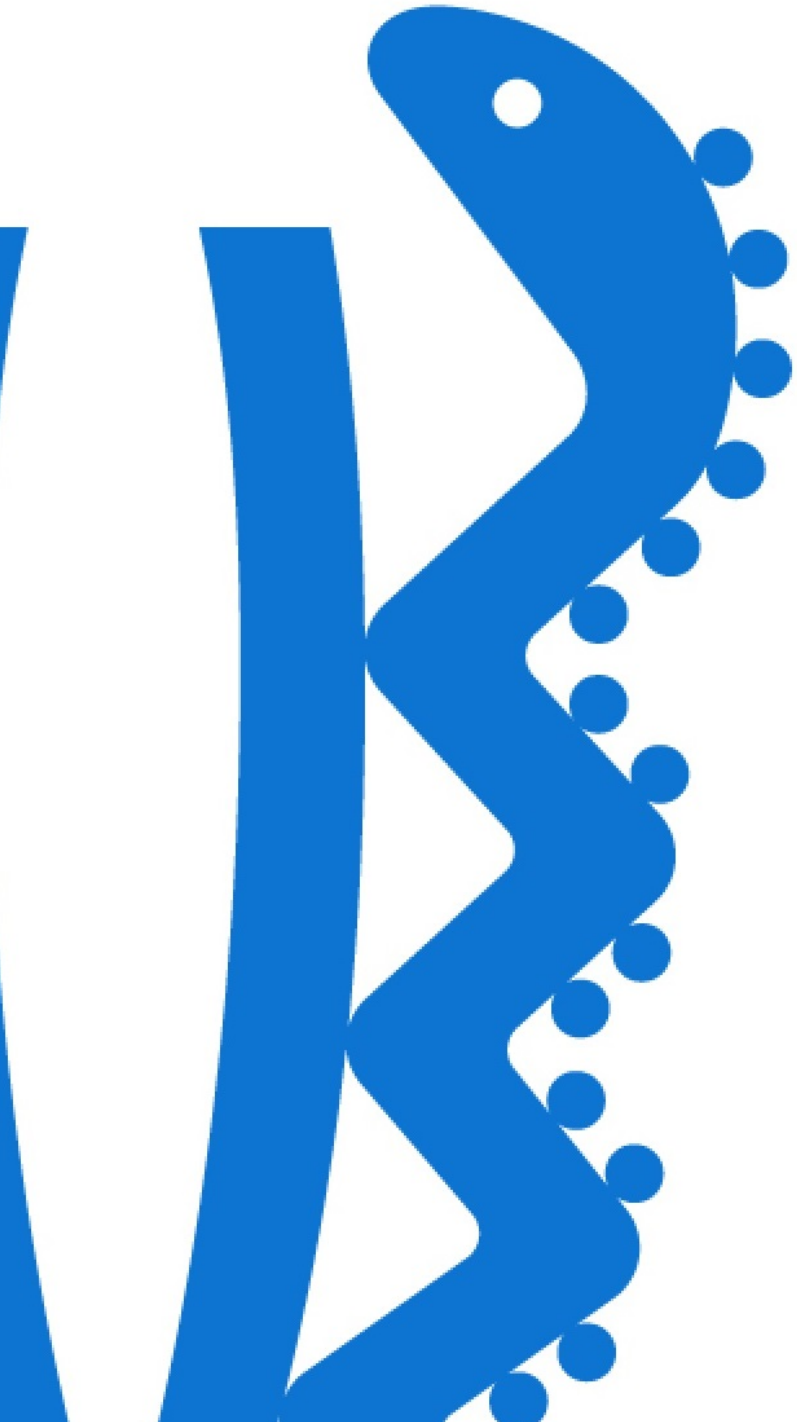
Universidad El Bosque

# Metodología para segmentación de un SARLAFT



Nombre del evaluado:

**Lincoln Ernesto Perez Perez**

Nombre del Tutor Empresarial :

**Brayan Ricardo Rojas Ormaza**

2020

# Declaración de autoría original y única

Yo, Lincoln Ernesto Perez Perez, declaro que este documento titulado *Metodología para segmentación de un SARLAFT* y los datos presentados son originales al igual que mi propio trabajo.

Confirmo que:

- Ninguna parte de este trabajo se ha presentado previamente para obtener un título en esta o en cualquier otra universidad.
- Las referencias al trabajo de otros han sido claramente reconocidas. Las citas del trabajo de otros han sido claramente indicadas y atribuidas a ellas.
- En los casos en que otros han contribuido a parte de este trabajo, dicha contribución ha sido claramente reconocida y distinguida de mi propio trabajo..
- Ninguno de estos trabajos se ha publicado anteriormente en otros lugares.

# Índice general

# Índice de figuras

# Índice de tablas

# Introducción

Dentro del área de *Financial Risk Management (FRM)* de KPMG[1] (en adelante la Firma), es una red global de firmas que presta servicios de Auditoría, Impuestos y Consultoría, que esta presente en 147 países; KPMG comenzó operaciones en Colombia en 1959, prestando los servicios de Auditoría, Impuestos y Asesoría, y en la actualidad la firma es una de las más reconocidas Firmas de Auditoría, Asesoría, Impuestos y Servicios Legales del país, prueba de esto es el importante portafolio de clientes nacionales y multinacionales; en Colombia la Firma cuenta con 5 oficinas, donde laboran más de 1.500 profesionales, entre los que están Contadores, Administradores de Empresa, Economistas, Estadísticos, Abogados, Ingenieros Industriales, Ingenieros de Sistemas, que apoyan el proceso de auditoría mediante la revisión y reestimación de diferentes procesos de las entidades financieras que son clientes de la firma, estos además plantean nuevos métodos que brinden resultados más precisos dentro del proceso de revisión. Algunos de los procesos revisados son los modelos de segmentación que implementan las entidades financieras; esta segmentación que hace parte del ciclo de gestión del riesgo, que se define en el *parte I del título IV del capítulo IV de la Circular Básica Jurídica de la Superintendencia Financiera de Colombia* [2], de siguientes manera: *« Es el proceso por medio del cual se lleva a cabo la separación de elementos en grupos homogéneos al interior de ellos y heterogéneos entre ellos. La separación se fundamenta en el reconocimiento de diferencias significativas en sus características (variables de segmentación)».*

[1] © 2020 KPMG S.A.S. y KPMG Advisory, Tax & Legal S.A.S., sociedades colombianas por acciones simplificadas y firmas miembro de la red de firmas miembro independientes de KPMG afiliadas a KPMG International Cooperative ("KPMG International"), una entidad suiza. Derechos reservados.

[2] https://www.superfinanciera.gov.co/inicio/normativa/normativa-general

La definición normativa anterior es ambigua, porque en las diferentes entidades financieras, al igual que en otras empresas, se presentan frecuentemente dudas respecti a ¿cómo segmentar?, ¿cuál es la mejor forma de segmentar? o ¿cómo selecciono las variables para ello?, entre otras. Estas mismas preguntas se las han motivado investigadores como Karin(2020), el cual busca mostrar los diferentes enfoques que pueden tener los métodos de agrupamiento estadístico, que están basados en aprendizaje profundo, en especial en el tema para bioinformática, o hay otros investigadores que van más allá, y crean nuevos algoritmos para resolver estas preguntas como Aljarah,Mararja(2020), inspirándose en la colaboración social y las actividades de caza de lobos grises.

En nuestro caso se quiere resolver dichas preguntas por medio de una metodología que sea práctica, sencilla y funcional, con el fin de que las empresas o entidades que requieran hacer segmentación en este campo financiero, puedan tener la una metodología mas robusta al igual que el control y verificación de dicha segmentación.

# Objetivo General

Establecer una metodología apropiada de segmentación para dar cumplimiento a la parte I del título IV del capítulo IV de la Circular Básica Jurídica de la Superintendencia Financiera de Colombia.

# Objetivos específicos

- Comparar los resultados de los diferentes algoritmos por medio de diferentes índices, que permitan validar la óptima selección del número de clusters y la pertenencia a estos.
- Medir el tiempo de ejecución y rendimiento de los algoritmos.

- Analizar la efectividad de la metodología propuesta, frente a diferentes metodologías utilizadas por entidades bancarias.

# Marco referencial

Dentro del marco normativo internacional y nacional; el Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación del Terrorismo(SARLAFT) en 1970 se expide las primeras disposiciones encaminadas a reprimir el lavado de activos en los Estados Unidos. Ya para 1988 se suscribió en Viena la "Convención de las Naciones Unidas contra el tráfico Ilícito de Narcóticos y Sustancias Sicotrópicas" ese mismo año se creó un mecanismo intergubernamental llamado Grupo de Acción Financiera Internacional (GAFI), que durante la década de los 90 consolido varias bases entre ellas las 40 recomendaciones, que hoy constituyen el estándar internacional sobre la lucha contra el lavado de activos(LA), el financiamiento al terrorismo(FT) y la Financiamiento a la proliferación de armas de destrucción masiva (PADM), en 1992 es emitida la norma en temas relacionados con el lavado de activos en Colombia con el Decreto 1872 del 23 de noviembre, en 1999 se creó la Unidad de Información y Análisis Financiero (UIAF), entidad encargada de prevenir y detectar operaciones que puedan ser utilizadas como instrumento para el LA/FT.El concepto de lavado de activos se limitaba al narcotráfico, por lo tanto habría que esperar hasta el año 2000, cuando en la ocasión de la Convención contra el crimen organizado transnacional, se amplían los delitos fuente de lavado de activos: Tráfico de armas, delitos contra el sistema financiero, la administración pública, o vinculados con el producto de los delitos objeto de un concierto para delinquir. Y en el 2002 la Ley 747 se modifica el Código Penal Colombiano se incluyen delitos como el tráfico de migrantes y la Trata de personas.

El trabajo de grado realizado en este documento es similar a la optimización del proceso de monitorio de transacciones (Correa Chaparro, 2015) que ayuda a documentar y formular propuestas de rediseño del procedimiento de monitoreo transaccional para

reducir los tiempos y costos; que en este caso, la metodología propuesta en este documento, sirve tanto para una reducción de tiempos, como para mejorar procesos de segmentación dentro del SARLAFT. La revisión propuesta en *"Data clustering: a review"* (Jain, Murty, y Flynn, 1999), donde se revisan resultados obtenidos y diferentes metodologías, que de manera comparativa con este documento los resultados y metodologías son suministradas por dos entidades financieras que por cuestión de privacidad de la información, esta se encuentran anonimizadas al igual que los datos suministrados, es por esto que para los datos se genero un ruido, que consta de una variable aleatoria normal, para que no que puedan ser identificadas dichas entidades. Dentro del documento se a revisado y probado diferentes metodologías de segmentación, similar a lo publicado en *An efficient k-means clustering algorithm: analysis and implementation*(Kanungo y cols., 2002) en el cual efectúan un análisis de la implementación del algoritmo K-Means; que dentro de este documento, también se revisara en cada escenario de ejemplo. Y por último, el documento *Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm* (Babichev, Lytvynenko, y Osypenko, 2017) guarda similitud con este documento, dado a que se busca implementar y probar que la metodología plantea, puede generar una solución factible en la segmentación, y en especial en el caso del SARLAFT.

# Justificación

La financiación del terrorismo (FT) como también el lavado de activos (LA) representan un gran riesgo para un país, dado que afecta en gran medida la estabilidad del sistema financiero, al igual que a la integridad de sus mercados; esto debido a las implicaciones globales y las redes utilizadas para el manejo de los recursos. Esta circunstancia destaca la importancia y urgencia de combatirlos, resultando esencial el papel que para tal propósito deben desempeñar las entidades vigiladas por la Superintendencia Financiera de Colombia

(SFC).[3]

La Superintendencia requiere que las entidades vigiladas implementen un Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT) con el fin de prevenir que sean utilizadas para dar apariencia de legalidad a activos provenientes de actividades delictivas o para la canalización de recursos hacia la realización de actividades terroristas. Con el desarrollo de los artículos 102 y siguientes del Estatuto Orgánico del Sistema Financiero (EOSF) y la consonancia con el articulo 22 de la Ley 964 de 2005, la SFC establece los criterios y parámetros mínimos que las entidades vigiladas deben atender en el diseño, implementación y funcionamiento del mencionado sistema. Este documento se centra en la primera fase, la cual describe dichos artículos, lo corresponde a la prevención del riesgo y cuyo objetivo es prevenir que introduzcan al sistema financiero recursos provenientes de actividades relacionadas con el lavado de activos o de la financiación del terrorismo (LA/FT). Por otro lado, dentro de la Firma el proceso del auditor realizará una revisión y seguimiento del SARLAFT configurado por las entidades y así verificar la correcta segmentación de estos, así como realizar recomendaciones que puedan mejorar los modelos implementados, puesto que esto asegura la continuidad del negocio de las entidades auditadas y el mantenimiento de la credibilidad de la firma auditora. De esta manera, al realizar este documento se sugiere una metodología con técnicas estadísticas, para ser aplicada como alternativa que se adapte a las características de las variables utilizadas dentro del SARLAFT de la entidad que lo requiera.

[3]https://www.superfinanciera.gov.co/inicio/normativa/normativa-general

# Capítulo 1

# Marco teórico

## 1.1. Clustering

Existe una amplia cantidad de técnicas de aprendizaje no supervisado, cuya finalidad es encontrar patrones o clúster dentro de un conjunto de observaciones. Las particiones se establecen de forma que, las observaciones que están dentro de un mismo clúster [1], que son similares entre ellas y distintas a las observaciones de otros clusters. El *clustering* es útil en diferentes disciplinas académicas, públicas, o de carácter privado como la economía, marketing, salud, política, entre otras; por lo tanto, se han desarrollado variantes y adaptaciones de sus algoritmos, los cuales pueden diferenciarse en tres grupos principalmente:

- Agrupamiento no jerárquico: *K-means*, *K-medoids*, CLARA, entre otros.
- Agrupamiento jerárquico: *Clustering* aglomerativo, entre otros.
- La combinación de los anteriores como *clustering* confuso *(fuzzy clustering)*.

Los métodos de *clustering* tienen en común que para poder obtener los *clusters* es necesario cuantificar la distancia entre las observaciones (datos), por lo que distancia que se emplea dentro del *clustering* como la cuantificación de dicha distancia entre observaciones

[1] Clúster es la hispanización del término de origen inglés *cluster*, que se traduce como 'rácimo', 'conjunto' o 'cúmulo'. Actualmente, su empleo es muy común en diferentes ámbitos como la informática, las ciencias y el mundo empresarial.

es muy relevante para este trabajo. Si se representan las observaciones en un espacio $P$ dimensional, siendo $N$ el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones estas serán más próximas, de esto es que nace el concepto de distancia.

## 1.2. Distancias

Una característica principal que hace de la segmentación un método adaptable y versátil, es que al probar diferentes escenarios estos pueden emplear un tipo de distancia diferente, es por esto que permite al estadístico escoger la mejor o la más adecuada para el trabajo en cuestión. Es por esto que a continuación, se describen algunas distancias que son frecuentemente utilizadas, y que dentro de la metodología fueron utilizadas y probadas.

### 1.2.1. Distancia euclídea

La distancia euclídea entre dos puntos x y y N-variados se define como la longitud del segmento que une ambos puntos(Lloyd, 1982). En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x,y), la distancia euclídea entre $x$ y $y$ viene dada por la ecuación:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{1.1}$$

### 1.2.2. Distancia máxima

Es la distancia máxima entre dos componentes de x y y (norma del supremo).

$$d(x, y) = \sum_{j=1}^{N} |x_j - y_j| \tag{1.2}$$

### 1.2.3. Distancia de Manhattan

Se le conoce también como *taxicab metric*, es la distancia entre dos puntos x y y como el sumatorio de las diferencias absolutas entre cada $N$ dimensión.(James, Witten, Hastie, y Tibshirani, 2013) Esta medida se ve menos afectada por outliers por lo que se puede decir que es más robusta que la distancia euclídea, debido a que no eleva al cuadrado las diferencias.

$$d_M(x, y) = \sum_{j=1}^{N} |(x_j - y_j) \tag{1.3}$$

### 1.2.4. Distancia de Canberra

Los términos con numerador y denominador cero se omiten de la suma y se tratan como si faltaran los valores.

$$d(x, y) = \sum_{j=1}^{N} \frac{|x_j - y_j|}{|x_j| + |y_j|} \tag{1.4}$$

### 1.2.5. Distancia de Minkowski

La norma $p$, la $p^{th}$ raíz de la suma de la $p^{th}$ poderes de las diferencias de los componentes.

$$d(x, y) = \left( \sum_{j=1}^{d} |x_j - y_j|^p \right)^{\frac{1}{p}} \tag{1.5}$$

## 1.3. Agrupamiento no Jerárquico

### 1.3.1. Método K-means

Este método se basa en agrupar observaciones en *K clusters*, donde $K$ es determinado por el estadístico antes de proceder con el algoritmo; al obtener los mejores *clusters*, cuya varianza interna sea lo más pequeña o baja posible (MacQueen, 1967).

Esto se puede reducir a un problema estrictamente de optimización, el cual se trata de repartir las observaciones (datos) en $K$ clusters de forma tal, que la suma de las varianzas internas de todos los clusters sea lo más baja posible. Antes de proceder con medidas para el calculo de la varianza, se debe tener en cuenta lo siguiente:

- Considere los siguiente $C_1, ..., C_k$ como los *data sets*, integrados por índices de las observaciones pertenecientes a cada cluster.
- $C_1 \cup C_2 \cup ... \cup C_k = \{1, ..., n\}$ que toda observación o dato, pertenece al menos a uno de los *K clusters*.
- $C_k \cap C_{k'} = \varnothing$ para todo $k \neq k^{'}$ , significa que ninguna observación pertenece a más de un *cluster* a la vez.

Para el calculo de la varianza interna de un *cluster* $(W(Ck))$ son dos las metodologías mas utilizadas:

- La suma de las distancias euclídeas al cuadrado entre cada observación$(x_i)$ y la media$(\mu)$ de su *cluster*. Equivale a la suma de cuadrados intra *cluster*:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{1.6}$$

- La suma de las distancias euclídeas al cuadrado entre todos los pares de observaciones

que forman el *cluster*, dividida entre el número de observaciones del *cluster*.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{1.7}$$

Para minimizar la suma total de varianza interna $\sum_{k=1}^{k} W(C_k)$ de forma exacta es encontrar el mínimo global, para ello tenemos el siguiente algoritmo:

1. Determinar el numero de *clusters(q)* que se quieren crear.
2. De forma aleatoria seleccionar $q$ observaciones como medias iniciales.
3. Se asignan las observaciones al la media o centroide más cercano.
4. Ahora para cada uno de los *K-clusters* se recalcular su media.
5. Repetir el 3 y 4 hasta que las observaciones no varíen o se alcance el número de iteraciones pre-establecido por el estadístico.

De manera gráfica se puede entender mejor la explicación del algoritmo:

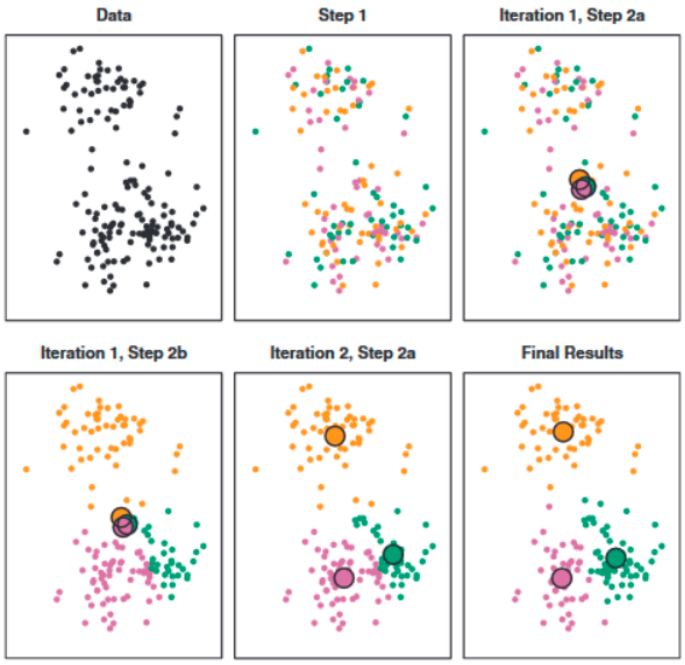



**Figura 1.1:** Proceso de iteración

### 1.3.2. Método K-medoids (PAM)

Es un método que se basa en agrupar las observaciones en $q$ clusters, y para este método $q$ es un valor dado por el estadístico, cada cluster ($Kq$) está representado por una observación presente en el *cluster (medoid)*, aqui podemos ver una diferencia con *K-means* dado a que en en esa metodología cada uno de los *cluster* es representado por su promedio.

Para esta metodología dicha observación representativa es llamada *medoid*, y este corresponde al elemento más central del *cluster*. Si se compara utilizar *medoids* en lugar de la media que se hace en *K-means*, hace que *K-medoids* sea método más robusto dado a se esta usando un valor que se encuentra en la información suministra, al igual que es menos afectada por datos atípicos.

Una de las diferencias más significativas con algoritmo *K-means*, es que en este se minimiza la suma total de cuadrados intra-cluster, y en el algoritmo PAM lo que se minimiza, es el promedio de diferencias de las observaciones respecto a su *medoid*.

Una analogía planteada en Kassambra (2017), es que la forma en la que podemos interpretar *k-means* es que sea la media, y *k-medoids* sea la mediana, de esta manera es más claro tener un punto de comparación y de la lógica del cálculo. Ahora para este trabajo el algoritmo a utilizar es llamado *Partitioning Around Medoids (PAM)*, el cual es uno de los más utilizado en *k-medoids* el cual tiene los siguientes pasos:

1. Tomar $q$ observaciones aleatorias, estas van a ser nuestros medoids iniciales.

2. Se calcula una matriz de las distancias entre los medoides y todas las observaciones.

3. Con la matriz se procede a asignar cada observación a su medoide más cercano.

4. Se guarda la distancia promedio en en cada cluster.

5. Se retorna al paso 2, si seleccionando otro medoide en el paso 1 este consiga reducir la distancia promedio en cada *cluster*, si esto sucede, guardar el que medoide que

genere la menor distancia promedio al *cluster*; de ser el medoide que menor distancia genere, se termina el algoritmo.

Para tener en cuenta el método de K-medoide es más utilizado cuando el estadístico o investigador conoce o se sospecha de que hay datos atípicos, y si este es el caso, la literatura nos recomienda utilizar la distancia *Manhattan* para la medida de similitud, dado a que esta es menos sensible que la distancia euclidiana.

### 1.3.3. Método CLARA

Es un método que tiene como idea base K-medoides (PAM) junto a técnicas de remuestreo, este es usado cuando se tienen grandes cantidades de datos; este método surge por las limitaciones del método K-medoides, dado a que este requiere de una gran cantidad de recursos tecnológicos (uno de ellos es la necesidad de gran cantidad de memoria RAM), lo cual puede suponer una limitación para el estadístico. Entonces para solventar esta limitación, es que el método para encontrar los medoids no va utilizar todos los datos, sino, va a requerir de una muestra aleatoria de tamaño $n$; una vez se tenga muestra, lo que resta es aplicar el algoritmo PAM ya definido en el 1.3.2.

Un punto a resaltar del método CLARA es al momento de medir la calidad de los medoides, dado a que esta medición inicia con la suma total de las distancias entre cada observación del cluster y su correspondiente medoide (distancias intracluster), luego se repite este proceso un número predeterminado de veces con el objetivo de reducir el sesgo muestral[2], y por último se seleccionan los clusters finales con aquellos medoids que han conseguido la menor suma total de distancias. Para tener mas claro y de manera ordenada el método, a continuación se describen los pasos del algoritmo:

1. El estadístico ha definido el tamaño de la muestra $n$, que por lo general se recomienda que el $n$ sea de 20 % de la base.

[2]En este caso sesgo muestral, se refiere a la variación que va a presentar mi calculo, debido a la forma en que es seleccionada los datos para la muestra

2. Se extrae la muestra de dados de base completa.

3. Se aplica el algoritmo PAM para identificar cuáles son los k-medoides.

4. Utilizando los medoids del paso anterior, se agrupan los datos.

5. Se calcula las distancias entre cada observación del cluster y su correspondiente medoides (distancias intracluster).

6. Se repite desde el paso 2 al paso 5, un numero de veces que defina el estadístico, y como cluster final se selecciona aquel que tenga menor suma total de distancias intracluster.

## 1.4. Índices de validación de *Clusters*

En la bibliografía de las librerías de R[3], existen múltiples formas para la identificación del número óptimo de clusters. En este caso se utilizará la función **NbClust()** del paquete **NbClust** que incorpora 30 índices distintos; esto nos brinda la posibilidad de calcularlos todos los métodos en un único paso, para poder identificar el valor en el que los índices más coinciden, dándonos una mayor seguridad en la elección del número de clusters.
Estos índices de *clusters* combinan información sobre la relación intragrupo y el aislamiento entre *clústers*, así como otros factores, como propiedades geométricas o estadísticas de los datos, el número de objetos de datos y las medidas de disimilitud o similitud.
A continuación, se presentan los índices implementados en el paquete **NbClust** y cómo seleccionar el número óptimo de **clústers** para cada índice. Para ello se define: $n =$ número de observaciones,
$p =$ número de variables
$q =$ número de conglomerados,
$X = x_{ij},\ i = 1, 2, \ldots, n\ , j = 1, 2, \ldots p,$

[3]R es un entorno de programación libre que se utiliza para el procesamiento y análisis estadístico de datos implementado en el lenguaje S de GNU.

$X = n$ x $p$ matriz de datos con $p$ variables medidas en $n$ observaciones independientes,

$\bar{X} = q$ x $p$ matriz de medias del conglomerado,

$\bar{x}$ = centro de la matriz X,

$n_k$ = número de objectos en el conglomerado $C_k$,

$c_k$ = centro del conglomerado $C_k$,

$x_i = p$ dimensiones del vector de observaciones de cada $i-esmo$ objeto en el conglomerado $C_k$,

$||x|| = (x'x)^{1/2}$,

$W_q = \sum_{k=1}^{q} \sum_{i \epsilon C_k} (x_i - c_k)(x_i - c_k)'$,es la matriz de dispersión dentro del grupo para datos agrupados en $q$ grupos

$Bq = \sum_{k=1}^{q} n_k (c_k - \bar{x})(c_k - \bar{x})'$ es la matriz de dispersión entre grupos para datos agrupados en q grupos,

$N_t$ = número total de pares de observaciones en el conjunto de datos:

$$N_t = \frac{n(n-1)}{2},$$

$N_w$ número total de pares de observaciones pertenecientes al mismo conglomerado:

$$N_w = \sum_{k=1}^{q} \frac{n_k(n_k - 1)}{2},$$

$N_b$ = total de pares de observaciones pertenecientes a un conglomerado diferente:

$$N_b = N_t - N_w,$$

$S_w$ = Suma de las distancias dentro del clúster:

$$S_w = \sum_{k=1}^{q} \sum_{i,j \epsilon c_k} d\left\{x_i, x_j\right\},$$

$S_b$ = Suma de las distancias entre cluster.

$$S_b = \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} \sum_{i \epsilon c_k} d\left\{x_i, x_j\right\},$$

### 1.4.1. Índice Ch

El Calinski y Harabasz (CH) indice (Caliński y Harabasz, 1974) esta definido en la ecuación 1.8 .

$$CH(q) = \frac{\text{traza}(B_q)/(q-1)}{\text{traza}(B_q)/(q-1)} \tag{1.8}$$

Se considera que el valor de $q$, que maximiza $CH(q)$, especifica el número de conglomerados en (Caliński y Harabasz, 1974).

### 1.4.2. Índice Duda

En Duda(1973) se propone un criterio de razón $Je(2) = Je(1)$ en la ecuación 1.9 , donde $Je$ (2) es la suma de los errores cuadrados dentro de los grupos cuando los datos se dividen en dos grupos, y $Je$ (1) da los errores al cuadrado cuando solo está presente un grupo.

$$Duda = \frac{Je(2)}{Je(1)} = \frac{W_k + W_l}{W_m}, \tag{1.9}$$

Se supone que los grupos $C_k$ y $C_l$ se fusionan para formar $C_m$.

En Gordon(1999), el número óptimo de conglomerados es el más pequeño $q$ tal que

$$Duda \geq 1 - \frac{1}{\pi p} - z\sqrt{\frac{2\left(1 - \frac{8}{\pi^2 p}\right)}{n_m p}} = \text{Punto Critico Duda}, \tag{1.10}$$

donde y $Z$ es un puntaje normal estandár. Se probaron varios valores para la puntuación estándar y los mejores resultados se obtuvieron cuando el valor se estableció en 3,20 .

### 1.4.3. Índice Pseudot2

Duda(1973) propuso otro índice, Pseudo $t2$, que sólo se puede aplicar a métodos jerárquicos. Se calcula usando la ecuación 1.11 .

$$Pseudot2 = \frac{V_{kl}}{\frac{W_k+W_l}{n_k+n_l-2}}. \tag{1.11}$$

donde $V_{kl} = W_m - W_k - W_l$, si $C_m = C_k \cup C_l$ .

Gordon(1999) especificó que el número óptimo de conglomerados es el más pequeño $q$ tal que:

$$Pseudot2 \leq \left(\frac{1 - \text{Punto Critico Duda}}{\text{Punto Critico Duda}}\right) \times (n_k + n_l - 2). \tag{1.12}$$

### 1.4.4. Cíndice

El Cíndice se revisó en Hubert (1970). Se calcula usando la ecuación 1.13

$$\text{Cindex} = \frac{S_w - S_{min}}{S_{max} - S_{min}}, S_{min} \neq S_{max}, \text{ Cindex } \epsilon\ (0,1), \tag{1.13}$$

donde

- $S_{min}$ es la suma de las $N_w$ distancias más pequeñas entre todos los pares de puntos en el conjunto de datos completo (hay $N_t$ pares de este tipo);
- $S_{max}$ es la suma de las $N_w$ mayores distancias entre todos los pares de puntos en el conjunto de datos completo.

El valor mínimo del índice se utiliza para indicar el número óptimo de conglomerados.

### 1.4.5. Índice Gamma

Este índice, calculado usando la Ecuación 1.14 , representa una adaptación de Goodman y Kriskal Estadística gamma para uso en situaciones de agrupamiento (Baker y Hubert,

1975). Se hacen comparaciones entre todas las diferencias dentro del clúster y todas las diferencias entre clústeres. similitudes. Se considera que una comparación es concordante $[s(+)]$ (resp. Discordante $[s()]$) si la disimilitud dentro del conglomerado es estrictamente menor (o estrictamente mayor) que una disimilitud entre conglomerados semejanza; las igualdades entre miembros de dos conjuntos de diferencias no se tienen en cuenta en la definición del índice (Gordon, 1999).

$$\text{Gamma} = \frac{s(+) - s(-)}{s(+) - s(-)}, \tag{1.14}$$

donde

- $s(+)$ = número de comparaciones concordantes,
- $s(-)$ = número de comparaciones discordantes.

El valor máximo del índice se toma para representar el número correcto de conglomerados (Milligan y Cooper, 1985) . En el paquete **NbClust** , este índice se calcula solo si el argumento del índice se establece en "gamma" o "alllong" debido a su alta demanda computacional.

### 1.4.6. Índice Beale

Beale(1969) propuso el uso de una F-ratio para probar la hipótesis de la existencia de $q1$ *versus* $q2$ agrupaciones en los datos ($q2 > q1$). El índice de Beale se calcula usando la ecuación 1.15

$$\text{Beale} = F = \frac{\left(\frac{V_{kl}}{W_k + W_l}\right)}{\left(\left(\frac{n_m - 1}{n_m - 2}\right) 2^{\frac{2}{p}} - 1\right)}, \tag{1.15}$$

donde $V_{kl} = W_m - W_k - W_l$. Se supone que los grupos $C_k$ y $C_l$ se fusionan para formar $C_m$. El número óptimo de conglomerados se obtiene comparando $F$ con una distribución $F_p, (n_m - 2)p$. La hipótesis nula de un solo grupo se rechaza para valores significativamente grandes de $F$ (Gordon, 1999). De forma predeterminada, el paquete **NBClust**, se usó el

nivel de significación del 10 % para rechazar el valor nulo hipótesis (alphaBeale = 0.1 en la función **NbClust** .

### 1.4.7. Índice CCC

El Criterio de agrupamiento cúbico (CCC) es la estadística de prueba proporcionada por el software SAS [4] paquete Sarle (1983). Se calcula usando la Ecuación 1.16.

$$CCC = \ln\left[\frac{1-E(R^2)}{1-R2}\right]\frac{\sqrt{\frac{np*}{2}}}{(0.001+E(R2))^{1.2}} \tag{1.16}$$

donde

$$R^2 = 1 - \frac{\text{traza}(X'X - \bar{X}'Z'Z\bar{X})}{\text{traza}(X'X)}$$

- X'X = matriz de suma de cuadrados y productos cruzados (SSCP) de muestra total $p \times p$,
- $\bar{X} = (Z'Z)^{-1}Z'X$
- $Z$ es una matriz de indicadores de conglomerados $n \times q$ con el elemento $z_{ik} = 1$ si la i-ésima observación pertenece al k-ésimo grupo y $z_{ik=0}$ en caso contrario

$$E(R^2) = 1 - \left[\frac{\sum_{j=1}^{p*}\frac{1}{n+u_j} + \frac{\sum_{j=p*+1}^{p}u_j^2}{n+u_j}}{\sum_{j=1}^{p}u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1+\frac{4}{n}\right]$$

- $u_j = \dfrac{s_j}{c}$,
- $s_j$ = raíz cuadrada del j-ésimo valor propio de $X'X/(n-1)$,
- $c = \left(\frac{v^*}{q}\right)^{\frac{1}{p*}}$ ,
- $v^* = \prod_{j=1}^{p^*} s_j$,

[4] es un paquete de software estadístico desarrollado por el Instituto SAS para la gestión de datos, análisis avanzado, análisis multivariado, inteligencia empresarial, investigación criminal y análisis predictivo.

- $p^*$ se elige para que sea el número entero más grande menor que $q$ de modo que $u_p^*$ no sea menor que uno.

El valor máximo del índice se utiliza para indicar el número óptimo de conglomerados en el conjunto de datos (Milligan y Cooper, 1985).

### 1.4.8. Índice Ptbiserial

Este índice, examinado por (Milligan, 1980), (Milligan, 1981) y (Kraemer, 1982) , es simplemente un punto biserial. correlación entre la matriz de disimilitud de entrada bruta y una matriz correspondiente que consiste de 0 o 1 entradas. Se asigna un valor de 0 si los dos puntos correspondientes están agrupados por el algoritmo. De lo contrario, se asigna un valor de uno (Milligan, 1980) . Dado que los valores positivos mayores re ect una mejor t entre los datos y la partición, el valor máximo del índice se utiliza para seleccionar el número óptimo de conglomerados en el conjunto de datos (Milligan y Cooper, 1985) . El coeficiente de correlación biserial puntual se calcula utilizando la Ecuación 1.17 (Milligan, 1981).

$$\text{Ptbiserial} = \frac{\left[\bar{S}_b - \bar{S}_w\right]\left[N_w N_b / N_t^2\right]^{1/2}}{s_d}, \tag{1.17}$$

donde

- $\bar{S}_w = S_w/N_w$,
- $\bar{S}_b = S_b/N_b$,
- $s_d =$ desviación estandár de todas las distancias.

### 1.4.9. Índice Gplus

Este índice fue revisado por Rouhlf (1974) y examinado por Milligan (1981). Se calcula usando la Ecuación 1.18 .

$$\text{Gplus} = \frac{2s(-)}{N_t(N_t - 1)}, \tag{1.18}$$

donde $s(-)$ es el número de comparaciones discordantes, es decir, el número de veces donde dos los puntos que estaban en el mismo grupo tenían una distancia mayor que dos puntos no agrupados juntos (Milligan, 1981). Los valores mínimos del índice se utilizan para determinar el óptimo número de conglomerados en los datos (Milligan y Cooper, 1985). En el paquete **NbClust** , este índice se calcula solo si el argumento del índice se establece en "gplus" o "todo", ya que es computacionalmente muy caro.

### 1.4.10. Índice DB

El índice de Davis y Bouldin (1979) es una función de la relación de suma de la dispersión dentro del grupo a la separación entre grupos. Se calcula usando la ecuación 1.19

$$\mathrm{DB}(q) = \frac{1}{q}\sum_{k=1}^{q} \max\left(\frac{\delta_k + \delta_l}{d_{kl}}\right) k \neq l, \tag{1.19}$$

donde

- $K, l = 1, \ldots, q =$ número de cluster,
- $d_{kl} = \sqrt[v]{\sum_{j=1}^{p} |c_{kj} - c_{lj}|^v} =$ distancia entre los centroides de los clusters $C_k$ y $C_l$ (para $v = 2, d_{kl}$ es la distancia euclidiana),
- $\delta_k = \sqrt[u]{\frac{1}{n_k}\sum_{i \epsilon C_k}\sum_{j=1}^{p} |x_{ij} - C_{kj}|^u} =$ medida de dispersión de un grupo $C_k$ (para $u = 2$, $\delta_k$ es la desviación estándar de la distancia de los objetos en el grupo $C_k$ al centroide de este cluster).

Se considera que el valor de q minimizando $DB(q)$ especifica el número de conglomerados Michigan y Cooper (1983), Dvies y Bouldin (1979)

### 1.4.11. Índice Frey

El índice propuesto por Frey y Van Groenewoud (1972), cuando introdujeron su k-método de agrupamiento, solo se puede aplicar a métodos jerárquicos. Como se muestra

en la Ecuación 1.20 , es la razón de las puntuaciones de diferencia de dos niveles sucesivos en la jerarquía. El numerador es la diferencia entre las distancias medias entre grupos, $\bar{d_b}$, de cada una de las dos jerarquías niveles (nivel $j$ y nivel $j+1$). El denominador es la diferencia entre la media dentro de distancias de racimo, $\bar{d_w}$, de los dos niveles (nivel $j$ y nivel $j+1$). Los autores propusieron, utilizando una puntuación de razón de 1.00, para identificar el nivel de agrupación correcto. Las proporciones a menudo variaban arriba y por debajo de 1.00.

Los mejores resultados se obtuvieron cuando se continuó con la agrupación hasta que la última proporción cayó por debajo de 1,00. En este punto, el nivel de clúster anterior a este se tomó como partición óptima. Si la proporción nunca cayó por debajo de 1,00, se asumió una solución de un grupo (Milligan y Cooper, 1985).

$$\text{Frey} = \frac{\bar{S}_{b_{j+1}} - \bar{S_{bj}}}{\bar{S}_{w_{j+1}} - \bar{S_{wj}}}, \tag{1.20}$$

donde

- $\bar{S_b} = S_b/N_b =$ distancia media entre cluster,
- $\bar{S_w} = S_w/N_w =$ distancia media dentro del cluster.

### 1.4.12. Índice Hartigan

The Hartigan index (Hartigan, 1975) is computed using Equation .

$$\text{Hartigan} = \left(\frac{\text{traza}(W_q) - 1}{\text{traza}(W_{q+1})}\right)(n - q - 1), \tag{1.21}$$

donde $q \,\epsilon\, \{1, \ldots, n-2\}$ . La diferencia máxima entre niveles jerárquicos se toma como indicando el número correcto de agrupaciones en los datos (Milligan y Cooper, 1985).

### 1.4.13. Índice Tau

El índice Tau, revisado por (Rohlf, 1974) y probado por (Milligan, 1981), se calcula entre entradas correspondientes en dos matrices. El primero contiene las distancias entre elementos y el la segunda matriz 0/1 indica si cada par de puntos está dentro del mismo grupo o no. El índice Tau se calcula usando la Ecuación 1.22.

$$\text{Tau} = \frac{s(+) - s(-)}{[(N_t(N_t-1)/2 - t)(N_t(N_t-1)/2 - t)]^{1/2}} \tag{1.22}$$

- $s(+)$ representa el número de veces en que dos puntos no agrupados distancia mayor que dos puntos que estaban en el mismo grupo, es decir,$i.w.,\ s(+)$ es el número de comparaciones concordantes,
- $s(-)$ representa el resultado inverso (Milligan, 1981), es decir, $i.e.,\ s(-)$ es el número de discordantes comparaciones .
- $N_t$ es el número total de distancias y $t$ es el número de comparaciones de dos pares de puntos donde ambos pares representan comparaciones de conglomerados o ambos pares están entre comparaciones de conglomerados.

Se considera que el valor máximo del índice indica el número correcto de conglomerados (Milligan y Cooper, 1985). En el paquete **NbClust** , este índice se calcula solo si index = "tau" o index = "alllong" , porque es computacionalmente muy caro.

### 1.4.14. Índice Ratkowsky

(Ratkowsky y Lance, 1978) propusieron un criterio para determinar el número óptimo de clústeres basados $\frac{\bar{S}}{q^{1/2}}$. El valor de $\bar{S}$ es el promedio de las razones de ($\text{BGSS}_j/\text{TSS}_j$) donde BGSS representa la suma de cuadrados entre los conglomerados (grupos) para cada variable y TSS para la suma total de cuadrados de cada variable (Hill, 1980) . El número óptimo de conglomerados es el valor de $q$ para el que $\frac{\bar{S}}{q^{1/2}}$ tiene su valor máximo (Milligan y Cooper, 1985). Si el valor de q se hace constante, Ratkowsky y Lance El criterio se

puede reducir de $\frac{\bar{S}}{q^{1/2}}$ a $\bar{S}$ (Hill, 1980). En el paquete **NbClust** , el índice de Ratkowsky y Lance se calcula utilizando la Ecuación 1.23.

$$\text{Ratkowsky} = \frac{\bar{S}}{q^{1/2}}, \tag{1.23}$$

donde

- $\bar{S^2} = \frac{1}{p}\sum_{j=1}^{p} \frac{\text{BGSS}_j}{\text{TSS}_j}$,
- $\text{BGSS}_j = \sum_{k=1}^{q} n_k (C_{kj} - \bar{x_j})^2$,
- $\text{TSS}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x_j})^2$.

### 1.4.15. Índice Scott

(Scott y Symons, 1971) introdujeron un índice basado en la Ecuación 1.24, donde $n$ es el número de elementos en el conjunto de datos, $T$ es la suma total de cuadrados y $W_q$ es la suma de cuadrados dentro los grupos $q$, como se de nió anteriormente.

$$\text{Scott} = n\log\frac{\det(T)}{\det(W_q)} \tag{1.24}$$

La diferencia máxima entre los niveles jerárquicos se utiliza para sugerir el número correcto de particiones.

### 1.4.16. Índice Marriot

(FHC, 1971) propuso el siguiente índice calculado utilizando la Ecuación 1.25.

$$\text{Marriot} = q^2\det(W_q) \tag{1.25}$$

La diferencia máxima entre niveles sucesivos se utiliza para determinar la mejor partición.

### 1.4.17. Índice Ball

(Ball y Hall, 1965) propusieron un índice basado en la distancia promedio de los elementos a su los respectivos centroides del cluster. Se calcula usando la ecuación 1.26

$$\text{Ball} = \frac{W_q}{q}. \tag{1.26}$$

La mayor diferencia entre niveles se utiliza para indicar la solución óptima.

### 1.4.18. Índice Trcovw

Este índice, examinado por (Milligan y Cooper, 1985), representa el rastro de dentro de los clusters matriz de covarianza agrupada. Se calcula usando la ecuación 1.27

$$\text{Trcovw} = \text{traza}(\text{cov}(W_q)) \tag{1.27}$$

Las puntuaciones máximas de diferencia entre niveles se utilizan para indican la solución óptima.

### 1.4.19. Índice Tracew

Este índice ha sido uno de los índices más populares sugeridos para su uso en contextos de agrupación. (Milligan y Cooper, 1985); (Edwards y Cavalli-Sforza 1965); (Friedman y Rubin, 1967); (Orloci 1967); (Fukunaga y Koontz 1970). Se calcula usando la ecuación 1.28

$$\text{Tracew} = \text{traza}(W_q) \tag{1.28}$$

Dado que el criterio aumenta monótonamente con soluciones que contienen menos grupos, la Las puntuaciones máximas de las segundas diferencias se utilizan para determinar el número de conglomerados en los datos.

### 1.4.20. Índice Friedman

Este índice fue propuesto por (Friedman y Rubin, 1967), como base para una estructura no jerárquica. método de agrupamiento. Se calcula usando la ecuación 1.29

$$\text{Friedman} = \text{traza}(W_q^{-1} B_q) \tag{1.29}$$

Se utiliza la diferencia máxima en los valores de este criterio. para indicar el número óptimo de conglomerados (Milligan y Cooper, 1985).

### 1.4.21. Índice McClain

El índice de McClain y Rao (McClain y Rao, 1975) consiste en la razón de dos términos Ecuación 1.30. El primer término es el promedio dentro de la distancia del grupo, dividido por el número dentro de las distancias del grupo. El valor del denominador es el promedio entre la distancia del grupo dividido por el número de distancias de racimo.

$$\text{McClain} = \frac{\bar{S_w}}{\bar{S_b}} = \frac{S_w/N_w}{S_b/N_b} \tag{1.30}$$

El valor mínimo del índice se utiliza para indicar el número óptimo de agrupaciones.

### 1.4.22. Índice Rubin

(Friedman y Rubin, 1967) propuso otro criterio basado en la razón del determinante de la suma total de cuadrados y la matriz de productos cruzados al determinante de la combinación dentro matriz de conglomerados. Este criterio se calcula utilizando la ecuación 1.31.

$$\text{Rubin} = \frac{\det(T)}{\det(W_q)} \tag{1.31}$$

El valor mínimo de las segundas diferencias entre niveles es utilizado para seleccionar el número óptimo de agrupaciones (Milligan y Cooper, 1985); (Dimitriadou et al. 2002).

### 1.4.23. Índice KL

El índice KL propuesto por (Krzanowski y Lai, 1988) se define mediante la Ecuación 1.32

$$\mathrm{Kl}(q) = \left| \frac{\mathrm{DIFF}_q}{\mathrm{DIFF}_{q+1}} \right|, \tag{1.32}$$

donde $\mathrm{DIFF}_q = (q-1)^{2/p}\mathrm{traza}(W_{q-1}) - q^{2/p}\mathrm{traza}(W_q)$ . El valor de q, maximizando KL (q), se considera que especifica el número óptimo de agrupaciones.

### 1.4.24. Índice Silhouette

(Rousseeuw, 1987) introdujo el índice de silueta calculado utilizando

$$\mathrm{Silhouette} = \frac{\sum_{i=1}^{n} S(i)}{n}, \mathrm{Silhouette}\epsilon[-1, 1], \tag{1.33}$$

Este método maximiza la media de los coeficientes de silhouette($s_i$) o índices silueta; este coeficiente busca cuantificar la buena asignación que se ha hecho de una observación($i$), comparando su similitud con el resto de observaciones dentro de su cluster, contra las observaciones de los otros clusters, y el valor puede estar entre -1 y 1. Para denotar o simplificar mejor esta idea los valores mas altos de un coeficiente o índice, implica que dicha observación ha sido asignada de manera correcta; el algoritmo para el calculo de cada uno de los coeficientes de silhouette es:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}} \tag{1.34}$$

1. Se calcula el promedio de las distancias ($a_i$) entre la observación $i$ con el resto de observaciones del mismo cluster. Y cuanto sea menor el $a_i$,es que es mejor su asignación de la observación.

$$a(i) = \frac{\sum_{j\ \epsilon\{C_r/i\}} d_{ij}}{n_r - 1} \tag{1.35}$$

2. Ahora se calcula la distancia promedio de la observación $i$ con los demas clusters.

3. Se define como $b_i$ como la menor de todas las distancias promedio entre $i$ con el resto de los clusters, de manera mas sencilla, es la distancia al cluster más próximo.

$$b(i) = \min_{s \neq r} \{d_i C_s\}, \tag{1.36}$$

4. $d_{iC_s} = \dfrac{\sum_{j \ \epsilon \ C_s} d_j}{n_s}$ es la disimilitud promedio del i-ésimo objeto con todos los objetos del grupo $C_s$.

El valor máximo del índice se utiliza para determinar el número óptimo de agrupaciones en el datos (Kaufman y Rousseeuw 1990). $S(i)$ no está definido para $k = 1$ (solo un cluster).

### 1.4.25. Índice Gap

La estadística de Gap estimada propuesta por (Tibshirani, Walther, y Hastie, 2001) se calcula utilizando Equa- ción 1.37

$$\text{Gap}(q) = \frac{1}{B} \sum_{b=1}^{B} \log W_{qb} - \log W_q, \tag{1.37}$$

donde $B$ es el número de conjuntos de datos de referencia generados mediante prescripción uniforme (Tibshirani y cols., 2001) y $W_{qb}$ es la matriz de dispersión interna definida como en el índice de Hartigan. los El número óptimo de conglomerados se elige encontrando el $q$ más pequeño de modo que:

$$\text{Gap}(q) \geq \text{Gap}(q+1) - S_{q+1}, \quad (q = 1, \ldots, n-2),$$

donde

- $S_q = sd_q \sqrt{1 + 1/B}$

- $sd_q$ es la desviación estandar de $\{\log W_{qb}\}$ , $b = 1, \ldots, B$ : $sd_q = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\log W_{qb-\bar{l}}\right)^2}$,
- $\bar{l} = \frac{1}{B}\sum_{b=1}^{B}\log W_{qb}$.

En el paquete **NbClust** , el índice Gap se calcula solo si método = "gap" o método = "alllong", debido a su alto costo computacional.

### 1.4.26. Díndice

El Díndice (Lebart, Morineau, y Piron, 2000) se basa en la ganancia de agrupamiento en la inercia intragrupo. La inercia intralúster mide el grado de homogeneidad entre los datos asociados con un clúster. Calcula sus distancias en comparación con el punto de referencia que representa el perfil del cluster, es decir, el centroide del cluster en general. Se puede definir mediante la ecuación 1.38

$$\omega(P^q) = \frac{1}{q}\sum_{k=1}^{q}\frac{1}{nk}\sum_{x_i \epsilon C_k} d(x_i, c_k) \tag{1.38}$$

Dadas dos particiones, $P^{k1}$ compuesto por $k-1$ grupos y $P_k$ compuesto por $k$ clusters, el La ganancia de agrupamiento en la inercia intra-grupo se define como se muestra en la Ecuación 1.39.

$$\text{Gain} = \omega(P^{q-1}) - \omega(P^q) \tag{1.39}$$

Esta ganancia de agrupamiento debe minimizarse.
La configuración óptima del cluster se puede identificar por la rodilla afilada que corresponde a un disminución significativa de las primeras diferencias de ganancia de agrupamiento versus el número de agrupaciones. Este codo o gran salto de los valores de ganancia se puede identificar por un pico signi cativo en segundos diferencias de ganancia de agrupamiento.

### 1.4.27. Índice Dunn

El índice de Dunn (Dunn, 1974) define la relación entre la distancia mínima entre clústeres y distancia máxima intracluster. Este índice viene dado por la ecuación 1.40

$$\text{Dunn} = \frac{\min\limits_{1\leq i\leq j\leq q} d(C_i, C_j)}{\max\limits_{1\leq k\leq q} \text{diam}(C_k)} \tag{1.40}$$

donde $d(C_i; C_j)$ es la función de disimilitud entre dos grupos $C_i$ y $C_j$ definidos como $d(C_i, C_j) = \min\limits_{x\ \epsilon C_i, y\ \epsilon C} d(x, y)$ y $\text{diam}(C)$ es el diámetro de un grupo, que puede considerado como una medida de la dispersión del racimo. El diámetro de un cluster $C$ se puede definir usando la ecuación 1.41

$$\text{Diam}(C) = \max_{x,y\ \epsilon\ C} d(x, y) \tag{1.41}$$

Si el conjunto de datos contiene grupos compactos y bien separados, el diámetro de los clusters es se espera que sea pequeña y se espera que la distancia entre los clusters sea grande. Así, El índice de Dunn debe maximizarse.

### 1.4.28. Hubert statistic

El estadístico de Hubert (L. Hubert y Arabie, 1985) es el coeficiente de correlación serial puntual entre dos matrices cualesquiera. Cuando las dos matrices son simétricas, $\Gamma$ se puede escribir en su forma bruta como mostrado por la ecuación 1.42

$$\Gamma(P, Q) = \frac{1}{N_t} \sum_{\substack{i=1 \\ i<j}}^{n-1} P_{ij} Q_{ij}, \tag{1.42}$$

donde

- $P$ es la matriz de proximidad del conjunto de datos,
- $Q$ es una matriz $n \times n$ cuyo elemento $(i, j)$ es igual a la distancia entre los representantes

puntos nativos $(v_{ci}, v_{cj})$ de los clusters a los que pertenecen los objetos $xi$ y $xj$.

Observamos que para $q = 1$ o $q = n$, el índice no está definido.
La definición del estadístico normalizado $\Gamma$ de Hubert viene dada por la ecuación 1.43

$$\bar{\Gamma} = \frac{\sum_{\substack{i=1 \\ i<j}}^{n-1}(P_{ij} - \mu_P)(Q_{ij} - \mu_Q)}{\sigma_P \sigma_Q}, \tag{1.43}$$

donde $\mu_P, \mu_Q, \sigma_P, \sigma_Q$ son las respectivos medias y varianzas de la $P$ y $Q$ matrices.
Este índice toma valores entre $-1$ y 1. Si $P$ y $Q$ no son simétricos, entonces todas las sumas se extienden sobre todas las entradas $n^2$ y $N_t = n^2$ (Bezdek y Pal 1998).

Los valores altos de las estadísticas $\Gamma$ normalizadas indican la existencia de conglomerados compactos. Así, en la gráfica de $\Gamma$ normalizada versus $q$ ($q$ es el número de agrupaciones), buscamos una rodilla significativa que corresponde a un aumento signi cativo de $\Gamma$ normalizado cuando $q$ varía de 2 a $q_max$, donde $q_max$ es el número máximo posible de clústeres. El número de grupos en los que la rodilla ocurre es una indicación del número de agrupaciones que subyacen a los datos (Halkidi, Batistakis, y Vazirgiannis 2001).
En el paquete **NbClust**, los valores de las segundas diferencias de las estadísticas $\Gamma$ normalizadas se grafican en ayudan a distinguir la rodilla de otras anomalías. Un pico signicativo en este gráfico indica el número óptimo de clústeres.

### 1.4.29. SDindex

La definición del índice de validez de **SD** se basa en los conceptos de dispersión *promedio para conglomerados* y *separación total entre grupos*. Se calcula usando la ecuación 1.44

$$\text{SDindex}(q) = \alpha \text{Scat}(q) + \text{Dis}(q) \tag{1.44}$$

El primer termino, Scat(q), se calcula usando la ecuación 1.45 indica el valor compacto promedio de grupos (es decir, distancia intra-grupo). Un valor pequeño para este término

indica grupos compactos.

$$\text{Scat}(q) = \frac{\frac{1}{q}\sum_{k=1}^{q}\left\|\sigma^{(k)}\right\|}{\|\sigma\|} \tag{1.45}$$

donde

- $\sigma$ es el vector de varianzas para cada variable en el conjunto de datos,
- $\sigma = (\text{var}(V_1), \text{var}(V_2), \ldots, \text{var}(V_p))$
- $\sigma^k$ es el vector de varianza para cada cluster $C_k$,
- $\sigma^k = (\text{var}(V_1^{(k)}), \text{var}(V_2^{(k)}), \ldots, \text{var}(V_p^{(k)}))$

El segundo término Dis(q), calculado usando la Ecuación 1.46, indica la separación total entre los grupos q (es decir, un indicador de la distancia entre grupos).

$$\text{Dis}(q) = \frac{D\text{max}}{D\text{min}}\sum_{k=1}^{q}\left(\sum_{z=1}^{q}\|c_k - c_z\|\right)^{-1} \tag{1.46}$$

donde

- $D_{\text{max}} = \max(\|c_k - c_z\|)\forall k, z\ \epsilon\{1, 2, 3, \ldots, q\}$ es la distancia máxima entre el clúster centros,
- $D_{\text{min}} = \min(\|c_k - c_z\|)\forall k, z\ \epsilon\{1, 2, 3, \ldots, q\}$ es la distancia minima entre el clúster centros,

$\alpha$ es un factor de ponderación igual a Dis($q_{max}$) donde $q_max$ es el número máximo de entradas clusters. El número de conglomerados, $q$, que minimiza el índice anterior, se puede considerar como un valor óptimo para el número de conglomerados presentes en el conjunto de datos.

### 1.4.30. Índice SDbw

La definición del índice de validez SDbw se basa en los criterios de compacidad y separación entre grupos. Se calcula usando la ecuación 1.48.

$$\text{SDbw}(q) = \text{Scat}(q) + \text{Density.bw}(q) \tag{1.47}$$

El primer término, Scat(q), es el mismo calculado en SDindex Ecuación 1.44.
El segundo término, Density.bw(q), es la densidad entre conglomerados. Evalúa la densidad media en la región entre conglomerados en relación con la densidad de los conglomerados y se calcula utilizando la ecuación 1.48

$$\text{Density.bw}(q) = \frac{1}{q(q-1)} \sum_{i=1}^{q} \left( \sum_{j=1, i \neq j}^{q} \frac{\text{densidad}(u_{ij})}{\max(\text{densidad}(c_i), \text{densidad}(c_j))} \right), \tag{1.48}$$

donde

- $u_{ij}$ es el punto medio del segmento de línea de nido por los centroides de los clústeres $c_i$ y $c_j$,
- density$(u - ij)$ is calculated using Equation 1.49.

$$\text{densidad}(u_{ij}) = \sum_{l=1}^{n_{ij}} f(x_l, u_{ij}), \tag{1.49}$$

donde

- $n_{ij}$ es el número de tuplas que pertenecen a los clústeres $C_i$ y $C_j$,
- $f(x_l, u_{ij})$ es igual a 0 si $d(x, u_{ij}) >$ Stdev y 1 en cualquier otro caso,
- Stdev, definida en la ecuación 1.50, es la desviación estándar promedio de los conglomerados.

$$\text{Stdev} = \frac{1}{q} \sqrt{\sum_{k=1}^{q} \|\sigma^{(k)}\|} \tag{1.50}$$

El número de clusters $q$ que minimiza SDbw se considera el valor óptimo para la número de grupos en el conjunto de datos (Halkidi y Vazirgiannis 2001).

Como se mencionó anteriormente, el número óptimo de clusters seleccionados por **NbClust** para cada índice es basado en valores máximos (o mínimos) del índice, diferencia máxima (o mínima) entre los niveles de jerarquía del índice ($\max_q(i_q - i_{q-1})$) , $q$ es el número de conglomerados y $i_q$ es el valor del índice para q clusters), valor máximo (o mínimo) de segundas diferencias entre niveles del índice ($\max_q(i_{q+1} - i_q) - (i_q - i_{q-1})$)) o por el uso de un valor crítico como en el caso del índice Gap y el índice Beale.
Si la medida aumenta a medida que aumenta el número de conglomerados, como en el caso de Díndice y el índice de Hubert, entonces simplemente encontrar el mínimo o el máximo en un gráfico ya no es suficiente. En cambio, un cambio local significativo en el valor de la medida, visto como un "codo" en el gráfico, indica los mejores parámetros para la agrupación.

### 1.4.31. Método Elbow

Este método tiene la intención de encontrar el numero de clusters óptimo, y la idea es evaluar al igual que ensayar un rango de valores para K que es el numero de clusters. Este método maneja de manera gráfica la representación de las observaciones que se agrupan de una forma tal que se minimiza la varianza total intracluster en función del número de clusters(K) y escoge como el cluster óptimo aquel que su diferencia al agregar K+1 apenas se note una mejoría.

En el paquete **NbClust**m , la rodilla es detectada por un pico local en el gráfico de segundas diferencias. entre niveles del índice. Por lo tanto, el número adecuado de grupos se elige mediante inspección visual. de la segunda parcela de diferencias. La ausencia de tal rodilla podría ser una indicación de que el conjunto de datos no posee una estructura de agrupamiento. La Tabla 2.1 resume los índices incluidos en el paquete **NbClust** . Da

el nombre de cada índice en referencias y en el paquete **NbClust** , y cómo seleccionar el número óptimo de clústeres.

## 1.5. Analisis de correspondencias

El AC Simple, o simplemente AC, es una herramienta para analizar las asociaciones entre las filas y columnas de una tabla de contingencia

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1q} \\ X_{12} & X_{22} & \cdots & X_{2j} & \cdots & X_{2q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{iq} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{pq} \end{bmatrix}_{(p\times q)} \tag{1.51}$$

### 1.5.1. Estadística $\chi^2$

Una forma de medir asociación entre las filas y columas de una tabla de contingencias es a través de la estadística.

$$\chi^2 = \sum_{i=1}^{p}\sum_{i=1}^{q}\left(\frac{X_{ij} - E_{ij}}{E_{ij}}\right)^2 \tag{1.52}$$

donde, $E_{ij} = \frac{\mathrm{x}_{i.}\mathrm{x}_{.j}}{\mathrm{x}_{..}}$ bajo la hipótesis de independencia $\chi^2 \sim \chi^2_{(p-1,q-1)}$

### 1.5.2. Análisis de Correspondencias Múltiples

- Extensión del Análisis de Correspondencias a más de dos variables categóricas.
- Considere un conjunto de $n$ individuos y $p$ variables. Cada variable con $c_j$ modalidades, $j = 1p$.
- El método consiste de un $AC$ a una matriz indicadora $X$ con elementos.

$$X_{ik} = \begin{cases} 1, & \text{si el individuo } i \text{ seleccinó la categoría } K \\ 0, & \text{en otro caso} \end{cases} \tag{1.53}$$

con $i = 1, \ldots, n,\ k = 1, \ldots, m$ y $m = \sum_{j=1}^{p} c_j$

### 1.5.3. Matriz indicadora

$$\left[\begin{array}{cccc|ccc|c|}
X_{11} & \cdots & X_{1c_1} & | & X_{1_{(c_1+1)}} & \cdots & X_{1_{(c_1+c_2)}} & | \cdots & | \\
X_{1m} & & & & & & & & \\
X_{21} & \cdots & X_{2c_2} & | & X_{2_{(c_1+1)}} & \cdots & X_{2_{(c_1+c_2)}} & | \cdots & | \\
X_{2m} & & & & & & & & \\
\vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots & | & | \\
\vdots & & & & & & & & \\
X_{n1} & \cdots & X_{nc_1} & | & X_{n_{(c_1+1)}} & \cdots & X_{n_{(c_1+c_2)}} & | \cdots & | \\
X_{nm} & & & & & & & &
\end{array}\right] \tag{1.54}$$

- El análisis de correspondencias múltiple consistirá de un análisis de correspondencias sobre la matriz $X$ o sobre las matrices $X^T X$.
- La matriz $X^T X$ se conoce como matriz de Burt y fuera de su bloque diagonal contiene las tablas de contingencia entre pares de variables.
- Un análisis de correspondencias sobre $X$ es equivalente a un análisis de correspondencias sobre $X^T X$ o $X^T$.

## 1.6. Cuantificación de categorías o transformaciones

### 1.6.1. Análisis de componentes principales no lineal

Teniendo una matriz de datos $n \times m$ con variables métricas, el análisis de componentes principales (ACP) es untécnica común para reducir la dimensionalidad del conjunto de datos, es decir, para proyectar las variables en un subespacio $\mathbb{R}^p$ donde $p \ll m$. El teorema de Eckart-Young establece que esta forma clásica de PCA lineal se puede formular mediante una función de pérdida. Su minimización conduce a una $n \times p$ matriz de

*puntuaciones de componentes* y una matriz $m \times p$ de *cargas de componentes*.
Sin embargo, al tener variables no métricas, se puede utilizar PACP no lineal (ACPNL). El término "no lineal" pertenece a las transformaciones no lineales de las variables observadas (de Leeuw 2006). En Gi, ACPNL puede definirse como análisis de homogeneidad con restricciones en el matriz de cuantificación $Y_j$. Denotemos $r_j \leq p$ como el parámetro de la restricción impuesta en la variable $j$. Si no se imponen restricciones, como por ejemplo, para una solución simple de homals, $rj = kj1$ iff $kj \leq p$, y $rj = p$ en caso contrario.
El caso simple de $r_j = 1$ para todo $j$. En este caso decimos que todas las variables *son únicas* y las restricciones de rango son impuestas por

$$Y_j = z_j a_j', \tag{1.55}$$

donde $z_j$ es un vector de longitud $k_j$ con cuantificaciones de categoría y $a_j$ un vector de longitud $p$ con pesas. Por lo tanto, cada matriz de cuantificación está restringida al rango 1, lo que permite existencia de puntuaciones de objetos con una cuantificación de categoría única.

### 1.6.2. Múltiples cuanticaciones

No es necesario que restrinjamos el rango de la matriz de puntuación a 1. Homals permite múltiples restricciones de rango. Simplemente podemos extender la Ecuación 1.55 a el caso general

$$Y_j = Z_j A_j' \tag{1.56}$$

donde de nuevo $1 \leq r_j \leq \min(k_j - 1, p)$, $Z_j$ es $k_j \times r_j$ y $A_j$ es $p \times r_j$. Se requiere, sin perdida de generalidad, que $Z_j' D_j Z_j = I$. Así, tenemos la situación de *cuantificaciones múltiples* que implica imponer una restricción adicional cada vez que se realiza un ACP.

Para establecer la función de pérdida para la versión con restricción de rango, escribimos

$r_\star$ por la suma del $r_j$ y $r_\bullet$ para su promedio. La matriz de bloques $G$ de variables ficticias ahora se convierte en

$$Q \stackrel{\triangle}{=} \left[ G_1 Z_1 \ \vdots \ G_2 Z_2 \ \vdots \ \cdots \ G_m Z_m \right]. \tag{1.57}$$

Reuniendo los $A_j's$ en una matriz de bloques también, el $p \times r_\bullet$ matriz.

$$A \stackrel{\triangle}{=} \left[ A_1 \ \vdots \ A2 \ \vdots \ \cdots \ A_m \right] \tag{1.58}$$

Entonces

$$\begin{aligned} \sigma(X; Z; A) = \textstyle\sum_{j=1}^{m} \mathrm{tr}(G_j Z_j A_j')' M_j (X - G_j Z_j A_j') = \\ = m \ \mathrm{tr}(X' M_\star X) - 2\mathrm{tr}(X' Q A) + \mathrm{tr}(A' A) = \\ = mp \ \mathrm{tr}(Q_X A)'(Q - XA) - \mathrm{tr}(Q'Q) = \\ = \mathrm{tr}(Q - XA)'(Q - XA) + m(p - r_\bullet) \end{aligned} \tag{1.59}$$

Esto muestra que $\sigma(X; Y1; , \ldots, Ym) \geq m(p - r_\bullet)$ y la pérdida es igual a este límite inferior si puede elegir $Z_j$ tal que $Q$ sea de rango $p$. De hecho, al minimizar 1.59 sobre $X$ y $A$ vemos ese

$$\sigma(Z) \stackrel{\triangle}{=} \min_{X,A} \sigma(X; Z; A) = \sum_{s=p+1}^{r_*} \lambda_s^2(Z) + m(p - r_\bullet), \tag{1.60}$$

donde $\lambda_s$ son los valores singulares ordenados.

### 1.6.3. Restricciones de nivel: escalamiento óptimo

Desde un punto de vista general, *el escalado óptimo* intenta hacer dos cosas simultáneamente: La transformación de los datos mediante una transformación apropiada para el nivel de escala (es decir, nivel restricciones), y la bondad de ajuste de un modelo a los datos transformados para dar cuenta de los datos. Así es un proceso simultáneo de transformación y representación de datos (Takane, 2005). En este trabajo tendremos en cuenta el nivel de escala de las variables en términos de restricciones dentro de $Z_j$. Para hacer esto, el punto

de partida es dividir la Ecuación 1.59 en dos términos separados. Usando $\hat{Y}_j = D_j^{-1} G_j' X$ esto conduce a

$$\begin{aligned} &\textstyle\sum_{j=1}^{m} \operatorname{tr}(X - G_j Y_j)' M_j (X - G_J Y_J) \\ &= \textstyle\sum_{j=1}^{m} \operatorname{tr}(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j)))' M_j (X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j))) \\ &= \textstyle\sum_{j=1}^{m} \operatorname{tr}(X - G_j \hat{Y}_j)' M_j (X - G_j \hat{Y}_j) + = \sum_{j=1}^{m} \operatorname{tr}(Y_j - \hat{Y}_j)' D_j (Y_j - \hat{Y}_j). \end{aligned} \tag{1.61}$$

Obviamente, las restricciones de rango $Y_j = Z_j A_j'$ afectar sólo el segundo término y por lo tanto, vamos a Proceda con nuestras explicaciones con respecto a este término en particular, es decir,

$$\sigma(Z; A) = \sum_{j=1}^{m} \operatorname{tr}(Z_j A_j' - \hat{Y}_j)' D_j (Z_j A_j' - \hat{Y}_j). \tag{1.62}$$

Ahora, se pueden imponer restricciones de nivel para variables nominales, ordinales, polinomiales y numéricas en $Z_j$ de la siguiente manera. Para las variables nominales, todas las columnas en $Z_j$ no están restringidas. La ecuación 1.62 se minimiza en las condiciones $u' D_j Z_j = 0$ y $Z_j' D_j Z_j = I$. Las ecuaciones estacionarias son:

$$\begin{aligned} A_j &= Y' D_j Z_j, \\ Y_j A_j &= Z_j W + \mathrm{uh}', \end{aligned} \tag{1.63}$$

con $W$ como una matriz simétrica de multiplicadores de Langrange. Resolviendo,se encontra

$$\mathrm{h} = \frac{1}{\mathrm{u'}D_j \mathrm{u}} A_j' Y_j' D_j \mathrm{u} = 0, \tag{1.64}$$

y así, dejar $\bar{Z}_j \triangleq D_j^{1/2} Z_j$ y $\bar{Y} \triangleq D_j^{1/2} Y_j$, resulta que

$$\bar{Y}_j \bar{Y}_j{}' \bar{Z}_j = \bar{Z}_j W. \tag{1.65}$$

Si $\bar{Y} j = K \Delta L'$ es la SVD de $\bar{Y} j$ , entonces vemos que $\bar{Z} j = K_r O$ con O como una rotación

arbitraria matriz y $K_r$ como los vectores singulares correspondientes a los $r$ valores singulares más grandes. Así, $Z_j = D_j^{1/2} K_r O$ y $A_j = \bar{Y_j}' \bar{Z_j} = L_r \Delta_r O$. Además, $Z_j A_j' = D_j^{-1/2} K_r \Delta_r L_r'$. Teniendo variables ordinales, la primera columna de $Z_j$ está restringida a ser creciente o decreciendo, el resto es gratis. Nuevamente 1.62 debe minimizarse bajo la condición $Z'jD_jZ_j = I$ (y opcionalmente condiciones adicionales en $Z_j$). Si minimizamos sobre $A_j$, también podemos resolver el problema $\text{tr}(Z_j' D_j Y_j Y_j' D_j Z_j)$. con $Z_j' D_j Z_j = I$ Para las restricciones polinomiales, la matriz $Z_j$ son los primeros polinomios ortogonales $r_j$. Así todo $p$ columnas de $Y_j$ son polinomios de grado $r_j$. En el caso de las variables numéricas, la primera La columna en $Z_j$ denotada por $z_j$ es fija y lineal con los números de categoría, el resto es libre. Por tanto, la función de pérdida en 1.62 cambia a

$$\sigma(Z, A) = \sum_{j=1}^{m} \text{tr}(Z_j A_j' + z_{j0} a_{j0}' - \hat{Y_j})' D_j (Z_j A_j' + z_{j0} a_{j0}' - \hat{Y_j}). \tag{1.66}$$

Como la columna $z_{j0}$ está fija, $Z_j$ es una matriz $k_j \times (r_j - 1)$ y $A_j$, con $a_{j0}$ como primera columna, es $p \times (r_j - 1)$. Para minimizar 1.66, $z_{j0}' D_j Z_j = 0$ es necesario como condición de minimización. Tenga en cuenta que las restricciones de nivel se pueden imponer además de las restricciones de rango. Si el conjunto de datos tiene variables con diferentes niveles de escala, el paquete **homals** permite establecer restricciones de nivel para cada variable j por separado. A diferencia de (Gifi, 1990) y Michailidis y de Leeuw (1998), No es necesario tener restricciones de rango 1 para permitir diferentes niveles de escala. Nuestra La implementación permite múltiples restricciones de nivel ordinales, numéricas múltiples, etc.

### 1.6.4. Análisis de correlación canónica no lineal

En la terminología Gifi, el análisis de correlación canónica no lineal (NLCCA) se llama "OVERALS' " (van der Burg, de Leeuw y Verdegaal 1988; van der Burg, de Leeuw y Dijksterhuis 1994). Esto se debe al hecho de que tiene la mayoría de los otros modelos Gi como casos especiales.

En esta sección se muestra la relación con el análisis de homogeneidad. El paquete **homals** permite de nición de conjuntos de variables y, por tanto, para el cálculo NLCCA entre $g = 1, \ldots, K$ conjuntos de variables.

Recuerde que el objetivo del análisis de homogeneidad es encontrar $p$ vectores ortogonales en $m$ indicador matrices $G_j$. Este enfoque se puede ampliar para calcular $p$ vectores ortogonales en $K$ matrices generales $G_v$, cada una de dimensión $n \times m_v$ donde $m_v$ es el número de variables $(j = 1, \ldots, mv)$ en el conjunto $v$. Así,

$$G_v \triangleq \left[ G_{v1} \ \vdots \ G_{v2} \ \vdots \ \cdots \ G_{vmv} \right]. \tag{1.67}$$

La función de pérdida se puede establecer como

$$\sigma(X; Y_1, \ldots, Y_k) \triangleq \frac{1}{K} \sum_{v=1}^{K} \operatorname{tr} \left( X - \sum_{j=1}^{m_v} G_{vj} Y_{vj} \right)' M_v \left( X - \sum_{j=1}^{m_v} G_{vj} Y_{vj} \right). \tag{1.68}$$

$X$ es la matriz $n \times p$ con puntuaciones de objetos, $G_{vj}$ es $n \times k_j$ y $Y_{vj}$ es la matriz $k_j \times p$ que contiene las coordenadas. Los valores faltantes se tienen en cuenta en $M_v$, que es el elemento mínimo de $M_j$ en el conjunto $v$. Las condiciones de normalización son $XM_\bullet X = I$ y $u'M_\bullet X = 0$ donde $M_\bullet$ es el promedio de $M_v$.

Dado que NLPCA se puede considerar como un caso especial de NLCCA, es decir, para $K = m$, todos los Las restricciones para diferentes niveles de escala se pueden aplicar directamente para NLCCA. diferente a análisis de correlación canónica clásica, NLCCA no se limita a dos conjuntos de variables, sino permite la definición de un número arbitrario de conjuntos. Además, si los conjuntos se tratan de manera asimétrica modelos predictivos como análisis de regresión y discriminante el análisis se puede emular. Para $v = 1, 2$ establece esto implica que $G1$ es $n \times 1$ y $G2$ es $n \times m - 1$.

Después de analizar las variables utilizadas en cada factor, se procede a realizar un análisis de homogeneidad, este también es conocido como análisis de correspondencia

múltiple,y normalmente se busca la reducción de la dimensionalidad, pero en este caso se utilizamos para convertir variables categóricas en variables numéricas, y en el trasfondo estoy utilizando todas las variables numéricas, ya sean ordinales o nominales, y para cada varible categorica ; por medio de la función **homals** que nos permite realizar cuantificaciones de categoría (o transformaciones), es decir, se convierte las variables categóricas por medio de variables numéricas, al final utilizamos las cuantificaciones basadas únicamente en los efectos o pesos principales. En el capitulo 3, se observan los ejemplos práctico de este pensamiento.

# Hipótesis

Los modelos de planteados por entidades financieras(en especial Bancos), realizan las segmentaciones para los diferentes factores de riesgo, hoy en día se tienen los siguientes factores de riesgo:

- Clientes
- Productos
- Jurisdicciones
- Canales de distribución

En este documento se observaron 2 entidades financieras con los resultados anonimizados, dado a motivos de confidencialidad; por otro lado las metodologías son las propuestas por cada entidad. La hipótesis inicial para este documento es utilizar un análisis de homogeneidad, o también conocido como análisis de correspondencia múltiple para las variables categóricas presentes por cada factor de riesgo, es por esto que realizaremos restricciones de rango en las cuantificaciones de categoría (o transformaciones) y restricciones de nivel (lo que permite tratar una variable como nominal, ordinal o numérica); todo esto, para poder utilizar las variables categóricas de una manera numérica(pesos por categoría), y así poder obtener un mejor resultado.

Por otro lado, a efectos prácticos y dada la cantidad de datos dentro de una entidad financiera, el mejor algoritmo para el calculo propuesto a manera de hipótesis es CLARA, y este sera contrastado con los métodos utilizados por las entidades en el factor de riesgo clientes, por que este es el que mas datos tiene y por lo tanto el ideal para probar la metodología planteada en este documento.

# Capítulo 2

# Metodología

## 2.1. Estadísticas descriptivas

Se realizó un análisis descriptivo, para tener un primer acercamiento de los datos, detectar si existen datos vacíos o datos erróneos, conocer la naturaleza de las variables (cualitativas o cuantitativas), obtener estadísticas para las variables cuantitativas como la media, la medianas, moda la desviación estándar, los quartiles etcétera, y para variables cualitativas, conocer cuantas categorías tiene cada variable, cual es la categoría más frecuento o menos frecuente, etcétera.

Antes de realizar una segmentación es fundamental realizar un análisis descriptivo para cada factor de riesgo. Cuando estemos en el capitulo 3, se abordará este tema de manera detallada, dado a que es la parte más importante en el documento, por que si se falla en este análisis todo lo demás sera erróneo; y se mostrara a manera de ejemplo, como se realizar un buen análisis descriptivo para una buena segmentación.

## 2.2. Cuantificaciones de categorías o transformaciones

Gigi, 1990 proporciona varias extensiones del análisis de homogeneidad y elabora conexiones a otros métodos multivariantes. El paquete **holmals** [1] permite imponer restricciones en los rangos y niveles de variables, así como en la definición de conjuntos de variables. Estas opciones ofrecen un amplio espectro de posibilidades adicionales para el análisis de datos multivariados más allá de la homogeneidad clásica análisis (cf. visión de sentido amplio en la Introducción).

## 2.3. Escalar las variables

Después de transformar las variables categóricas, notamos que escalar cada variable nos arroja mejores resultados por lo que se procede a utilizar la función , la cual realiza la siguiente operación haciendo la siguiente operación:

$$VS(x)_{ij} = \frac{x_{ij} - \bar{X}_j}{Sd_j} \qquad Sd_j = \frac{(x_{ij} - \bar{X}_{ij})^2}{n-1} \qquad \bar{X}_j = \frac{\sum_{i=1}^{n} x}{n} \tag{2.1}$$

## 2.4. Calcular el número de cluster

Para calcular el numero de cluster recurrimos al paquete **NbClust** [2], el cual nos proporciona 30 índices, entre los que estén medidas de distancia y métodos de iteración, para así determinar el número de clusters.

[1] `https://cran.r-project.org/web/packages/homals/index.html`
[2] `https://cran.r-project.org/web/packages/NbClust/NbClust.pdf`

| Nombre del índice en NBClust | Número optimo de clusters |
|---|---|
| 1. “ch ” (Caliński y Harabasz, 1974) | Max. valor del índice |
| 2. “duda ” (Duda, Hart, y cols., 1973) | Menor clústeres tal que el índice >valor crítico |
| 3. “pseudot2 ” (Duda y cols., 1973) | Menor de clústeres tal que el índice <valor crítico |
| 4. “cindex ” (L. J. Hubert y Levin, 1976) | Min. valor del índice |
| 5. “gamma ” (Baker y Hubert, 1975) | Max valor del índice |
| 6. “beale ” (Beale, 1969) | Número de clusters tal que el valor crítico sea $\geq \alpha$ |
| 7. “ccc ” (Sarle, 1983) | Max valor del índice |
| 8. “ptbiserial ” (Milligan, 1980) | Max valor del índice |
| 9. “gplus ” (Milligan, 1981) | Min valor del índice |
| 10. “ db ” (Davies y Bouldin, 1979) | Min valor del índice |
| 11. “ frey” (Frey y Van Groenewoud, 1972) | Nivel de clúster antes valor de índice <1.00 |
| 12. “hartigan ” (Hartigan, 1975) | Max. Dif entre niveles de jerarquía del índice |
| 13. “ tau ” (Rohlf, 1974) | Max valor del índice |
| 14. “ ratkowsky ” (Ratkowsky, 1978) | Máximo valor del índice |
| 15. “ scott ” (Scott y Symons, 1971) | Dif. max entre niveles de jerarquía del índice |
| 16. “marriot ” (Marriott, 1971) | Max. valor de 2das dif entre niveles del índice |
| 17. “ball ” (Ball y Hall, 1965) | Dif. max entre niveles de jerarquía del índice |
| 18. “trcovw ” (Milligan y Cooper, 1985) | Dif. max entre niveles de jerarquía del índice |
| 19. “tracew ” (Milligan y Cooper, 1985) | Max. valor de 2das dif entre niveles del índice |
| 20. “ friedman ” (Friedman y Rubin, 1967) | Dif. max entre niveles de jerarquía del índice |
| 21. “mcclain ” (McClain y Rao, 1975) | Min valor del índice |
| 22. “rubin ” (Friedman y Rubin, 1967) | Valor mínimo de 2das dif entre niveles |
| 23. “kl ” (Krzanowski y Lai, 1988) | Max valor del índice |
| 24. “silhouette ” (Rousseeuw, 1987) | Max valor del índice |
| 25. “gap ” (Tibshirani, 2001) | Menor número clusters tal que valor crítico $\geq 0$ |
| 26. “dindex ” (Lebart, 2000) | Método gráfico |
| 27. “dunn ” (Dunn, 1974) | Max valor del índice |
| 28. “ hubert ” (L. Hubert y Arabie, 1985) | Método gráfico |
| 29. “sdindex ” (Halkidi, 2000) | Min valor del índice |
| 30. “sdbw ” (Halkidi y Vazirgiannis, 2001) | Min valor del índice |

**Tabla 2.1:** Resumen de los índices implementados en el paquete NbClust.

## 2.5. Aplicación de CLARA

Este paso es el mas sencillo, dado a que se aplica el algoritmo que en el capitulo 1 ya definimos y podemos obtener resultados. En este paso, para poder comparar resultados obtenidos en los diferentes escenarios propuestos en el capitulo 3.

# Capítulo 3

# Aplicación practica

En este capitulo se revisaron dos escenarios diferentes, los cuales muestran las segmentaciones realizadas para el factor de riesgo clientes, por parte de dos diferentes Entidades, encontraste lo realizado por KPMG.

## 3.1. Escenario 1

### 3.1.1. Resultados de la Entidad 1

Para este escenario revisaremos los resultados obtenidos por la Entidad Financiera 1 (EF1), para el factor de riesgo "Clientes", el cual contiene 3´340.291 registros(filas) por 5 variables(columnas) las cuales son las siguientes:

- Frecuencia de transacciones
- Ciudad de origen
- Transacciones:
    - Efectivo (Suma Ingresos, Egresos y patrimonio).
    - Cheque (Suma Ingresos, Egresos y patrimonio).
    - Transferencia (Suma Ingresos, Egresos y patrimonio).

Los resultados obtenidos por la EF1, por medio de la metodología de tablas cruzadas, están dispuestos en la tabla 3.1 de manera que en la primera columna se encuentran los grupos, y para el resto de columnas los promedios(Prom) por cada una de las variables.

| Cluster | Prom Efectivo | Prom Cheque | Prom Transferencia |
|---|---|---|---|
| Bogotá D.C | $ 537.307.062.325 | $ 81.297.763.155 | $ 2.402.878.305 |
| Medellin | $ 47.529.986.320 | $ 1.327.417.837 | $ 934.758.707 |
| Otras Ciudades | $ 1.242.898.153 | $ 723.894.577 | $ 8.235.311.355 |

**Tabla 3.1:** Tabla de resultados de la Entidad 1

La metodología de tablas cruzadas para este caso, fue filtrar por cada ciudad y promediar el Efectivo, Cheque y Transferencia; lo cual, no es considerado como un trabajo de segmentación óptimo, dado a que no estamos garantizando de que las operaciones en cada segmento o cluster, sean lo mas similares en el cluster y lo más diferentes entre los clusters. Esto es un claro ejemplo, en el que el cumplimiento de la norma es diferente al resultado de la misma, dado a que se realizaron segmentos por Ciudad, pero estas no me garantizan que sean independientes de las otras en forma de cluster.

Además de la metodología utilizada, resulta carente de índices y técnicas que garanticen una óptima segmentación, y tenemos que tener en cuenta, que los resultados obtenidos por la EF1, carecen de un análisis descriptivo que es necesario en una base de datos, al menos en la documentación suministrada; al igual, que no se detalla ninguna alusión a la calidad de los datos, ni una revisión de datos atípicos, o la tipificación los mismos, como objetivo de esta segmentación.

### 3.1.2. Resultados desde una metodología clásica

Dado el trabajo básico y documentación tan escasa entregada por la Entidad, se decide realizar un segmentación alterna utilizando K-Means, que es una de las mas frecuentes dentro del los análisis de segmentación. Utilizando los datos suministrados por la entidad:

| Variable | Media | Mediana | SD |
|---|---|---|---|
| Efectivo | $ 254.844.996.020 | $ 293.039.973.399 | $ 95.142.131.847 |
| Cheque | $ 33.072.697.259 | $ 41.674.537.784 | $ 17.840.473.643 |
| Transferencia | $ 2.916.036.132 | $ 5.786.474.183 | $ 1.651.320.155 |
| Transacciones | 1381 | 4029 | $ 789 |

**Tabla 3.2:** Tabla de descriptiva para las variables cuantitativas la Entidad 1

| Variable | Número de categorías | Más frecuente | Frecuencia |
|---|---|---|---|
| Ciudad | 17 | BOGOTA.D.C | 76988 |

**Tabla 3.3:** Tabla de descriptiva para las variables cualitativas la Entidad 1

Es necesario realizar algunas estadísticas básicas, con el objetivo de mirar cómo es el comportamiento, y si existe alguna anomalía en las mismas. Para este caso no se evidencia error o inconsistencias en la información suministrada . Una vez realizado un descriptivo podemos seguir con la segmentación, y se aclara que no se realiza ningún análisis con valores atípicos dado a que, se quiere identificar esos atípicos con el modelo de segmentación. Y en este caso, utilizando solamente las variables numéricas, dado que las entidades financieras frecuentemente no utilizan variables categóricas en metodologías como K-Means, K-Mediods, entre otras.

Los resultados obtenidos por **NbClust**, la cual nos permite calcular en número óptimo de clusters son los siguientes:

| Índice | Clusters | Value |
|---|---|---|
| Hubert | 0 | 0 |
| Dindex | 0 | 0 |
| Frey | 2 | - |
| CCC | 3 | -46,345 |
| DB | 3 | 0,8988 |
| Silhouette | 3 | 0,1344 |
| Duda | 3 | 0,9742 |
| PseudoT2 | 3 | 25,6556 |
| Beale | 3 | 0,0639 |
| McClain | 3 | 0,0467 |
| Dunn | 3 | 0,1751 |
| SDindex | 3 | 26,9098 |
| Scott | 4 | 224,2248 |

**Tabla 3.4:** Tabla de índices parte 1

| Índice | Clusters | Value |
|---|---|---|
| Ball | 4 | 0,3399 |
| Marriot | 5 | 0,1499 |
| TrCovW | 5 | 0,0419 |
| TraceW | 5 | 0,033 |
| Rubin | 5 | -0,0086 |
| KL | 6 | 1,6055 |
| Hartigan | 6 | 14,0828 |
| CH | 7 | 31,2326 |
| Friedman | 7 | 0,3038 |
| Cindex | 7 | 0,3907 |
| Ratkowsky | 7 | 0,1506 |
| PtBiserial | 7 | 0,266 |
| SDbw | 7 | 0,2261 |

**Tabla 3.5:** Tabla de índices parte 2

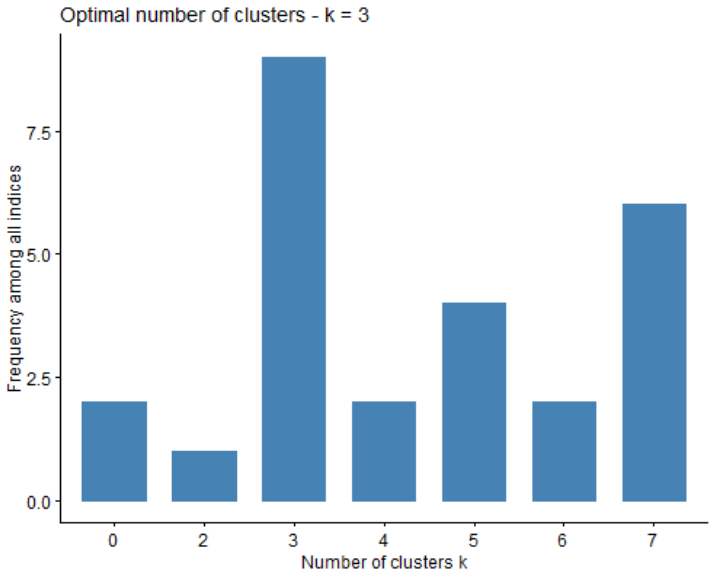



**Figura 3.1:** Histograma del numero de clusters óptimos

De los índices anteriores al igual que la gráfica obtenida, podemos decir que el número de clusters óptimos es 3, dado que es el más frecuente como se presenta en la Figura 3.1.

Los resultados de manera gráfica en dos dimensiones con el algoritmo de K-Means son los siguientes :

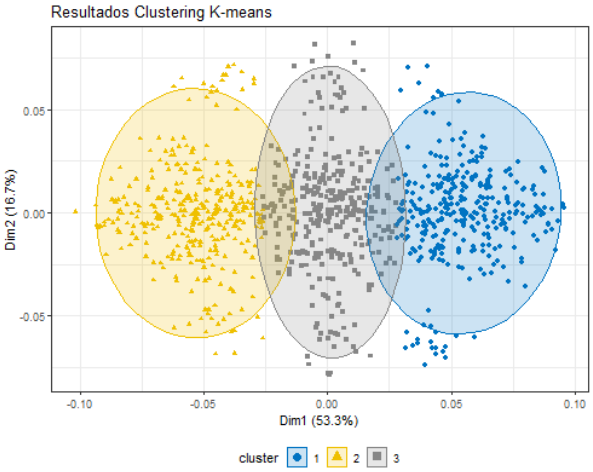



**Figura 3.2:** Gráfica de K-Means

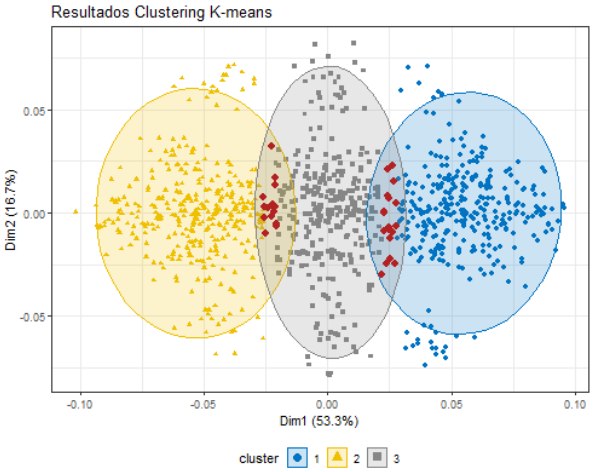



**Figura 3.3:** K-Means con atípicos

Esos atípicos son identificadas con el índice de *Silhouette*, que también puede indicar aquellas cuentas que al parecer no pertenecen a su cluster correspondiente, por lo que se ve en fronteras de los clusters circundantes en la Figura 3.3. Por otro lado, no es el objetivo identificar estos atípicos, dado a que este trabajo es propio de la EF1 y por cuestiones de privacidad de la información, no se puede mostrar datos que sirvan para identificar estas situaciones atípicas en este documento.

| Cluster | Prom Efectivo | Prom Cheque | Prom Transferencia |
|---|---|---|---|
| Cluster 1 | $ 419.683.390.680 | $ 102.825.410.838 | $ 1.575.407.541 |
| Cluster 2 | $ 130.533.263.211 | $ 1.978.918.080 | $ 2.837.595.926 |
| Cluster 3 | $ 3.606.393.280 | $ 915.581.860 | $ 4.713.892.219 |

**Tabla 3.6:** Tabla de resultados de EF1 por un método clásico

Los resultados de la segmentación son los siguientes:

- El número de clusters óptimos son 3, según las tablas 3.4 y 3.5, evidenciado también en la 3.1

- El tiempo gastado en esta segmentación es de 27.6 horas, en un computador de 32 giga-bytes de RAM. Como dato adicional se realizó un estudio de sensibilidad en las conclusiones.

### 3.1.3. Resultados desde la metodología propuesta

Utilizando los datos suministrados por la EF1, se realiza la metodología descrita en el Capitulo 2, iniciando con unas estadísticas descriptivas:

| Variable | Media | Mediana | SD |
|---|---|---|---|
| Efectivo | $ 254.844.996.020 | $ 293.039.973.399 | 35.142.131.847 |
| Cheque | $ 33.072.697.259 | $ 41.674.537.784 | 7.840.473.643 |
| Transferencia | $ 2.916.036.132 | $ 5.786.474.183 | 1.051.320.155 |
| Transacciones | 1381 | 4029 | 78 |

**Tabla 3.7:** Tabla de descriptiva para las variables cuantitativas la EF 1

| Variable | Numero de categorías | Mas frecuente | Frecuencia |
|---|---|---|---|
| Ciudad | 17 | BOGOTA.D.C | 1.602.376.988 |

**Tabla 3.8:** Tabla de descriptiva para las variables cualitativas la EF 1

Con respecto a la variable "Ciudad" que en este caso será tomada como una variable categórica con la naturaleza de ser un factor; utilizando la función **homals** ya descrita su funcionalidad en la sección 2.2, la cual permite realizar la transformación de la misma por medio de las diferentes variables numéricas mencionadas en la tabla 3.7.

Una vez realizada la transformación de la variable "Ciudad", según (MacQueen, 1967) mejora la segmentación si se escala las variables como se indica en en la sección 2.3.

Después de realizar todo lo anterior, sigue el paso de calcular el numero de clusters óptimos como se es presentado en la sección 2.4, por medio de la función **NbClust** y los resultados son los siguientes:

| Índice | Clusters | Value |
|---|---|---|
| Hubert | 0 | 0 |
| Dindex | 0 | 0 |
| Frey | 2 | - |
| CCC | 3 | -46,345 |
| DB | 3 | 0,8988 |
| Silhouette | 3 | 0,1344 |
| Duda | 3 | 0,9742 |
| PseudoT2 | 3 | 25,6556 |
| Beale | 3 | 0,0639 |
| McClain | 3 | 0,0467 |
| Dunn | 3 | 0,1751 |
| SDindex | 3 | 26,9098 |
| Scott | 4 | 224,2248 |

**Tabla 3.9:** Tabla de índices parte 1

| Índice | Clusters | Value |
|---|---|---|
| Ball | 4 | 0,3399 |
| Marriot | 5 | 0,1499 |
| TrCovW | 5 | 0,0419 |
| TraceW | 5 | 0,033 |
| Rubin | 5 | -0,0086 |
| KL | 6 | 1,6055 |
| Hartigan | 6 | 14,0828 |
| CH | 7 | 31,2326 |
| Friedman | 7 | 0,3038 |
| Cindex | 7 | 0,3907 |
| Ratkowsky | 7 | 0,1506 |
| PtBiserial | 7 | 0,266 |
| SDbw | 7 | 0,2261 |

**Tabla 3.10:** Tabla de índices parte 2

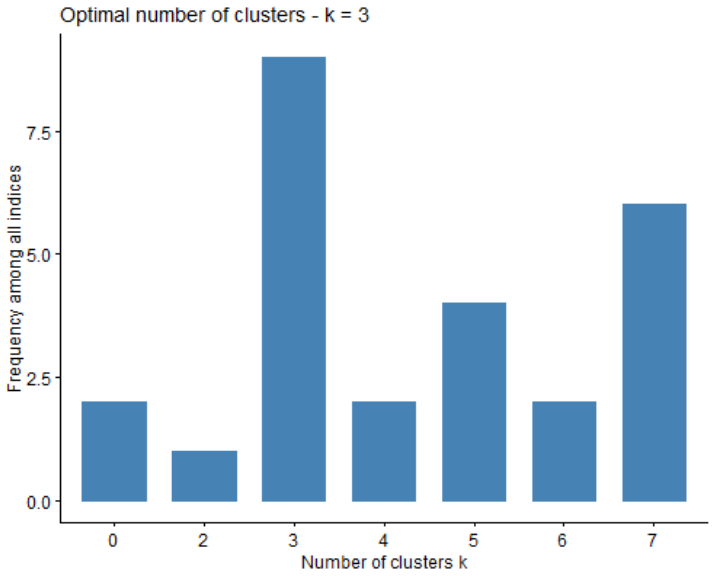



**Figura 3.4:** Histograma del numero de clusters óptimos

De los índices anteriores al igual que la gráfica obtenida, podemos decir que el número de clusters óptimos es 3, dado a que nos basamos en el mas frecuente como se presenta

en la Figura 3.4. Ya con los el número de clusters óptimos, se utiliza el algoritmo CLARA citado en sección 2.5 y explicado a detalle en la sección 1.4.3, de este algoritmo resulta una representación gráfica en dos dimensiones que es la siguiente:

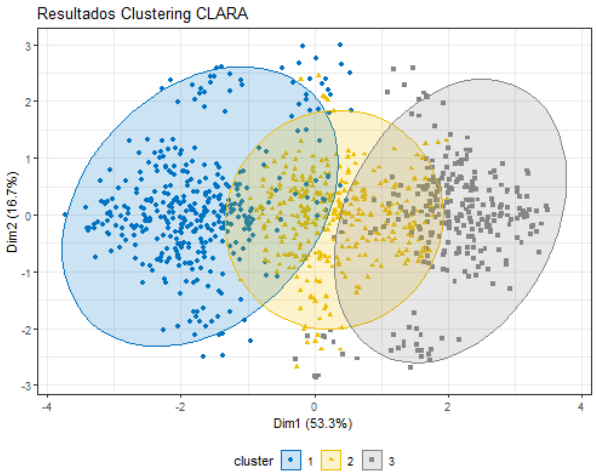



**Figura 3.5:** Escenario 1 - Gráfica de CLARA

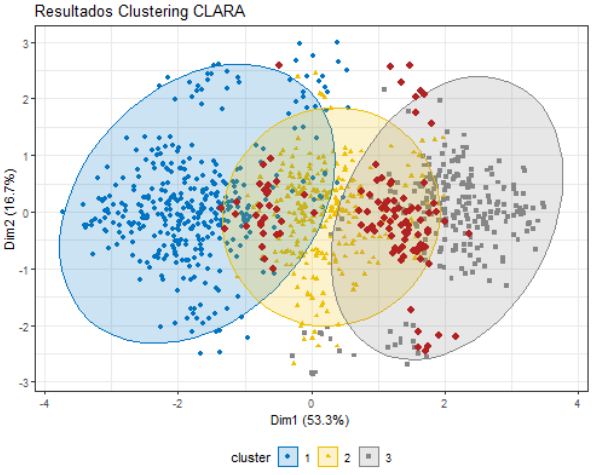



**Figura 3.6:** Escenario 1 - CLARA con atípicos

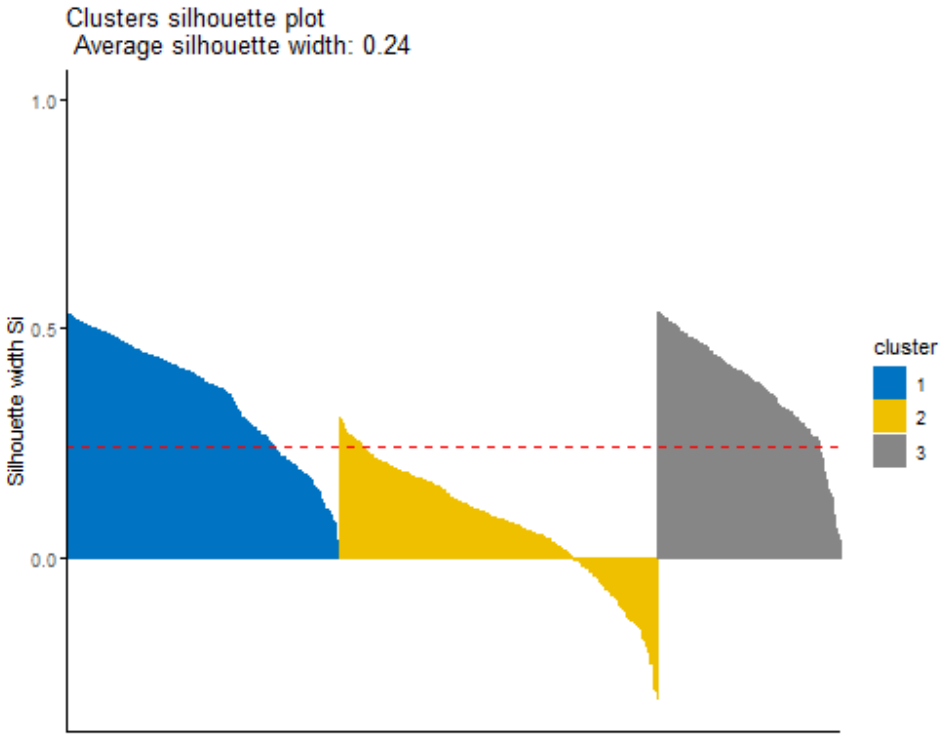



**Figura 3.7:** Escenario 1 -Atípicos Silhouette

Tener en cuenta que los atípicos son identificadas con el índice de Silhouette, este nos sirve identificar aquellas cuentas que al parecer no pertenecen a su cluster asignado, como se aprecia en la Figura 3.6 y Figura 3.7 , con las fronteras de los clusters circundantes. Se repite que no es el objetivo identificar estos atípicos, dado a que este trabajo tiene que ser propio de la Entidad financiera y por cuestiones de confidencialidad, no es posible mostrar las situaciones atípicas identificadas en este documento.

| Cluster | Efectivo | Cheque | Transferencia |
|---|---|---|---|
| Cluster 1 | $ 559.659.036.117 | $ 56.680.800.471 | $ 5.631.385.595 |
| Cluster 2 | $ 43.318.829.532 | $ 1.898.101.313 | $ 647.675.126 |
| Cluster 3 | $ 1.456.428.055 | $ 1.035.111.333 | $ 5.790.082.707 |

**Tabla 3.11:** Tabla de resultados por el método CLARA

Los resultados de la segmentación son los siguientes:

- El número de clusters óptimos son 3, según las tablas 3.9 y 3.10, evidenciado también en la figura 3.4
- El tiempo gastado en esta segmentación es de 40 minutos, en un computador de 32 giga-bytes de RAM. Análisis de sensibilidad en el Anexo 1.

Para concluir este Escenario hay que identificar que la entidad no utiliza una metodología con la cual avale sus resultados desde el campo estadístico, la Superintendencia Financiera de Colombia lo aprueba por el momento, pero dicho ente sigue investigando y mejorando sus análisis por lo que en un futuro descartara dicha metodología. Por otro lado, se presento dos escenarios, el primero utilizando K-Means que dada la cantidad de datos es muy demorada la segmentación, por lo que no es una opción viable desde un campo empresarial en este momento, además de que no utiliza todas las variables que contiene la base de datos; para la segunda metodología propuesta en este documento en la sección 2, se plantea el algoritmo CLARA pero con la gran oportunidad de utilizar las variables categóricas de la Entidad, por lo que no estamos despreciando información que puede ser aprovechable dentro de la segmentación.

## 3.2. Escenario 2

### 3.2.1. Resultados de la Entidad 2

Para este escenario revisaremos los resultados obtenidos por la Entidad Financiera 2 (EF2), para el factor de riesgo "Clientes", el cual contiene 27´916.452 registros(filas) por 7 variables(columnas); estas 7 variables son las siguientes:

- Frecuencia de transacciones
- Ciudad de origen
- Transacciones:
  - Efectivo (Suma Ingresos, Egresos y patrimonio).
  - Cheque (Suma Ingresos, Egresos y patrimonio).
  - Transferencia (Suma Ingresos, Egresos y patrimonio).
- Sector o persona natural
- Años de antigüedad

La EF2 utiliza el algoritmo de K-Means para segmentar y el método Elbow para la selección del numero de clusters, el cual calcula la varianza total intra-cluster con respecto al número de clusters, al final se selecciona el numero de clusters en el que el valor apenas consigue mejorar. Y en este caso la entidad la Entidad decide que el numero de clusters es 4, con la siguiente gráfica:

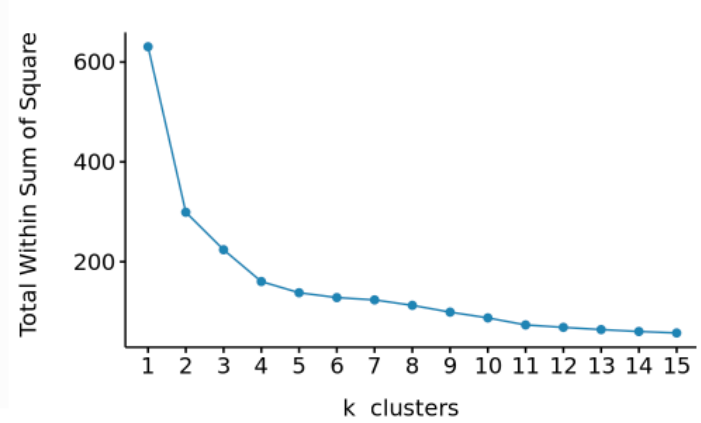



**Figura 3.8:** Gráfica construida con información de la Entidad 2

| Cluster | Efectivo | Cheque | Transferencia |
|---|---|---|---|
| Cluster 1 | $ 1.198.732.056.047 | $ 41.510.637.866 | $ 10.067.274.156 |
| Cluster 2 | $ 1.060.446.326.061 | $ 39.424.321.940 | $ 9.955.188.657 |
| Cluster 3 | $ 406.039.399.479 | $ 9.127.779.547 | $ 3.193.219.489 |
| Cluster 4 | $ 58.545.811.558 | $ 823.620.371 | $ 512.374.944 |

**Tabla 3.12:** Tabla de resultados de la Entidad 2

El algoritmo de K-Means es muy utilizado en entidades financieras, pero no utilizan variables categóricas y solo utilizan variables cuantitativas, por lo que en este escenario no utilizan ni el sector ni ciudad de origen; por lo que el algoritmo de K-Means solo va a utilizar las variable Efectivo, Cheque y Transferencia. Lo anterior refleja que la entidad realiza un trabajo de segmentación óptimo, dado a que esta garantizando las operaciones en cada segmento o cluster, sean lo mas similares en el cluster y lo mas diferentes entre los clusters. A un que se resalta de que no utilizaron sus variables categóricas, pero al fin y al cabo, se da cumplimiento de la norma que a su vez esta en sincronía con los resultados obtenidos.

Resultados dentro de la revisión de los resultados de la entidad, se puede extraer lo siguiente:

- La EF2 no utilizó ningún los índices de verificación del número de clusters, ya que la selección del número de clusters esta ligada a un selección de experto, dado a que esta basada en una gráfica generada por el método Elbow, y no se evalúa un estudio de sensibilidad haber que sucede con 3 o con 5 clusters

- La EF2 no realizó un análisis descriptivo para por entender un poco más los datos y poder generar un primer acercamiento a la información.

- La EF2 notifica que utilizando el algoritmo de K-Means, la entidad se demora 83.18 horas, en un computador de 64 giga-bytes de RAM.

### 3.2.2. Resultados desde la metodología propuesta

Utilizando datos obtenido por la EF2, para el factor de riesgo "Clientes", se realiza la metodología descrita en el Capitulo 2, iniciando con estadísticas descriptivas:

| Variable | Media | Mediana | SD |
|---|---|---|---|
| Efectivo | $ 927.787.019.944 | $ 961.881.796.572 | 72.376.359.314 |
| Cheque | $ 33.786.387.099 | $ 35.243.179.862 | 6.088.635.972 |
| Transferencia | $ 7.401.961.668 | $ 8.864.028.623 | 872.805.724 |
| Años de antigüedad | 11.2 | 23.7 | 5.1 |
| Frecuencia de Transacciones | 8381 | 12362 | 91 |

**Tabla 3.13:** Tabla de descriptiva para las variables cuantitativas la Entidad 2

| Variable | Numero de categorías | Mas frecuente | Frecuencia |
|---|---|---|---|
| Ciudad | 15 | BOGOTA.D.C | 1.829.652.091 |
| Sector | 5 | PERSONA-NATURAL | 983.602.782 |

**Tabla 3.14:** Tabla de descriptiva para las variables cualitativas la Entidad 2

Con respecto a las variables "Ciudad" y "Sector" que en este caso serán tomadas como variables categóricas con la naturaleza de ser factores; utilizando la función **homals** ya descrita su funcionalidad en la sección 2.2, la cual permite realizar la transformación de la misma por medio de las diferentes variables numéricas mencionadas en la tabla 3.13.

Una vez realizada la transformación de la variable "Ciudad" y de "Sector", según MacQueen, 1967 mejora la segmentación si se escala las variables como se indica en en la sección 2.3.

Después de realizar todo lo anterior, sigue el paso de calcular el numero de clusters óptimos como se es presentado en la sección 2.4, por medio de la función **NbClust** y los resultados son los siguientes:

| Índice | Clusters | Value |
|---|---|---|
| Hubert | 0 | 0 |
| Dindex | 0 | 0 |
| KL | 3 | 81.554 |
| CH | 3 | 9.963.509 |
| CCC | 3 | -109.941 |
| Silhouette | 3 | 0.3813 |
| Duda | 3 | 0.9997 |
| PseudoT2 | 3 | 0.196 |
| Beale | 3 | 0.0008 |
| Ratkowsky | 3 | 0.433 |
| PtBiserial | 3 | 0.7304 |
| McClain | 3 | 0.5108 |
| Dunn | 3 | 0.1419 |

**Tabla 3.15:** Tabla de índices parte 1

| Índice | Clusters | Value |
|---|---|---|
| Scott | 4 | 174.451 |
| TrCovW | 4 | 0.0005 |
| Friedman | 4 | 0.035 |
| Ball | 4 | 0.1796 |
| SDindex | 5 | 380.168 |
| Marriot | 6 | 0.0171 |
| TraceW | 6 | 0.0006 |
| Rubin | 6 | -0.0006 |
| DB | 6 | 0.6993 |
| Hartigan | 7 | 49.996 |
| Cindex | 7 | 0.0483 |
| SDbw | 7 | 0.1281 |
| Frey | - | - |

**Tabla 3.16:** Tabla de índices parte 2

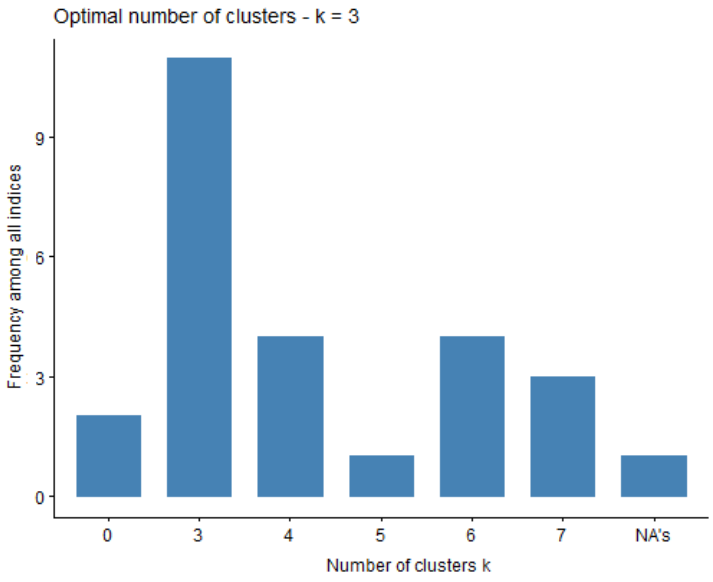



**Figura 3.9:** Histograma del numero de clusters óptimos

De los índices anteriores al igual que la gráfica obtenida, podemos decir que el número de clusters óptimos es 3, dado a que el más frecuente como se presenta en la Figura 3.8.

Con el número de clusters óptimos, se utiliza el algoritmo CLARA citado en sección 2.5 y explicado a detalle en la sección 1.4.3, de este algoritmo resulta una representación gráfica en dos dimensiones que es la siguiente:

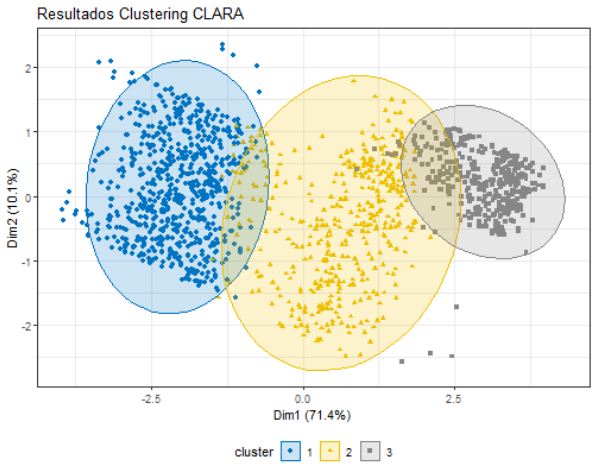



**Figura 3.10:** Escenario 2 - Gráfica de CLARA

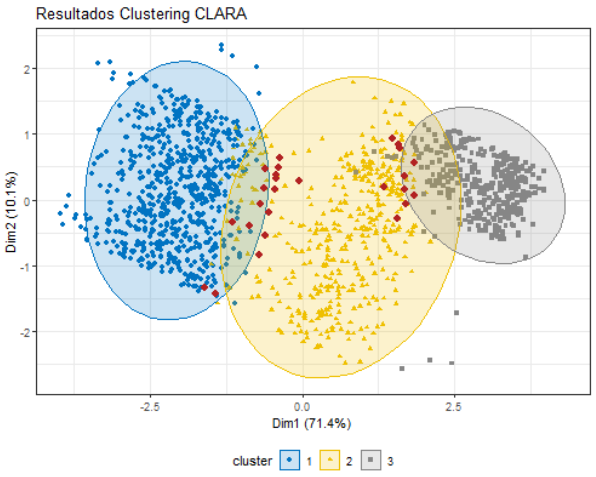



**Figura 3.11:** Escenario 2 - CLARA con atípicos

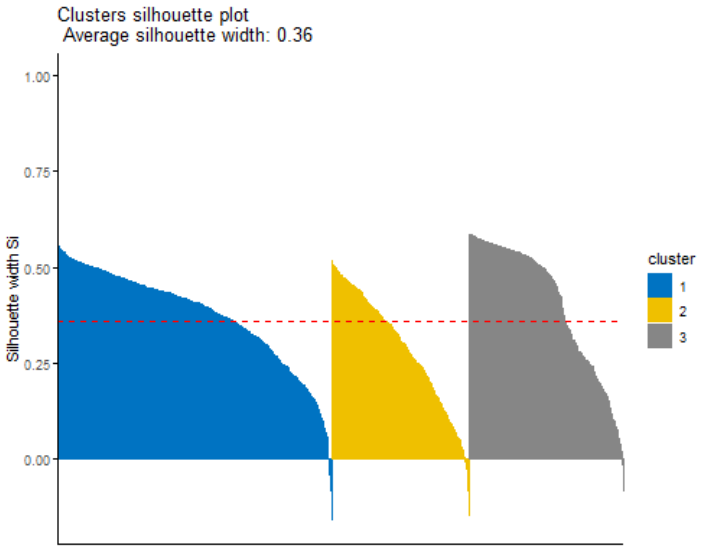



**Figura 3.12:** Escenario 2 - Atípicos Silhouette

Por medio de el índice de Silhouette es posible identificar datos atípicos una vez

realizada la segmentación; esto es un procedimiento propio de la Entidad Financiera y por cuestiones de confidencialidad, no es posible mostrar las situaciones atípicas identificadas en este documento. Por otro lado, los valores cluster obtenidos son los siguientes:

| Cluster | Efectivo | Cheque | Transferencia |
|---|---|---|---|
| Cluster 1 | $ 1.337.784.974.548 | $ 48.567.446.303 | $ 11.778.710.762 |
| Cluster 2 | $ 791.177.974.237 | $ 30.902.473.240 | $ 7.302.926.701 |
| Cluster 3 | $ 70.781.886.173 | $ 963.635.834 | $ 782.933.664 |

**Tabla 3.17:** Tabla de resultados por el método CLARA

Los resultados de la segmentación son los siguientes:

- El número de clusters óptimos son 3, según las tablas 3.15 y 3.16, evidenciado también en la figura 3.8
- El tiempo gastado en esta segmentación es de 1.2 horas, en un computador de 32 giga-bytes de RAM. Como dato adicional se realizó un estudio de sensibilidad del tiempo gastado en la segmentación en el Anexo 1.

Para concluir el Escenario 2 los resultados obtenidos desde el campo estadístico, son afines con la Superintendencia Financiera de Colombia, al igual que los realizados bajo la metodología de este trabajo de grado, descrita en la sección 2. Se concede a la metodología de este documento ser mas robusta a la presentada por la Entidad Financiera 2, dado a la gran cantidad de índices y pruebas realizadas para su selección. Si planteamos el escenario tanto de revisoría fiscal como podría ser KPMG o desde el campo de la Entidad Financiera la comparación en los tiempos de corrida y verificación de los datos es significativo.

# Conclusiones

La metodología implementada en las secciones 3.1.3 y 3.2.2, en dos diferentes escenarios, muestra una adecuada segmentación, en donde se evidencian, múltiples índices que permiten tener una certeza mayor del comportamiento de los diferentes factores de riesgo, como en este caso se evidencia en el factor clientes, dado a que es unos de los principales y es de los que mas información posee. Con lo anterior, se indica que se daría cumplimiento al a parte I del título IV del capítulo IV de la Circular Básica Jurídica de la Superintendencia, y a su vez una revisión estadística más completa y más robusta. Por otro lado, con respecto a la eficacia del algoritmo CLARA dentro de la metodología planteada, se realiza un estudio de comuto donde se prueba el tiempo de cálculo obtenido en diferentes oportunidades dado a que inicia con una semilla aleatoria como se observa a continuación:

| Método | Computador (RAM) | Tiempo Promedio Gastado | Intentos |
|---|---|---|---|
| K-Means | 32 giga-bytes | 1656 minutos | 3 |
| | 8 giga-bytes | 8492 minutos | 2 |
| K-Medoids | 32 giga-bytes | 1947 minutos | 2 |
| | 8 giga-bytes | 10529 minutos | 1 |
| CLARA | 32 giga-bytes | 40 minutos | 10 |
| | 8 giga-bytes | 288 minutos | 4 |

**Tabla 3.18:** Estudio de simulación Escenario 1

En este primer estudio para el Escenario 1, la cantidad de datos es muy grande, pero se realizando múltiples corridas con diferentes algoritmos para poder tener una idea de cuanto se podría demorar el algoritmo en procesar, y en en la Tabla 3.18 se observa que el mejor de todos en velocidad es CLARA.

| Método | Computador (RAM) | Tiempo Promedio Gastado | Intentos |
|---|---|---|---|
| K-Means | 32 giga-bytes | 4869 minutos | 1 |
| | 8 giga-bytes | 28614 minutos | 1 |
| K-Medoids | 32 giga-bytes | 4957 minutos | 1 |
| | 8 giga-bytes | 25348 minutos | 1 |
| CLARA | 32 giga-bytes | 72 minutos | 9 |
| | 8 giga-bytes | 362 minutos | 4 |

**Tabla 3.19:** Estudio de simulación Escenario 2

Para el Escenario 2, la cantidad de datos supera a la del Escenario 1, realizando múltiples corridas que para este caso en particular el algoritmo se demora días en su ejecución. En este caso el ganador también es CLARA como se evidencia en la Tabla 3.19.

Con este trabajo de grado y su metodología planteada en el capitulo 2, se puede desarrollar de manera óptima y efectiva la segmentación necesaria dentro del SARLAFT, al igual que abre la opción de mejorar las metodologías presentes en las diferentes Entidades Financieras, con el fin de mejorar cada vez mas el Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación al Terrorismo (SARLAFT) y la efectividad del mismo desde el tema que se abordo de segmentación.

# Referencias

Babichev, S., Lytvynenko, V., y Osypenko, V. (2017). Implementation of the objective clustering inductive technology based on dbscan clustering algorithm. En *2017 12th international scientific and technical conference on computer sciences and information technologies (csit)* (Vol. 1, pp. 479–484).

Baker, F. B., y Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, *70*(349), 31–38.

Ball, G. H., y Hall, D. J. (1965). *Isodata, a novel method of data analysis and pattern classification* (Inf. Téc.). Stanford research inst Menlo Park CA.

Beale, E. (1969). *Cluster analysis.* Scientific Control Systems, London.

Caliński, T., y Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Correa Chaparro, D. (2015). Optimización del proceso de monitoreo de transacciones (sarlaft).

Davies, D. L., y Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224–227.

Duda, Hart, P. E., y cols. (1973). *Pattern classification and scene analysis* (Vol. 3). Wiley New York.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, *4*(1), 95–104.

FHC, M. (1971). Practical problems in a method of cluster analysis. *Biometrics*, 501–514.

Frey, T., y Van Groenewoud, H. (1972). A cluster analysis of the d2 matrix of white spruce stands in saskatchewan based on the maximum-minimum principle. *The Journal of Ecology*, 873–886.

Friedman, H. P., y Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, *62*(320), 1159–1178.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Gordon, A. (1999). *Classification* (2nd ed.). Chapman Hall/CRC, London.

Halkidi, M., y Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a data set. En *Proceedings 2001 ieee international conference on data mining* (pp. 187–194).

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hill, R. S. (1980). A stopping rule for partitioning dendrograms. *Botanical Gazette*, *141* (3), 321–324.

Hubert, L., y Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2* (1), 193–218.

Hubert, L. J., y Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, *83* (6), 1072.

Jain, A. K., Murty, M. N., y Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31* (3), 264–323.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. doi: 10.1007/978-1-4614-7138-7

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., y Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24* (7), 881–892.

Kraemer, H. (1982). Biserial Correlation. *John Wiley & Sons.*. Descargado de `http://support.sas.com/kb/24/991.html`

Krzanowski, W. J., y Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 23–34.

Lebart, L., Morineau, M., y Piron, M. (2000). Estadística exploratoria multidimensional. *París, Francia: Dunod*.

Lloyd, S. (1982, marzo). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28* (2), 129–137. doi: 10.1109/TIT.1982.1056489

MacQueen, J. (1967). Some methods for classification and analysis of multivariate

observations.

Marriott, F. (1971). Practical problems in a method of cluster analysis. *Biometrics*, 501–514.

McClain, J. O., y Rao, V. R. (1975). Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, 456–460.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *psychometrika*, *45*(3), 325–342.

Milligan, G. W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, *46*(2), 187–199.

Milligan, G. W., y Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.

Ratkowsky, D., y Lance, G. (1978). Criterion for determining the number of groups in a classification.

Rohlf, F. J. (1974). Methods of comparing classifications. *Annual Review of Ecology and Systematics*, *5*(1), 101–113.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Sarle, W. (1983). Sas technical report a-108,cubic clustering criterion.". *SAS Institute Inc.Cary, NC.*.

Scott, A. J., y Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 387–397.

Takane, Y. (2005). Optimal scaling. *Wiley StatsRef: Statistics Reference Online*.

Tibshirani, R., Walther, G., y Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.