedentes de un caldo de enriquecimiento de coliformes a temperaturas superiores a las normales (44-45ºC) dependiendo del método. Tales cultivos

de enriquecimiento contienen por lo general un alto porcentaje de *E. coli* y son, por ello, muy indicativos de una probable contaminación de origen fecal del alimento.

La principal bacteria es la *Escherichia coli* cuya presencia en los alimentos indica una posible contaminación fecal por lo cual el consumidor en caso de ingerir ese alimento podría estar expuesto a bacterias entéricas (Haller, Pote, Loizeau, & Wildi, 2009). *E. coli* reúne las condiciones del indicador ideal de contaminación fecal: está presente universalmente en las heces y en las aguas residuales, no puede crecer en las aguas naturales y es fácilmente detectable por métodos rápidos. Muchas cepas de *E. coli* son causantes de enfermedad en humanos y animales. La detección de contaminación fecal se debe realizar de forma rápida y precisa para proteger la salud humana y el medio ambiente (Paruch & Mæhlum, 2012).

E. coli es el indicador por excelencia de la contaminación fecal en el agua, como consecuencia, resulta importante conocer cuáles son aquellos predictores que inciden en la presencia de esta bacteria para evaluar la calidad del agua.

Los árboles de decisión son modelos con estructuras arbóreas. El recorte de las ramas puede resolver el problema del sobreentrenamiento. En general, los árboles de decisión más grandes son más expresivos y pueden tener más poder de predicción, pero cuanto más pequeño es un árbol de decisión, más fuerte es su simplicidad. La construcción de CART se basó en el índice de Gini, con la mejor variable independiente elegida para el corte binario en cada rama. Por lo tanto, es probable que cada variable independiente (campo) se utilice repetidamente en diferentes nodos. El objetivo es obtener dos subconjuntos lo más homogéneos como sea posible en cada partición.

Los resultados que presentamos en este apartado se basan en la identificación de variables que predicen el indicador por excelencia de contaminación fecal denominado E.coli. La variable E.coli se utilizó dicotomizada considerando como punto de corte la mediana igual a diez. Para valores superiores a diez se consideró que hay contaminación fecal, y para los valores inferiores que no. Como variables independientes o predictoras las siguientes variables físico-químicas: alcalinidad, clorofila, cloruro, conductividad, dureza, potasio, nitrato, oxígeno, ortofosfatos, ph, sulfato, sólidos totales disueltos, temperatura y turbiedad.

Para ello hemos utilizado en cada una de las estaciones de calidad de agua en el embalse Alhajuela una muestra de agua a 0.5 metros de la superficie, con una periodicidad mensual (desde el año **2008 hasta el 2015**). La muestra está compuesta por 456 puntos de muestreo.

La Tabla 8 describe las variables seleccionadas en el árbol, el tipo de variable y el valor mínimo y máximo.

Tabla 8. Descripción de las variables incluidas en el árbol

Variables	Descripción de la variable	Tipo de variables	Mínimo- Máximo
ALCT/Alcalinidad	Mide la capacidad del agua para neutralizar ácidos.	Continua	23-92
CHLA/Clorofila a	Pigmento por el cual las plantas realizan fotosíntesis, medida de la biomasa de fitoplancton.	Continua	0.00-52.80
CL/Cloruro	Ión que resulta de la combinación del cloro con un metal.	Continua	3.50-11.40
COND/conductividad	Medida de sales disueltas en una solución.	Continua	69 -210
DUREZA	Suma de la dureza del calcio y magnesio.	Continua	14.80-99.30
K/Potasio	Potasio de origen volcánico.	Continua	0.18-2.13
NNO3/Nitrato	Nutrientes	Continua	0.00-1.46
OD/Oxígeno disuelto	Medida de la concentración de oxígeno gaseoso en el agua.	Continua	3.87-10.43
PPO4/Ortofosfatos	Nutrientes	Continua	0.00-0.10
PH	Indicador de la acidez o basicidad del agua.	Continua	5.88-8.70
SO4/Sulfato	Ión de la sal de ácido sulfúrico	Continua	0.25-14.80
STD/ Solido totales disueltos	Medida de sales disueltas luego de removidos los sólidos suspendidos.	Continua	48-153
TEMP/Temperatura	Medida del contenido térmico del agua.	Continua	23,00-31.00
TURB/Turbidez	Apariencia del agua provocada por partículas en suspensión.	Continua	0.00-382.00

En primer lugar, construimos los conjuntos de entrenamiento y validación del modelo. Para ello utilizamos los 350 primeros casos (puntos de muestreo) para el entrenamiento y el resto para la validación.

En la muestra de validación el 76,42% de los casos se han clasificado bien.

Resultados:

	NCF	CF
NCF	64	18
CF	7	17

NCF: No hay contaminación fecal; CF: Si hay contaminación fecal

Posteriormente, analizamos todos los casos de la muestra disponible para construir el clasificador:

Resultados:

	NCF	CF
NCF	291	52
CF	15	98

La visualización del árbol corresponde a la ilustración de la Figura 9.

Los nodos terminales, hojas del árbol, muestran la clase predominante (contaminación fecal NO/SI), la proporción de puntos de muestreo con contaminación fecal = CF y no contaminación fecal = NCF dentro del nodo.

En el nodo raíz partimos de los 456 puntos de muestreo, de los cuales el 33% tienen probabilidad de no tener contaminación fecal. En el segundo nodo se utiliza la variable

turbidez para segmentar con el punto corte 8,6. Este nodo lo conforman 377 puntos de muestreo de los cuales 23% tienen probabilidad de no tener contaminación fecal.

El análisis de segmentación subdivide a la muestra en 12 segmentos terminales descritos en la figura 9. Las variables que resultaron significativas son la turbidez, oxígeno, PH, dureza y clorofila.

En la tabla siguiente se presenta el perfil de cada nodo terminal:

Tabla 9. Perfil de cada nodo terminal

Nodo	Perfil del nodo
Nodo 4	Valores de turbidez<8,6; turbidez <3,2; y <1,7
Nodo 6	Valores de turbidez>=1,7 y oxígeno>=7,3
Nodo 8	Valores de turbidez>=1,7; oxígeno<7,3 y dureza >=47
Nodo 9	Valores de turbidez>=1,7; oxígeno<7,3 y dureza <47
Nodo 13	Valores de turbidez>=3,2; PH>=7,4 y oxígeno<8,5
Nodo 14	Valores de turbidez>=3,2; PH>=7,4 y oxígeno>=8,5
Nodo 16	Valores de turbidez>=3,2; PH <7,4 y clorofila<2,3
Nodo 17	Valores de turbidez>=3,2; PH<7,4 ; clorofila >=2,3 y <4,52
Nodo 18	Valores de turbidez>=3,2: PH <7,4 ; clorofila>=2,3 y >= 4,52
Nodo 21	Valores de turbidez>=8,6 ; clorofila<0,6
Nodo 22	Valores de turbidez>=8,6; clorofila>=0,6 y dureza<39
Nodo 23	Valores de turbidez>=8,6; clorofila<0,6 y dureza>=39

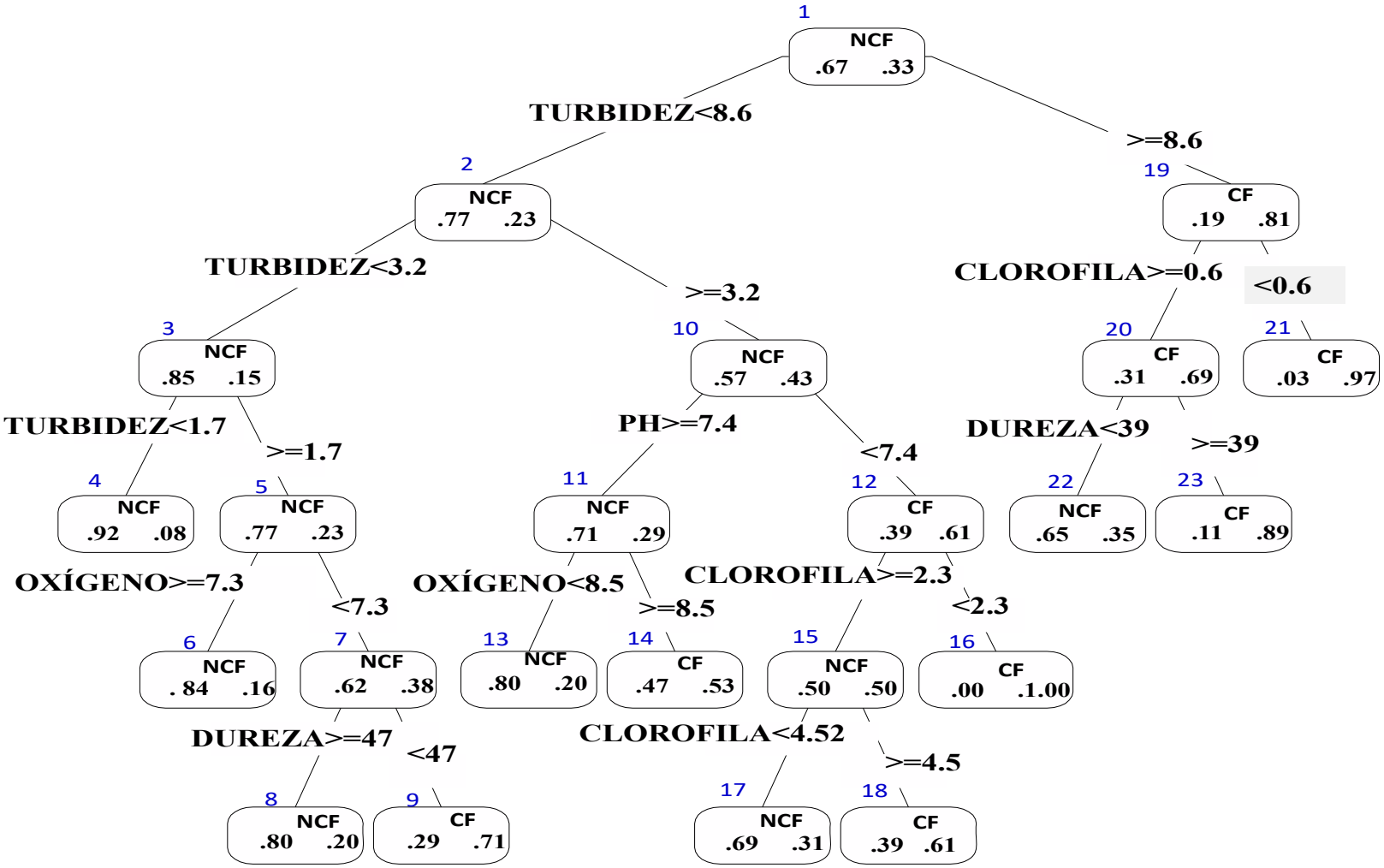

Figura 9. Visualización del árbol.

Nota: NCF: No contaminación fecal; CF: Contaminación fecal.

CONCLUSIONES

1.- Los métodos cluster se recogen en la literatura especializada de Data Mining y Machine Learning, como métodos no supervisados (simétricos utilizando la terminología estadística) basados en algoritmos cuyo objetivo directo es agrupar las unidades taxonómicas a clasificar, bajo diferentes criterios; sin embargo, en la era actual en la que cada día es más frecuente el manejo de grandes masas de datos y de un elevado número de variables, estos métodos son insuficientes.

2.- El Algoritmo K-means de Forgy, propuesto en 1965, y modificado por Mcqueen en 1967, sigue siendo el método más utilizado, probablemente porque es fácil de implementar y de interpretar sus resultados, es rápido y eficiente en términos de coste computacional y el procedimiento está basado en cálculos medias; sin embargo, es sensible a la existencia de outliers, converge pero encuentra mínimos locales de la función de coste, es sensible a la inicialización, no existe una solución única para un número K de clusters, todos las unidades tienen que pertenecer a un cluster determinado, no puede determinar clusters no convexos o clusters con formas inusuales y produce clusters de tamaño similar, aunque en la estructura de los datos existan agrupaciones de diferentes tamaños.

3.- Las limitaciones del K-means han dado lugar a distintas alternativas que tratan de mejorar diferentes aspectos: Fuzzy C-means (Dunn, 1974) donde no se considera la pertenencia de forma dicotómica sino en términos probabilísticos; K-medoids (PAM) (Kaufman & Rousseeuw, 1990) más robusto al ruido y a valores grandes de los datos, donde cada *cluster* está representado por una observación presente en el *cluster* (*medoid*), mientras que en *K-means* cada *cluster* está representado por su centroide.

Puede trabajar con variables categóricas y con cualquier medida de distancia, aunque es computacionalmente más costoso y puede converger a mínimos locales.

4.- Las soluciones para datos de grandes dimensiones, generalizan el Algoritmo PAM, generando Medoids óptimos Globales con el algoritmo CLARA (Kaufman & Rousseeuw, 1990) y parten de submuestras aleatorias iniciales diferentes en varias iteracciones, en la propuesta CLARANS (Ng & Han, 2002).

5.- Cuando se dispone de muchas variables es posible trabajar sobre la HJ-bigeometría proyectada en un subespacio y en ese espacio de baja dimensión crear clusters con diferentes algoritmos. Trabajar de esta forma presenta la ventaja de que no solo es posible descubrir patrones, sino que se pueden identificar las variables responsables de esas agrupaciones y las combinaciones lineales de variables que proporcionan máxima discriminación entre los clusters.

6.- El cluster HJ-BIPLOT es una representación gráfica multivariante donde los clusters se definen, maximizando la Inercia Entre clusters y minimizando la Inercia Dentro de cada clúster, mientras que el Clusplot es una representación gráfica donde los cluster se generan utilizando el algoritmo PAM, el cual considera la mínima suma de las disimilitudes entre puntos de un cluster, en lugar de la disimilitud promedio que es en lo que se basa el algoritmo K-means.

7.- Es posible crear algoritmos que integran la búsqueda de direcciones de máxima inercia, respetando a su vez la estructura de los clusters, y también es posible representar los clusters y sus centroides sobre subespaciones generados por componentes principales disjuntas.

8.- Es posible crear un espacio multidimensional sobre estructuras sparse y proyectar en diferentes subespacios centroides y cluster.

9.- Todos estos métodos de clusters, admiten alternativas supervisadas a partir de árboles de clasificación y regresión, basadas en particiones binarias recursivas, o en algoritmos divisivos basados en criterios de entropía (Modelo CART).

10.- Mientras que en el modelo CART los clusters solo son compatibles con estructuras dicotómicas, en el CLUSTER HJ-BIPLOT los clusters pueden tener cualquier estructura.

11.- En el caso de tener una estructura respuesta multivariante y muchas variables explicativas, se pueden generar clusters que capturan la estructura de la respuesta a partir de clases latentes y seleccionan las variables que intervienen en la creación de los clusters, a partir de coeficientes de predictividad que son la base del algoritmo TAID (Castro, 2005), el cual genera árboles ternarios sobre los que se definen los clusters.

BIBLIOGRAFÍA

Abonyi, J., & Feil, B. (2007). *Cluster Analysis for Data Mining and System Identification*. Springer Science & Business Media.

Aboubi, Y., Drias, H., & Kamel, N. (2016). BAT-CLARA: BAT-inspired algorithm for Clustering LARge Applications. *IFAC-PapersOnLine*, *49*(12), 243–248. https://doi.org/10.1016/j.ifacol.2016.07.607

ACP. (2010a). Informe de Calidad de Agua 2008-2009. *Departamento de Ambiente, Agua y Energía. División de Agua. Unidad de calidad de Agua. Panamá.*

ACP. (2010b). Informe de calidad de agua de la cuenca del canal. *Panama: Autoridad del Canal de Panama*.

Anderson, R. J., & Landis, J. . (1982). Catanova for multidimensional contingency tables: Nominal-scale response. *Communications in Statistics-Theory and Methods*, *9*(11), 257–270. https://doi.org/10.1080/03610928008827952

Avila, C. A. (1996). Una Alternativa al Análisis de Segmentación Basada en el Análisis de Hipótesis de Independencia Condicionada. *[Tesis Doctoral]. Universidad de Salamanca.*

Bachoon, D., Markand, S., Otero, E., Perry, G., & Ramsubaugh, A. (2010). Assessment of non-point sources of fecal pollution in coastal waters of Puerto Rico and Trinidad. *Marine Pollution Bulletin*, *60*(7), 1117–1121. https://doi.org/10.1016/j.marpolbul.2010.04.020

Beh, E. J. (2008). Simple correspondence analysis of Nominal-Ordinal contingency tables. *Journal of Applied Mathematics & Decision Sciences*, *1*, 17–34.

Benzécri, J. P. (1973). *L'Analyse des Données. Tome I: Taxinomie.* Dunod. Paris.

Bezdek, J. C. (1974). Cluster validity with fuzzy sets. *J. Cybernet.*, *3*, 58–73.

Borah, S., & Ghose, M. K. (2009). Performance analysis of AIM-K-Means and K-Means in quality cluster generation. *J. Computer.*, *1*, 175–178.

Breiman, L., Friedman, J., Olsen, R., & Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Bucheli, Herbert & Thompson Wayne (2009). *Statistics and Machine Learning at Scale New Technologies Apply Machine Learning to Big Data*. SAS

Cardot, H., Cénac, P., & Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics and Data Analysis*, *56*, 1434–1449. https://doi.org/10.1016/j.csda.2011.11.019

Cardot, H., Cénac, P., & Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient approach. *Bernoulli, 19(1)*, 18-43.

Carmichael, J. W., & Sneath, P. H. A. (1969). Taxometric maps. *Systematic Zoology*, *18*, 402–415.

Carrasco, G., Molina, J. L., Patino-Alonso, M. C., Castillo, M. D. C., Vicente-Galindo, M. P., & Galindo-Villardón, M. P. (2019). Water quality evaluation through a multivariate statistical HJ-Biplot approach. *Journal of Hydrology*, *577*, 1–9. https://doi.org/10.1016/j.jhydrol.2019.123993

Castro, C. (2005). Contribuciones a la detección y análisis de variables relevantes en las tablas de contingencia multivariantes. *[Tesis Doctoral]. Universidad de Salamanca.*

Chae, S. S., & Warde, W. D. (2006). Effect of using principal coordinates and principal components on retrieval of clusters. *Computational Statistics and Data Analysis*, *50*(6), 1407–1417. https://doi.org/10.1016/j.csda.2005.01.013

Chen, H., Gnanadesikan, R., & Kettenring, J. R. (1974). Statistical methods for grouping corporations. *Sankhyā: The Indian Journal of Statistics*, *Series B*, 1–28.

Clogg, C. (1988). Latent class models for measuring. In *Latent trait and latent class models* (pp. 173–205). Springer, Boston, MA.

Cox, D. R., & Wermuth, N. (2014). *Multivariate Dependencies. Models, Analysis and Interpretation*. London.: CRC Press.

Cuesta-Albertos, J. A., Gordaliza, A., & Matran, C. (1997a). Rimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, *25*(2), 553–576.

Cuesta-Albertos, J. A., Gordaliza, A., & Matran, C. (1997b). Trimmed k-Means: An Attempt to Robustify Quantizers. *The Annals of Statistics*, *25*(2), 553–576. https://doi.org/10.1214/aos/1031833664

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistics Society*, *41*(1), 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In *New Approaches in Classification and Data Analysis* (pp. 212–219). Springer, Heidelberg. https://doi.org/10.1007/978-3-642-51175-2_24

Der, G., & Everitt, B. S. (2005). *Statistical Analysis of Medical Data using SAS*. CRC Press.

DeSarbo, W. S., Jedidi, K., Cool, K., & Schendel, D. (1990). Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, *2*(2), 129–146. https://doi.org/10.1007/BF00436033

Dorado, A. (1998). Métodos de Búsqueda de Variables Relevantes en Análisis de

Segmentación :Aportaciones desde una Perspectiva Multivariante. *[Tesis Doctoral]. Universidad de Salamanca.*

Dunham, M. (2002). Data Mining: Introductory and Advanced Topics. *1st Edn., Prentice Hall USA*.

Dunn, J. C. (1973). A fuzzy relative of ISODATA process and its use in detecting compact, well separated clusters. *Journal of Cibernetics*, *3*, 95–104. https://doi.org/10.1080/01969727308546046

Eddy, W. F. (1977). A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software (TOMS)*, *3*(4), 398–403. https://doi.org/10.1145/355759.355766

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classification. *Biometrics, 21(3)*, 768.

Frank, R., Massy, W., & Wind, Y. (1972). *Market Segmentation*. New Jersey: Prentice-Hall.

Gabriel, K., & Odoroff, C. (1990). Biplots in biomedical research. *Statistics in Medicine*, *9*(5), 469–485. https://doi.org/10.1002/sim.4780090502

Gabriel, K. R. (1971). The biplot-grafhic display of matrices with application to principal component analysis. *Biometrika*, 58, 453-467. https://doi.org/10.1093/BIOMET/58.3.453

Galindo, M. P. (1986). Una alternativa de representación simultánea:HJ-Biplot. *Qüestiió:Quaderns d'estadística i Investigación Operativa*, *10*(1), 13–23.

Glover, F., & Laguna, M. (1997). T*abu Search*, Kluwer Academic Publishers.

Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications III: Approximate sampling theory. *Journal of the American Statistical Association*, *268*, 732–764. https://doi.org/10.2307/2281536

Gordaliza, A. (1991a). Best approximations to random variables based on trimming procedures. *J.Approx. Theory*, *64*, 162–180.

Gordaliza, A. (1991b). On the breakdown point of multivariate location estimators based on trimming procedures. *Statist.Probab.Lett.*, *11*, 387–394.

Gower, J. C., & Harding, S. A. (1988). Nonlinear biplots. *Biometrika*, *75*(3), 445–455. https://doi.org/10.1093/biomet/75.3.445

Güngör, Z., & Ünler, A. (2007). K-harmonic means data clustering with simulated annealing heuristic. *Applied Mathematics and Computation*, *184*(2), 199–209. https://doi.org/10.1016/j.amc.2006.05.166

Gupta, T., & Panda, S. (2018). A Comparison of K Means Clustering Algorithm a nd Clara Clustering Algorithm o n Iris Dataset. *International Journal of Engineering & Technology*, *7*(4), 4766–4768. https://doi.org/10.14419/ijet. v7i4.21472

Hair, J. F., Anderson, R. E., Tathaman, R. I., & Black, W. (1999). *Análisis Multivariante* (5 Edición). Madrid: Editorial Prentice Hall.

Haller, L., Pote, J., Loizeau, J., & Wildi, W. (2009). Distribution and survival of faecal indicator bacteria in the sediments of the Bay of Vidy, Lake Geneva, Switzerland. *Ecological Indicators*, *9*(3), 540–547.

Hammerly, G., & Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. In: Proceedings of the eleventh international conference on Information

and Knowledge Management (pp. 600–607).

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Elsevier.

Hartigan, J. (1975). Clustering Algorithms. *John Wiley &Sons, New York*.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136 A KMeans Clustering Algorithm. *Applied Statistics*, *28*(1), 100–108.

Holsheimer, M., & Siebes, A. (1994). *Data mining: The search for knowledge in databases.* In: *CWI Technical Report CS-R9406*. Amsterdam, The Netherlands.

Höppner, F., Klawonn, F., Rudolf, K., & Runkler, T. (1999). *Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition*. John Wiley & Sons.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques. Regression, classification and manifold learning*. New York: Springer.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. *Prentice Hall Inc.*, 320.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review *ACM. Comput.Surveys.*, *31*, 264–323.

James, G., Witten, D., Hastie, T, Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, *29*(2), 127–199.

https://doi.org/10.2307/2986296

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In *Statistical Data Analysis Based on the L1 Norm and Related Methods* (pp. 405–416). North-Holland.

Kaufman, L., & Rousseeuw, P. J. (1990a). *Finding Groups in Data. An Introduction to Cluster Analysis*. Jonh Wiley & Sons,New York.

Kaufman, L., & Rousseeuw, P. J. (1990b). Finding Groups in Data. An Introduction to Cluster Analysis. *Jonh Wiley & Sons,New York*.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Khan, S. S., & Ahmad, A. (2004). Cluster center initialization algorithm for K-means clustering. *Patt. Recog. Lett.*, *25*, 1293–1302.

Kirkpatrick, S., Gerlatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Klastoria, T. D. (1985). The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Science*, *31*(1), 84–95. https://doi.org/10.1287/mnsc.2019.3415

Klein, W., & Dubes, R. C. (1989). Experiments in projection and clustering by simulated annealing, Pattern Recognition. *Pattern Recognition*, *22*(2), 213–220. https://doi.org/10.1016/0031-3203(89)90067-8

Kondo, Y., Salabian-Barrera, M., & Zamar, R. (2012). A robust and sparse K-means clustering algorithm. *ArXiv :1201.6082v1*, 1–20.

Kottler, P. (1988). *Marketing Management*. New Jersey: Prentice-Hall.

Larrea, J., Rojas, M., Heydrich, M., Romeu, B., Rojas, N., & Lugo, D. (2009). Evaluación de la calidad microbiológica de las aguas del Complejo Turístico Las Terrazas, Pinar del Río (Cuba). *Hig Sanid Ambient*, *9*, 492–504.

Lauro, N. C., & D'Ambra, L. (1984). L'analyse non symétrique des correspondances. *Data Analysis and Informatics*, *3*, 433–446.

Leiva-Valdebenito, S. A., & Torres-Avilés, F. J. (2010). Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados: un estudio comparativo. *Revista Colombiana de Estadística*, *33*(2), 321–339.

Light, R., & Margolin, B. (1963). An analysis of variance for categorical data. *Journal of the American Statistical Association*, *66*(335), 534–544. https://doi.org/10.1080/01621459.1971.10482297

Ligth, R. J., & Margolin, B. H. (1974). An analysis of variance for categorical data II. Small samples comparisons with chi-square and other competitors. *Journal of the American Statistical Association*, *69*, 755–764. https://doi.org/10.2307/2286014

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129–136.

Luby, S., Gupta, S., Sheikh, M., Johnston, R., Ram, P., & Islam, M. (2008). Tubewell water quality and predictors of contamination in three flood-prone areas in Bangladesh. *Journal of Applied Microbiology*, *105*(4), 1002–1008.

MacQueen, J. (1967b). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probabilities* (pp. 281-297). Berkeley, California.

Martínez, A., Fonseca, K., Ortega, J., & García-Luján, C. (2009). Monitoreo de la calidad microbiológica del agua en la cuenca hidrológica del río Nazas, México. *Química Viva*, *8*(1), 35–47.

Massart, D. L., Plastria, F., & Kaufman, L. (1983). Non-HierarchicalClustering with Masloc. *Pattern Recognition.*, *16*(5), 507–516.

McGarigal, K., Cushman, S., & Stanford, S. (2000). Multivariate Statistics for Wildlife and Ecology Research. *Springer Verlag*.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204. https://doi.org/10.1007/BF01897163

Morgan, J., & Sonquist, J. (1963). Problems in the Analysis of Survey Data and A Proposal. *Journal of the American Statistical Association*, *58*(302), 415–434. https://doi.org/10.1080/01621459.1963.10500855

Mushi, D., Byamukama, D., Kirschner, A., Mach, R., Brunner, K., & Farnleitner, A. (2012). Sanitary inspection of wells using risk-of-contamination scoring indicates a high predictive ability for bacterial faecal pollution in the peri-urban tropical lowlands of Dar es Salaam, Tanzania. *J Water Health*, *10*(2), 236–243. https://doi.org/10.2166/wh.2012.117

Ng, R., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th international conference on very large databases,* (pp. 144–155). Santiago, Chile.

Ozdemir, O., & Kaya, A. (2018). K-medoids and fuzzy C-means algorithms for clustering CO2 emissions of turkey and other OECD countries. *Applied Ecology and*

Environmental Research, *16*(3), 2513–2526. https://doi.org/10.15666/aeer/1603_25132526

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Park, H.-S., Lee, J. S., & Jun, C.H. (2006). AK-means like algorithm for K-medoids clustering and its performance. In *Proceedings of ICCIE* (pp. 102-117).

Paruch, A., & Mæhlum, T. (2012). Specific features of Escherichia coli that distinguish it from coliform and thermotolerant coliform bacteria and define it as the most accurate indicator of faecal contamination in the environment. *Ecological Indicators*, *23*, 140–142. https://doi.org/10.1016/j.ecolind.2012.03.026

Perafan-López, J., & Sierra-Pérez, J. (2020). An unsupervised pattern recognition methodology based on factor analysis and a genetic-DBSCAN algorithm to infer operational conditions from strain measurements in structural applications. *Chinese Journal of Aeronautics*. In press.

Pison, G., Rousseeuw, P., & Struyf, A. (1999). Displaying a clustering with Clusplot. *Computational Statistics & Data Analysis.*, *30*(4), 381–392. https://doi.org/10.1016/S0167-9473(98)00102-9

Qin, A. K., & Suganthan, P. N. (2005). Initialization insensitive LVQ algorithm based on cost-function adaptation,Pattern Recognit., *38*, 773–776.

Quinn, G., & Keough, M. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge university press.

Rakhlin, A., & Caponnetto, A. (2007). Stability of K-Means clustering. *Adv. Neural Infor. Process. Syst.*, *12*, 216–222.

Ramírez, G. (1995). Contribuciones al análisis de segmentación. *[Tesis Doctoral]. Universidad de Salamanca.*

Rao, M. R. (1971). Cluster Analysis and Mathematical Programming. *The American Statistical Association*, *66*, 622–626.

Rose, K., Winslow, L., Read, J., & Hansen, G. (2016). Climate-induced warming of lakes can be either amplified or suppressed by trends in water clarity. *Limnology and Oceanography Letters 1(1),* 44–53. https://doi.org/10.1002/lol2.10027

Rossen, A., Rodríguez, M., Ruibal, A., Fortunato, M. Bustamante, A., Ruiz, M., Angelaccio, C., & Korol, S. (2008). Indicadores bacterianos de contaminación fecal en el embalse San Roque (Córdoba, Argentina). *Higiene, Sanidad y Ambiente*, *8*, 325–330.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105

Rousseeuw, P. (1985). Representing Data Partitions,. *Proceedings of the Statistical Computing Section of the American Statistical Association.*, 275–280.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rousseeuw, P., & Ruts, I. (1997). The Bagplot: a bivariate box-and-whiskers plot. *The American Statistician*, *53(4)*, 382-387.

Rousseeuw, P., & Van Driessen, K. (1997). A fast algorithm for the minimum covarience determinant estimator. *Technometrics*, *41*(3), 212–223.

https://doi.org/10.1080/00401706.1999.10485670

Russell Stuart, Norvig, Peter (2010). Artificial Intelligence: A Modern Approach, Global Edition. New Jersey: Prentice Hall.

Schaffer, C. M., & Green, P. E. (1996). An empirical comparison of variable standardization methods in cluster analysis. *Multivariate Behavioral Research*, *31*(2), 149–167. https://doi.org/10.1207/s15327906mbr3102_1

Siciliano, R., & Mola, F. (1997). Ternary Classification Trees: A Factorial Approach.. In *Visualization of categorical data* (pp. 311–323). Academic Press.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *13*, 238–241.

Smith, W. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, *21*(1), 3–8. https://doi.org/10.1177/002224295602100102

Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research*, *13*(5), 899–906.

Schubert, E., & Rousseeuw, P. J. (2019). Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In *International Conference on Similarity Search and Applications* (pp. 171-187). Springer, Cham.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological*, *58*(1), 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X

Titterington, D. (1976). Algorithms for computing D-optimal design on nite design

spaces. In *Proc. Conf. on Information Science and Systems* (pp. 213–16). Baltimore: Johns Hopkins University.

Tsiptsis, K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM*. John Wiley & Sons.

Vaccari, M., Collivignarelli, C., Tharnpoophasiam, P., & Vitali, F. (2009). Well sanitary inspection and water quality monitoring in Ban Nam Khem (Thailand) 30 months after 2004 Indian Ocean Tsunami. *Environmental Monitoring and Assessment*, *161*(1–4), 123–133. https://doi.org/10.1007/s10661-008-0732-5

Vicente-Tavera, S. (1992). Las técnicas de representación de datos multidimensionales en el estudio del Indice de Producción Industrial en la C.E.E. *[Tesis Doctoral]. Universidad de Salamanca*.

Vicente-Tavera, S., Molina-Ballesteros, E., Vicente, M. ., & Garcia-Talegon, J. (1999). Determination of the origin and evolution of building stones as a function of their chemical composition using the inertia criterion based on an HJ-Biplot. *Chemical Geology*, *153*, 37–51. https://doi.org/10.1016/S0009-2541(98)00151-X

Vicente-Villardón, J. L. (2017). MultBiplotR: Multivariate Analysis using Biplot. Retrieved from http://biplot.usal.es/classicalbiplot/multbiplot-in-r/

Vichi, M. (2000). Double k-means clustering for simultaneous classification of objects and variables. In *Advances in classification and data analysis* (pp. 43-52). Springer, Berlin, Heidelberg.

Vichi, M., & Kiers, H. (2001). Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis*, *37*, 49–64. https://doi.org/10.1016/S0167-9473(00)00064-5

Vichi, M., & Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, *53*(8), 3194–3208. https://doi.org/10.1016/j.csda.2008.05.028

Vinod, H. D. (1969). Integer Programming and the Theory of Grouping. *The American Statistical Association.*, *64*, 506–519.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley Publishing.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*(490), 713–726. https://doi.org/10.1198/jasa.2010.tm09415

Wolska, L., Sagajdakow, A., Kuczynska, A., & Namiesnik, J. (2007). Application of ecotoxicological studies in integrated environmental monitoring: possibilities and problems. *TrAC Trends in Analytical Chemistry*, *26*(4), 332–344. https://doi.org/10.1016/j.trac.2006.11.012

Xiong, H., Wu, J., & Chen, J. (2009). K-Means clustering versus valitation measures: A data distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39(2)*, 318–331.

Yu, S. S., Chu, S. W., Wang, C. M., Chan, Y. K., & Chang, T. C. (2018). Two improved k-means algorithms. *Applied Soft Computing Journal*, *68*, 747–755. https://doi.org/10.1016/j.asoc.2017.08.032

Zhang, B., Hsu, M., & Dayal, U. (1999). K-harmonic means-a data clustering algorithm, Technical Report HPL-1999-124, Hewlett-Packard Laboratories.

Zhang, B., Hsu, M., & Dayal, U. (2000). K-Harmonic Means. In *Proc. of International*

Workshop on Temporal,Spatial and Spatio-Temporal. Data Mining,TSDM2000. Lyon, France.