Modelo de categorización usando técnicas de clustering nítido y fuzzy para las agencias de una cooperativa de ahorro y crédito

Trabajo de grado para optar al título de

Especialista en Estadística

Claudia Cristina Aparicio Rueda

Directora:

Tulia Esther Rivera Flórez

Magister en Estadística

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

2020

AGRADECIMIENTOS

- En primera instancia agradecer a Dios por guiar mis pasos y permitirme culminar este proyecto.

- A mi familia, quienes siempre serán el pilar fundamental en mi vida depositando toda su confianza en mí y brindándome su apoyo en todo momento.

- Al profesor Henry Lamos Diaz, quien confió en mí y se tomó el arduo trabajo de transmitirme su amplio conocimiento, que junto con su paciencia y apoyo logró encaminarme por el camino correcto para lograr esta meta.

- A Financiera Comultrasan por darme la oportunidad de crecer personal y profesionalmente.

- A todas aquellas personas que fueron parte de nuestra formación académica, muchas gracias.

Claudia Cristina Aparicio Rueda

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Resumen

Título: Modelo de categorización usando técnicas de clustering nítido y fuzzy para las agencias de una cooperativa de ahorro y crédito*

Autor: Claudia Cristina Aparicio Rueda**

Palabras Clave: clúster fuzzy, análisis de conglomerados, cooperativa de ahorro y crédito

Descripción: La investigación se fundamenta en aplicar una clasificación de agencias de una Cooperativa de Ahorro y Crédito usando un algoritmo no supervisado nítido y uno difuso con el objetivo de verificar que la categorización actual corresponde a la misma al aplicar un modelo estadístico usando estas técnicas y que permita realizar una clasificación e identificación de aquellas agencias que puedan pertenecer a una o varias categorías, además de realizar una comparación con los resultados de categorización que hoy en día se maneja en la cooperativa de ahorro y crédito.

En la actualidad la cooperativa utiliza la clasificación en grupos para las agencias, porque otorga información general de cómo operan en cada una de las ciudades en las que está presente la cooperativa y cuáles son sus comportamientos frente al mercado financiero, con la aplicación de los métodos nítido y difuso se busca tener una alternativa de clasificación para las agencias, diferente a la que se realiza actualmente en la cooperativa de ahorro y crédito, que permita ubicar las agencias en clústeres o grupos homogéneos, esto sin llegar a cambiar la metodología existente pero que permita a la cooperativa tener una visión diferente y en un tiempo prudencial se utilice esta propuesta de clasificación.

---

* Trabajo de Grado

** Facultad de Ciencias. Escuela de Matemáticas. Director: Tulia Esther Rivera Florez. Magister en Estadística. Codirector: Henry Sebastián Rangel Quiñonez. Master de Ciencias de la Estadística

Abstract

Title: Categorization model using clustering fuzzy techniques for a savings and credit cooperative agencies*

Author: Claudia Cristina Aparicio Rueda**

Keywords: fuzzy cluster, conglomerate analysis, savings and credit cooperative

Description: The research is based on applying a classification of agencies of a Savings and Credit Cooperative using a clear and diffuse unsupervised algorithm with the aim of verifying that the current categorization corresponds to it when applying a statistical model using these techniques and that allows to perform a classification and identification of those agencies that may belong to one or more categories , in addition to making a comparison with the categorization results that are managed today in the savings and credit cooperative.

Currently the cooperative uses group classification for agencies, because it provides an overview of how they operate in each of the cities in which the cooperative is present and what their behaviors are against the financial market, with the application of the clear and diffuse methods it seeks to have a classification alternative for the agencies , different from that currently carried out in the savings and credit cooperative, which allows agencies to be placed in clusters or homogeneous groups, this without changing the existing methodology but that allows the cooperative to have a different vision and in a reasonable time this classification proposal is used.

---

* Degree work

** Faculty of Science. School of Mathematics. Director: Tulia Esther Rivera Florez. Master in Statistics. Co-director: Henry Sebastián Rangel Quiñonez. Master of Statistics Sciences

Introducción

La clasificación es el proceso que tiene como objetivo dividir un conjunto de objetos o sujetos (individuos) en clases con etiquetas predefinidas de antemano, o en clases aun sin etiquetar usando para esto un conjunto de características o atributos de los individuos. Existen numerosos algoritmos computacionales y estadísticos en la literatura que se usan dependiendo del propósito de la clasificación; los algoritmos se clasifican en algoritmos supervisados (se conocen las clases a la cuales pertenecen los individuos) o no supervisados (no se conocen las clases).

El objetivo de este trabajo es explorar el uso de métodos estadísticos para proponer una clasificación de agencias de una Cooperativa de Ahorro y Crédito usando un algoritmo no supervisados nítido y difuso. Actualmente la cooperativa de ahorro y crédito tiene una metodología para la categorización de las agencias con el fin de analizar periódicamente su comportamiento respecto al volumen del negocio; una vez asignada la categoría a la que pertenece la agencia se asignan recursos para su funcionamiento, remuneración del gerente y asistentes, entre otros.

El análisis actual se realiza a partir de una serie de variables de rendimientos financieros de los productos que se manejan como son: el saldo de crédito, el saldo de ahorro, el saldo de Cdat, el saldo de plan de ahorro programado (PAP) y saldo de aportes que reflejan el Total negocio; por otro lado, las variables como ingresos de cartera, gastos de personal, gastos generales, gastos administrativos, costos depósitos y provisiones reflejan lo que se denomina Excedentes.

De acuerdo con las variables anteriores la metodología tradicional ha clasificado a las agencias en 8 categorías en función de los rendimientos obtenidos. No obstante, considerando que existe cierta subjetividad en la clasificación, se ha querido por medio del presente trabajo construir una clasificación usando algoritmos de clasificación no supervisada, como el algoritmo de *k-means* para el caso nítido y el *c-means* para el caso difuso, con el fin de contrastar las dos clasificaciones producidas.

Desde un punto de vista práctico, se considera que una nueva forma de clasificar las agencias podría apoyar a las Directivas de la Cooperativa en lo que tiene que ver con el diseño de estrategias tanto para la asignación de recursos como para la movilidad de una agencia, entendida esta como el paso de una categoría a otra superior.

1. Antecedentes

La lógica difusa o borrosa (fuzzy) trata de imitar la manera en la que las personas toman decisiones y permite en cierta manera que las decisiones sean muy precisas. Se define como una metodología multivaluada que al contrario a la lógica nítida solo permite dos opciones: pertenencia o no pertenencia de un elemento a un conjunto, la difusa permite valores intermedios para poder definir evaluaciones entre las opciones principales.

En el trabajo de Azar, El-Said, & Hassanien (2013) utilizan técnicas de agrupación para un conjunto de datos de enfermedades de la tiroides, los autores se proponen encontrar una cantidad apropiada de grupos utilizando los algoritmos nítido y difuso, realizando a su vez una comparación de los resultados de estos con el fin de demostrar los resultados de cada algoritmo y los valores de cada característica que logran identificar para la enfermedad de la tiroides que son utilizadas como entrada para el sistema. Dentro del proceso señalan un número diferente de grupos para cada ejecución con el fin de obtener un número óptimo de grupos, para ello aplican el criterio de codo revelando que para el ejercicio la cantidad de agrupamientos es 3.

De otro lado, el documento de Campello & Hruschka (2006) es una generalidad del caso difuso del criterio de ancho de silueta promedio para ilustrar diferentes escenarios y permite utilizarlas juntas en el análisis de conglomerados difusos.

Por otra parte, Artola, Morettini y Blanco (2014) proponen la aplicación de técnicas de clustering para clasificar las universidades públicas de Argentina de acuerdo con su tamaño y a la participación que tengan con relación a los recursos presupuestarios nacionales. Dentro de su trabajo utilizan el método K-medias para compararlos con los del método borroso *Fuzzy 'C-means* el cual permite que cada universidad pueda pertenecer a más de un grupo y no dejar sólo la lógica nítida para que decida si pertenece o no pertenece.

Las aplicaciones de la teoría de conjuntos difusos son muy amplias, hoy en día son indispensables en la descripción de las cosas no nítidas. La teoría borrosa (difusa, *fuzzy*) es capaz de resolver problemas relacionados con la incertidumbre de la información o del conocimiento, proporciona un método formal para la expresión del conocimiento de una forma entendible y comprensible.

2. Justificación

Actualmente, la Cooperativa de Ahorro y Crédito que sirve de base, realiza una categorización de agencias con base en las variables Saldo de Crédito, Saldo de ahorros, Saldo de Cdat, Saldo de plan de ahorro programado (PAP) y Saldo de aportes y excedentes en los últimos 12 meses dando una ponderación del 70% para los saldos del mes y 30% para el peso de los excedentes. Cada una de estas variables tienen unos rangos para cada una de las categorías y estos se aumentan anualmente con base en un indicador de referencia que puede ser el IPC o el PIB, tomando siempre el porcentaje más alto, para ajustar los rangos de cada categoría.

Las agencias actualmente están categorizadas en 8 grupos; la categoría 8 corresponde a aquellas agencias que presentan un mayor rendimiento en las dos variables, mientras que las agencias que se encuentran ubicadas en la categoría 1, corresponde a las agencias que presentan un rendimiento menor. Un ejemplo de la información que se maneja actualmente se da en la siguiente tabla, allí se muestran los rangos para total negocio y excedentes que tendría cada una de las categorías.

Tabla 1.
Rangos de las variables Total negocio y Excedentes para las categorías (miles de pesos).

Total Negocio				Excedentes			
Categorías	**Intervalo ($ millones de pesos)**			**Categorías**	**Intervalo ($ millones de pesos)**		
8	<=	407.749 >	291.249	**8**	<=	8.388 >	6.640

7	<=	291.249	>	198.050
6	<=	198.050	>	128.150
5	<=	128.150	>	81.550
4	<=	81.550	>	48.930
3	<=	48.930	>	37.280
2	<=	37.280	>	23.300
1	<=	23.300	>	0

7	<=	6.640	>	5.126
6	<=	5.126	>	3.728
5	<=	3.728	>	2.563
4	<=	2.563	>	1.538
3	<=	1.538	>	1.048
2	<=	1.048	>	466
1	<=	466	>	0
0	<=	0		

Por consiguiente, una agencia que por sus resultados en la variable total negocio se ubica en la categoría 4 y en excedentes se ubica en la categoría 3, produciría un promedio ponderado de 3.7 que se aproxima a 4, de acuerdo con el cálculo:

Total negocio: 4 * 70% = 2.8

Excedentes: 3 * 30% = 0.9

3,7 → se aproxima a 4

Así, la agencia del ejemplo será ubicada en la categoría 4, resultado muy coherente pues hay casi un acuerdo entre las dos categorías, por lo cual se considera que para la cooperativa sería de gran ayuda contar con otra herramienta para la construcción de una taxonomía como es la basada en métodos de clustering nítido y *fuzzy* en función.

Con la aplicación de los métodos nítido y *fuzzy* se busca tener una alternativa de clasificación para las agencias diferente a la que se realiza actualmente en la cooperativa de ahorro y crédito para así ubicar las agencias en clústeres o grupos homogéneos que le permitan a la cooperativa tener una

visión diferente y en un tiempo prudencial se pueda oficializar esta metodología para la clasificación periódica que hace la empresa.

3. Objetivos

3.1 Objetivo general

Proponer una clasificación para las agencias de una cooperativa de ahorro y crédito usando técnicas de clustering nítido y difuso.

3.2 Objetivos específicos

- Comparar los resultados obtenidos por medio de las técnicas clustering usuales con los resultados de las soluciones del análisis *fuzzy*.
- Comparar métodos de elección de número de grupos para determinar la mejor solución clustering.
- Validar la solución obtenida por medio de la comparación del modelo de categorización propuesto con la categorización que maneja la cooperativa actualmente.

4. Metodología

Para la realización del presente trabajo se plantea una serie de etapas para cumplir con los objetivos propuesto.

4.1 Etapa 1. Integración y recopilación de la información

Esta etapa está conformada por la selección e identificación de las variables que se usan para el estudio mediante la gestión de bases de datos en las respectivas agencias. Actualmente se cuenta con 60 agencias de la cooperativa de ahorro y crédito, algunas de las variables que se usan en la construcción del modelo se describen a continuación:

Tabla 2
Variables a tener en cuenta para cada agencia.

Variables	Descripción3
Saldo de Crédito	Corresponde al saldo capital de los créditos vigentes por agencia
Saldo de Ahorros	Corresponde al saldo de las cuentas vigentes de ahorro por agencia
Saldo de Cdat	Corresponde al saldo de los Cdat vigentes por agencia
Saldo de Plan de Ahorro programado	Corresponde al saldo de los planes de ahorro programado vigentes por agencia
Saldo de Aportes	Corresponde al saldo de los aportes de los asociados vigentes por

	agencia
Excedentes	Corresponde a los ingresos, costos y gastos que genera cada agencia mensualmente

4.2 Etapa 2. Análisis y construcción del modelo de clasificación

En esta etapa mediante los softwares estadísticos R y SPSS se procesarán los datos para los diferentes análisis estadísticos implementados. A continuación, se describen las actividades que se desarrollaron para la construcción del modelo no supervisado:

- Aplicar y comparar varios algoritmos jerárquicos y no jerárquicos, con el fin de evaluar cuál sería el preferible para el proceso de analítica a desarrollar.
- Aplicar criterios para medir el rendimiento de los algoritmos.

4.3 Etapa 3. Comparación de Índices de validación del agrupamiento

En esta etapa se comparan diversos índices de agrupamiento como coeficiente de partición, índice silueta en agrupamiento nítido, silueta difusa, coeficiente de partición, Entropia de clasificación, índice de partición, índice de separación, para luego seleccionar la mejor partición.

4.4 Etapa 4. Evaluación e interpretación de la información

El objetivo de esta fase es presentar los resultados y descubrir qué puede funcionar en la cooperativa para revisar qué agencias están cerca de una categoría superior diferente a la que se

encuentran actualmente. Se usa el software estadístico R para el análisis de los datos y construcción del modelo.

5. Marco Teórico

En la categoría de reconocimiento de patrones, existe el tipo de problema no supervisado (clasificadores no supervisados) que consiste en descubrir la estructura del conjunto de datos si los hay. Esto generalmente significa que el usuario desea saber si hay grupos en los datos, y qué características hacen que los objetos sean similares dentro de un grupo. La elección de un algoritmo es una cuestión de características del problema. Diferentes algoritmos pueden presentar diferentes estructuras para el mismo conjunto de datos. Una característica de este tipo de aprendizaje es que no hay ninguna verdad fundamental contra la cual comparar los resultados. La única indicación de qué tan bueno es el resultado es la estimación subjetiva del usuario (Kuncheva, 2004).

El proceso de agrupamiento o análisis de Clúster consiste en organizar una colección de elementos en un conjunto de grupos homogéneos. Dichos elementos están representados por un vector de valores de atributos, es decir, son puntos de algún espacio multidimensional. Estos valores también se suelen denominar atributos, componentes o simplemente variables. Intuitivamente, dos elementos pertenecientes a un agrupamiento válido deben ser más parecidos entre sí que aquellos que estén en grupos distintos. Partiendo de esta idea se desarrollan las técnicas de agrupamiento. Estas técnicas dependen de cómo sean los datos de partida, de qué medidas de semejanza se estén utilizando y de qué clase de problemas se estén resolviendo.

Los algoritmos de agrupamiento se clasifican en dos grandes grupos, algoritmo de agrupamiento Jerárquico y el de agrupamiento no Jerárquico o particional. En esta sesión se va a discutir solamente el algoritmo no jerárquico *k-means* y el *c.means*.

5.1 Algoritmos de Agrupamiento Jerárquico.

Los métodos jerárquicos obtienen una clasificación a partir del cálculo de la matriz de distancia entre los objetos (individuos/sujetos). Con esta técnica en cada etapa se forman grupos de manera aglomerativa o mediante un proceso de división. En las técnicas aglomerativas se parte de n grupos, cada uno formado por un solo individuo; las clases o clústeres cercanos se van agrupando por pasos sucesivos hasta formar un solo grupo o conglomerado con todos los individuos. Las técnicas de división, a diferencia de las aglomerativas inician con un solo grupo formado por todos los individuos; este grupo es dividido en dos, tres y más grupos hasta que finalmente cada objeto forma un solo grupo.

Una desventaja de los métodos jerárquicos es que después que un individuo se asigna a un grupo, él siempre va a pertenecer al grupo en cualquier otro paso posterior. La asignación de un individuo a un grupo no se puede modificar a través de los pasos en la construcción de los clústeres.

Las técnicas jerárquicas son utilizadas en una primera etapa y sus resultados de la clasificación pueden ser visualizados en un diagrama de árbol o dendrograma en el cual las hojas terminales representan a cada uno de los individuos y las ramas son la clase o conglomerado

formado por todos los individuos. El número de conglomerados decrece a medida que se aumenta la altura del árbol.

5.1.1 Método de Ward

El método de Ward busca la mínima variabilidad entre los posibles conglomerados que se pueden formar en cada una de las etapas. En cada etapa el número de grupos se reduce en uno, por fusión de dos grupos que tienen el más pequeño incremento en la suma total de la variabilidad entre los grupos. Obsérvese que ahora se parte de los elementos directamente en lugar de la matriz de distancias.

En el trabajo De la Fuente (2011) se indica que el objetivo del método de Ward es la minimización de la variación intra grupal de la estructura formada a través del criterio:

$$SS(W) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ig} - \bar{x}_g)'(x_{ig} - \bar{x}_g)$$

El criterio comienza con cada dato forma en un grupo, g=n y, por tanto, $SS(W)$ es cero. A continuación, se unen los elementos que produzcan el incremento mínimo de $SS(W)$. Obviamente, esto implica tomar los más próximos con la distancia euclídea. En la siguiente etapa, se tiene n-1 grupos, n-2 de un elemento y uno de dos elementos. Se une de nuevo dos grupos para que $SS(W)$ crezca lo menos posible con lo que pasamos a n-2 grupos y así sucesivamente, hasta tener un único grupo.

5.1.2 Algoritmos de agrupamiento no jerárquico o particional.

Dado D, un conjunto de datos de n objetos, y k, el número de clústeres a formar, un algoritmo de partición organiza los objetos en k particiones ($k \leq n$), donde cada partición representa un clúster. Los clústeres se forman para optimizar un criterio de partición objetivo, como una función de disimilitud basada en la distancia, de modo que los objetos dentro de un clúster sean similares, mientras que los objetos de diferentes clústeres sean diferentes en términos de los atributos del conjunto de datos. (Han et al., 2011).

5.1.3 Agrupamientos nítidos o duros.

Cada dato debe asignarse exactamente a un grupo. Estos métodos clásicos producen particiones exhaustivas pero la asignación de datos al grupo puede ser inadecuado en presencia de puntos de datos que son igualmente distantes a dos o más grupos. Una partición dura obliga a la asignación completa de tales puntos de datos a uno de los grupos (Döring et al., 2006).

5.1.4 Método de k-medias

El algoritmo de agrupamiento más frecuentemente utilizado en este tipo de agrupamiento es el K-medias (en inglés, *k-means*), conocido por la sensibilidad ante datos atípicos, pero es eficiente en términos del tiempo computacional empleado para el análisis de la información (Azar, El-Said, & Hassanien, 2013) . Este algoritmo busca una partición para minimizar la suma del error

cuadrado de todos los K grupos, es decir el error entre μ_k μ_k (media del cluster w_k) w_ky los puntos en un cluster w_k; se define por la siguiente ecuación (Jain, 2010):

$$J(W)) \sum_{k=1}^{K} \sum_{x_i \in w_k}^{.} ||x_i - \mu_k||^2$$

Por consiguiente, el método se basa en la minimización de la suma de las distancias de los objetos asignados a un conglomerado al centroide de dicho conglomerado. En forma breve se describe los pasos para el agrupamiento de un conjunto de *n* individuos en *k* grupos:

1. Se escogen los centroides de cada uno de los grupos; $\mu_i(0), i = 1,2,..,K$, donde $\mu_i(l)$ representa el i -ésimo centroide del cluster en la l -ésima iteración. Para la primera iteración, por ejemplo, los valores de los centroides se seleccionan de manera aleatoria.

2. Se calcula la distancia entre los individuos y cada centroide de los clústeres.

3. Se asigna cada individuo al grupo cuyo centroide esté más cercano a él. Así, se asigna el vector muestra x_i al clúster w_j si

$\| \mu_j(k) - x_i \| < \| \mu_i(k) - x_i \|$ para $i = 1,2..K$, $i \neq j$,

donde

$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$ que corresponde a la distancia euclidiana

4. Se calculan los nuevos centroides de los grupos formados en el paso anterior.

$$\mu_j(k+1) = \frac{1}{n} \sum_{x \in w_j} x_i .$$

Donde, n es el número de elementos en el clúster w_j

5. Se verifica el criterio de convergencia. Por ejemplo, una condición para la convergencia es que ningún centroide se cambie durante la evaluación del paso 4. La condición puede ser expresada como $\mu_j(l+1) = \mu_j(l),\ j = 1,2,..K$

Si la condición anterior se satisface entonces el algoritmo converge, en caso contrario repetir los pasos 2, 3 y 4 hasta que la suma de las distancias de los objetos al centroide del grupo al cual pertenecen sea mínima.

5.2 Selección del número óptimo de Clústeres.

Se usan una serie de método que ayudan a la elección del número óptimo de grupos para la solución del problema de agrupamiento. A continuación, se explica dos métodos de uso común.

5.2.1 Elbow method

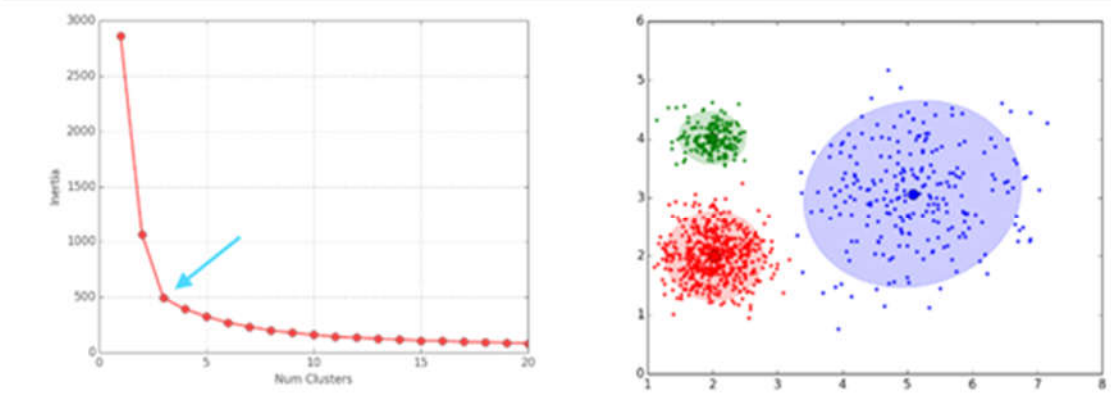
El método utiliza los valores de la función $J(W)$ (llamada inercia) una vez se ha aplicado *K-means* a cada uno de los diferentes clústeres (desde 1 a N clústeres), en la ecuación se presenta la inercia como la suma de las distancias al cuadrado de cada objeto del clúster a su centroide:

$$Inercia = \sum_{i=0}^{N} ||x_i - \mu||^2$$

Luego se construye un diagrama de líneas con los valores de la inercia en relación con el número de clústeres como se muestra en la Figura 1. En el diagrama se debe observar un punto en el cual la variación o donde tiene un cambio brusco en la evolución de la inercia y se asemeja al codo de un brazo doblado, justo ese punto donde se presenta el cambio brusco en la inercia indica el número óptimo de clústeres a seleccionar para la muestra de datos.

En la figura se muestra una modificación del método de Elbow adaptado de Moya (2016).

Figura 1
Elbow Method



Nota: Tomado de: http://noesis.uis.edu.co/bitstream/123456789/11077/1/175093.pdf

5.2.2 Silhouette method

El método de silueta promedio calcula la silueta promedio de las observaciones para diferentes valores de k. El número óptimo de conglomerados k es el que maximiza la silueta de promedios sobre un rango de valores posibles para k (Kaufman y Rousseeuw, 1990).

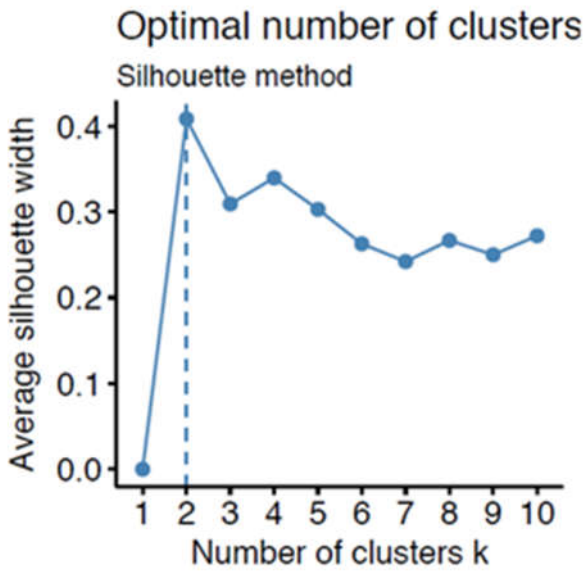
El algoritmo es similar al método del codo y se puede calcular de la siguiente manera:

- Calcular el algoritmo de agrupamiento (por ejemplo, *k-means* clustering) para diferentes valores de k. Por ejemplo, variando k de 1 a 10 clusters.

- Para cada k, calcular la silueta promedio de las observaciones (avg.sil).

- Trazar la curva de avg.sil de acuerdo con la cantidad de conglomerados k.

Figura 2
Silhouette method - Numero optimo de clusteres



Nota: Tomado de: http://noesis.uis.edu.co/bitstream/123456789/11077/1/175093.pdf

5.3 Agrupamientos suaves o difusos.

La matemática difusa (borrosa, *fuzzy*) se encarga de modelar fenómenos que son difíciles de describir de manera precisa; tienen implícitos cierto grado de imprecisión. La imprecisión se podría asociar con la forma, la posición, el momento, el sabor, el olor, e incluso en el lenguaje que

describe lo que son. En muchos casos el mismo concepto puede tener diferentes grados de imprecisión en diferentes contextos o tiempo. Este tipo de imprecisión asociado continuamente a los fenómenos es común en todos los campos de estudio: sociología, marketing, medicina, ingenierías, etc.

El análisis de clúster nítido discutido arriba tiene la particularidad que cada individuo pertenece a un único clúster, sin embargo, hay situaciones donde puede existir cierta ambigüedad sobre algún individuo que podría estar tanto en un grupo como en otro debido a que comparte elementos comunes.

La idea de Zadeh (1965) del conjunto difuso permite ampliar la definición de un conjunto clásico en el marco que la relación de pertenencia que exista de un elemento x en un conjunto A deja de ser una relación de pertenencia o no, dicho de otra manera para un subconjunto B de un conjunto universal U, un elemento X ∈ U y tiene solo dos posibilidades, X ∈ U o X ∉ U. El grado de pertenencia por lo tanto va a depender del problema al cual haga referencia. Por lo tanto, la idea es flexibilizar el grado de pertenencia de los elementos en los conjuntos y permitir que un elemento pueda pertenecer parcialmente a un conjunto dado, es por ello por lo que Zadeh introduce el concepto de conjunto difuso para modelar matemáticamente un conjunto.

Un conjunto difuso subconjunto de A de un conjunto universal $U \neq \emptyset$ se describe mediante una función $\mu: A \rightarrow [0,1]$, o bien, puede identificarse como μ un subconjunto de $A \times [0,1]$, esto es, dela forma $\{\{x, \mu(x)\} | x \in A\}$. La función de pertenencia representa al conjunto y por tanto

contiene todas las características de dicho conjunto y es un elemento de estudio inmediato en la teoría difusa.

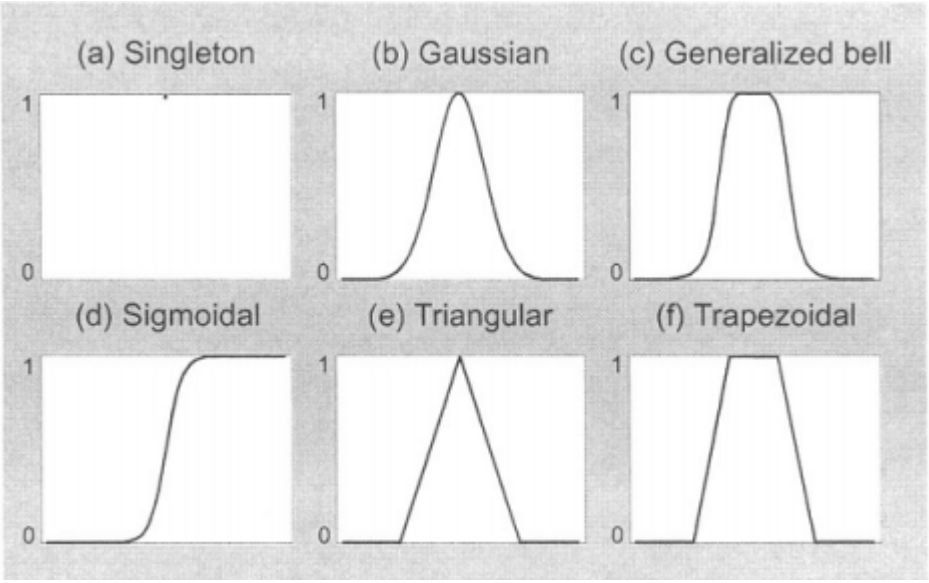
El análisis de conglomerados difuso permite asociaciones graduales de puntos de datos a grupos en el intervalo [0, 1], lo que da la flexibilidad para expresar que los puntos de datos pertenecen a más de un grupo al mismo tiempo (Döring et al., 2006).

Los conjuntos difusos proporcionan una búsqueda a través de los datos orientada y especificada en términos lingüísticos que ayuda a descubrir dependencias entre los datos en un formato cualitativo o semi-cualitativo, además estos conjuntos permiten incluir fácilmente información contextual e incluso orientar el proceso de búsqueda con información lingüística Hernandez, Ramirez & Ramirez (2004).

Se define la familia de conjuntos difusos $\{\tilde{A}_i, i = 1,2, \dots, c\}$ en un universo de puntos X, es posible asignar una pertenencia a los distintos datos en cada conjunto difuso, de tal forma que el punto k tiene la siguiente pertenencia a la clase *i:*

$$u_{ik} = u_{\tilde{A}_i}(x_k) \in [0,1].$$

Si $u_{\tilde{A}_i}(x_k)$ tiende a cero hay bajo grado de pertenencia, y un valor cercano a 1 significa alto grado de pertenencia del dato x_k al clúster $\tilde{A}_i$. Para determinar este valor o nivel de pertenencia existen varias funciones tal como se describe en la imagen a continuación.



Nota: Tomado de https://cse.iitkgp.ac.in/~dsamanta/courses/archive/sca/Archives/Chapter%203%20Fuzzy%20Membership%20Functions.pdf

En el paso siguiente, dado que las particiones a los clústeres son difusas, no hay una simple etiqueta que indique cuáles datos pertenecen a los clústeres, en su lugar, los métodos de clustering difuso asocian un vector etiqueta difuso a cada dato x_k que indica su pertencia a los c clústeres.

Entonces la matriz c x n, $U = (u_{ik}) = (\overrightarrow{u_1}, \ldots, \overrightarrow{u_n})$ es llamada matriz de partición difusa, y el conjunto de matrices de particiones difusas de X es denotado por:

$$M_{fc} = \left\{ U_{cxn} | u_{ik} \in [0,1]; \sum_{i=1}^{c} u_{ik} = 1,; \; 0 < \sum_{k=1}^{n} u_{ik} < n \right\},$$

$$i = 1,2,\dots,c \;,\; k = 1,2,\dots,n.$$

5.3.1 Algoritmos de agrupamiento difuso.

El agrupamiento difuso permite una pertenencia gradual entre 0 y 1 de cada uno de los puntos de datos a cada uno de los grupos, y ofrece más detalles del modelo de datos. Además, este grado de pertenencia también expresa la ambigüedad o definitiva pertenencia de un punto de datos a un grupo, así que el concepto de estos grados de pertenencia está respaldado por la teoría de conjuntos difusos (Azar et al., 2013).

5.3.2 Clúster suaves o difusos

Este algoritmo surge debido a la necesidad de resolver una deficiencia en el agrupamiento exclusivo, donde considera que cada elemento se puede agrupar inequívocamente con los elementos de su clúster y que no se asemeja al resto de los elementos.

Para Bezdek, Ehrlich, & Full, (1984), tras la introducción de la lógica difusa por Zadeh en 1965 surgió una solución para este problema, caracterizando la similitud de cada elemento a cada uno de los grupos. El algoritmo difuso k-medias fue presentado inicialmente por Dunn (1973), y completado por Bezdek (Bezdek et al., 1984), con el fin de minimizar la función objetivo:

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{m} ||x_i - c_i||^2$$

Se puede definir la familia de conjuntos difusos $\{A_i\ i = 1, 2, \ldots, c\}$ en un universo de puntos X, donde es posible que se asigne una pertenencia a los diferentes datos en cada conjunto difuso, de tal forma que el punto k tiene la pertenencia a la clase *i*:

$$U_{ik} = U_{\tilde{A}i}(X_k) \in [0, 1]$$

Entonces, si $U_{\tilde{A}i}(X_k)$ tiende a cero significa que hay un bajo grado de pertenencia y si tiende a 1 significa que existe un alto grado de pertenencia del dato X_k al cluster $\tilde{A}i$. Luego dado que las particiones a los clústeres son difusas, no hay una simple etiqueta que indique cuales datos pertenecen a cuál clúster. Por lo que los métodos de clustering difuso asocian un vector difuso a cada dato X_k que indica su pertenencia a los n clústeres.

Por lo tanto la matriz c x n, $U = (U_{ik}) = (U_1, \ldots, U_n)$ es denominada la matriz de partición difusa y su conjunto de matrices de particiones de X es denotado por:

$$M_{fc} = \left\{ U_{cxn} | u_{ik} \in [0,1]; \sum_{i=1}^{c} u_{ik} = 1,; \ 0 < \sum_{k=1}^{n} u_{ik} < n \right\}$$
$$i = 1,2,\ldots,c \ , \ \ k = 1,2,\ldots,n.$$

De acuerdo con Döring et al., (2006), el análisis difuso permite crear asociaciones graduales de puntos de datos a clústeres en el intervalo [0, 1] que da la flexibilidad para mostrar

que los puntos de datos pueden pertenecer a más de un grupo al mismo tiempo. Es decir, proporcionan una búsqueda a través de los datos que ayudan a descubrir dependencias entre los datos en una forma cualitativa.

Como lo comenta Azar et al., (2013), el agrupamiento difuso permite una pertenencia gradual entre 0 y 1 de cada uno de los puntos de datos a cada uno de los grupos y ofrece más detalles del modelo de datos. Este grado de pertenencia también expresa la ambigüedad o pertenencia definitiva de un punto de datos a un grupo, así que estos grados de pertenencia está respaldado por la teoría de conjuntos difusos.

6. Resultados

6.1 Análisis Estadístico Descriptivo

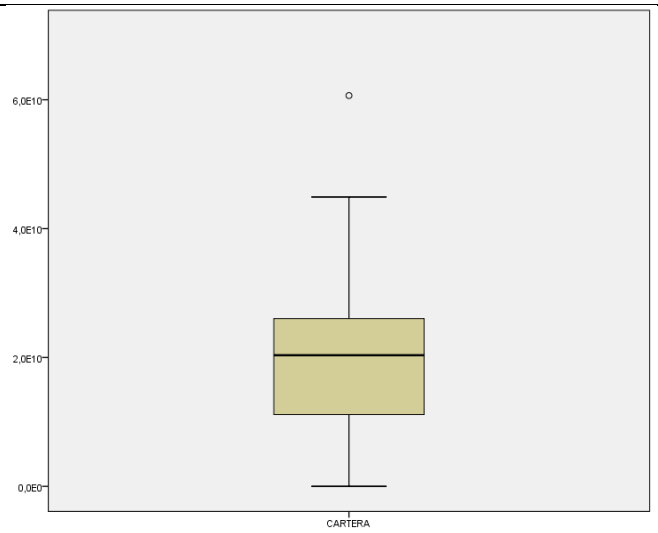
En esta sección se presenta un análisis descriptivo con las variables de estudio de la base de datos, utilizando la herramienta SPSS. En el caso que la variable sea métrica se calcula la media, mediana, desviación estándar, mínimo y máximo con el uso de diagramas de cajas, de dispersión, barras para la visualización de las variables. Para el caso de una variable no métrica se presenta tablas con las respectivas proporciones para cada modalidad.

6.2 Variables Cuantitativas

6.2.1 Saldo de Crédito

Figura 3
Tabla estadísticos descriptivos y diagrama de caja de la variable saldo de crédito

Media		20.534.396.718,86
Mediana		20.353.279.373,88
Desviación estándar		12.672.968.875,41
Mínimo		,00
Máximo		60.654.833.147,81
Percentiles	25	10.978.462.966,06
	50	20.353.279.373,88
	75	26.071.485.470,73



Nota: cooperativa, base de datos con corte a diciembre de 2019.

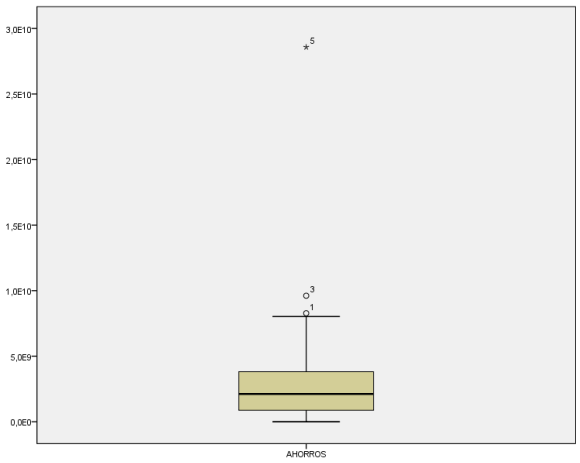
La variable saldo de crédito por agencia, corresponde al saldo total de cada crédito a corte de diciembre de 2019 proporcionado por la cooperativa en estudio. El valor del crédito promedio se ubica en $20.353.279.393 pesos, pero la variable describe una alta variabilidad (rango de variación es cercano a los 60.000 millones de pesos. La Figura 3 evidencia que los datos no parecen seguir una distribución normal, ya que se presenta una concentración de datos por debajo de $26.000 millones, percentil 75, y la mediana. Así mismo se presenta un dato atípico que corresponde a un valor de más de 60.000 millones. De las 60 agencias analizadas, se encuentran 20 agencias que tienen saldo de crédito por encima de $25.000 millones de pesos que corresponden a las agencias que tienen en el mercado un tiempo superior a los 16 años de antigüedad.

6.2.2 Saldo de ahorros

Figura 4

Tabla estadísticos descriptivos y diagrama de caja de la variable saldo de ahorros

Media		2.930.038.071,85
Mediana		2.120.395.378,93
Desviación estándar		4.040.032.130,50
Mínimo		,00
Máximo		28.575.623.823,05
Percentiles	25	855.180.017,77
	50	2.120.395.378,93
	75	3.918.536.367,28



Nota: cooperativa, base de datos con corte a diciembre de 2019.

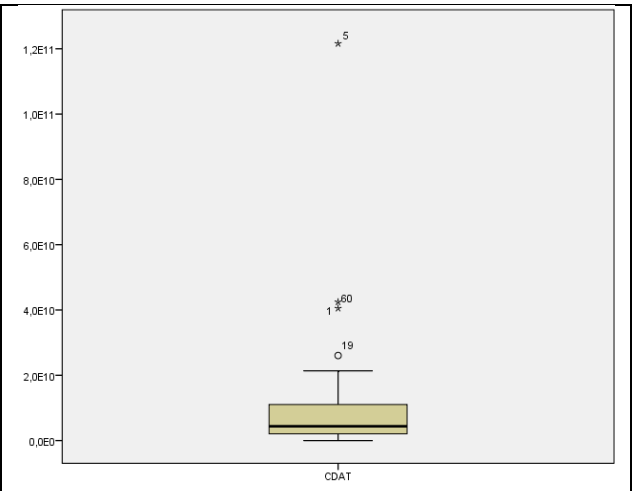
La variable saldo de ahorros por agencia, se refiere a los recursos que reportan las agencias por cuentas de ahorros correspondientes a los ahorradores adscritos a cada agencia. En esta variable se presenta valores atípicos y un saldo máximo de $28.000 millones; de las 60 oficinas, el 75% tienen depósitos por debajo de $4.000 millones de pesos, pero también hay un 25% de las agencias que tienen un saldo bajo comparado con el saldo promedio que se ubica en cerca de los 2.000 millones de pesos.

6.2.3 Saldo de Cdat

Figura 5

Tabla estadísticos descriptivos y diagrama de caja de la variable saldo de Cdat

Media		9.524.227.288,43
Mediana		4.380.870.766,14
Desviación estándar		17.174.825.168,93
Mínimo		,00
Máximo		121.588.653.684,52
Percentiles	25	2.040.813.202,50
	50	4.380.870.766,14
	75	11.096.527.150,75



Nota: cooperativa, base de datos con corte a diciembre de 2019.

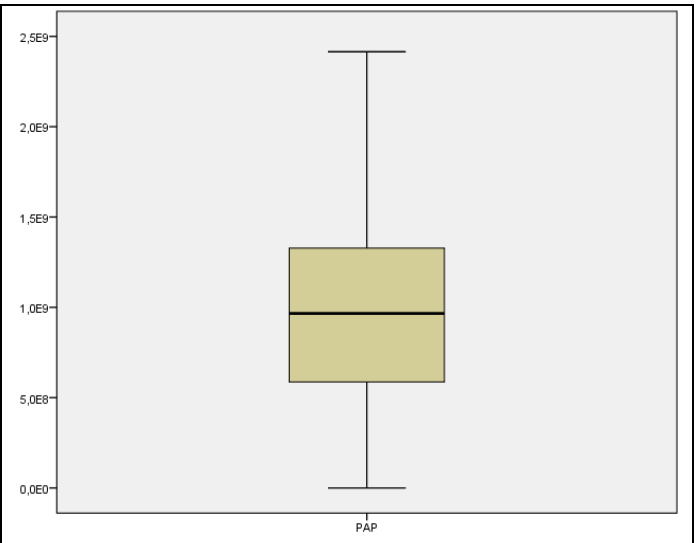
Este producto es característico por sus diferentes tasas de interés, por lo que es el producto más utilizado por los asociados. La distribución es sesgada a la derecha con presencia de datos atípicos en algunas agencias debido a sus montos de inversión. El saldo mediano de la variable Cdat se ubica alrededor de los $ 4.000 millones con un rango intercuartílico cercano a los $9.000 millones. Hay un valor extremo que supera los $120.000 millones que claramente afecta el cálculo del valor promedio de la variable ($9.824 millones) por lo cual éste deja de ser representativo, ya que se tiene que el 50% de las oficinas tienen un valor inferior a los $4.380 millones en CDAT.

6.2.4 Saldo de PAP (Plan de Ahorro Programado)

Figura 6

Tabla estadísticos descriptivos y diagrama de caja de la variable saldo de PAP

Media		943.617.233,55
Mediana		966.419.842,56
Desviación estándar		573.259.661,52
Mínimo		,00
Máximo		2.415.122.020,27
Percentiles	25	582.314.939,28
	50	966.419.842,56
	75	1.328.434.513,41



Nota: cooperativa, base de datos con corte a diciembre de 2019.

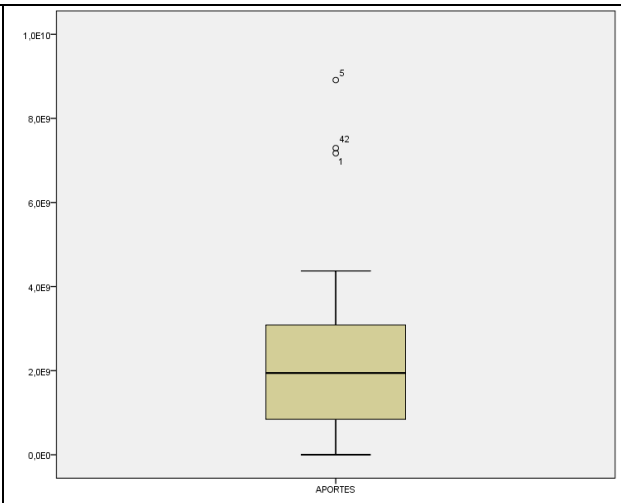
El producto PAP por sus siglas Plan de Ahorro Programado, es un producto donde el asociado pacta un plazo y un monto que se compromete a consignar. Estas cuentas de ahorro programado totalizado por agencia son derivados de los pagos consensuados anticipadamente y que de manera mensual abona a esta cuenta. El saldo máximo de PAP es de más de $2.000 millones, pero el saldo promedio es de $943 millones. Dada la asimetría en la variable conviene observar mejor el comportamiento de los cuartiles para describir la distribución, de esta manera podemos decir que hay un 50% de los valores se ubican entre los $582 millones y los $ 1.328 millones.

6.2.5 Saldo de Aportes

Figura 7

Tabla estadísticos descriptivos y diagrama de caja de la variable saldo de aportes

Media		2.144.276.926,48
Mediana		1.943.375.779,30
Desviación estándar		1.770.095.929,217
Mínimo		,00
Máximo		8.912.591.202,13
Percentiles	25	810.564.809,39
	50	1.943.375.779,30
	75	3.101.666.518,69



Nota: cooperativa, base de datos con corte a diciembre de 2019.

Para el saldo de aportes por agencia, la media es de $ 2.100 millones con una mediana muy cercana de $ 1.900 millones. En cuanto a dispersión, el coeficiente de variación es del 83% lo que indica que la variable saldo de aportes tiene una alta variabilidad pero hay que observar que hay presencia de valores atípicos, así, aunque el saldo máximo es de más de $8.000 millones, el 75% de las agencias tienen un saldo de aportes menor a los $ 3.101 millones.

6.2.6 Excedentes

Figura 8

Tabla estadísticos descriptivos y diagrama de caja de la variable excedentes

Media		518.708.513,13
Mediana		321.855.959,88
Desviación estándar		762.281.558,13
Mínimo		-784.978.202,25
Máximo		3.722.051.489,30
Percentiles	25	-60.435.108,33
	50	321.855.959,88
	75	909.383.940,45

Nota: cooperativa, base de datos con corte a diciembre de 2019.

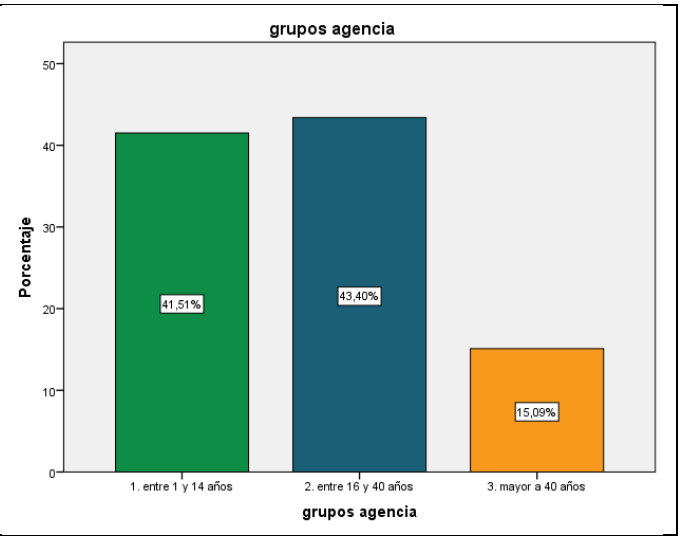
Para la variable Excedentes por agencia, la media es de $518 millones con una mediana de $ 321 millones. El saldo máximo por agencia es de más de $3.700 millones, se observa una alta variabilidad, rango de variación de cerca de 3.000 millones . El 75% de las agencias tienen un saldo de aportes inferior a $909 millones.

6.3 Variables Cualitativas

6.3.1 Antigüedad en años de las agencias

Figura 9
Tabla de frecuencia y diagrama de barras. Variable antigüedad

Frecuencia		Porcentaje	% acum
Entre 1 y 14 años	22	41,5	41,5
Entre 16 y 40 años	23	43,4	84,9
Mayor a 40 años	8	15,1	100,0



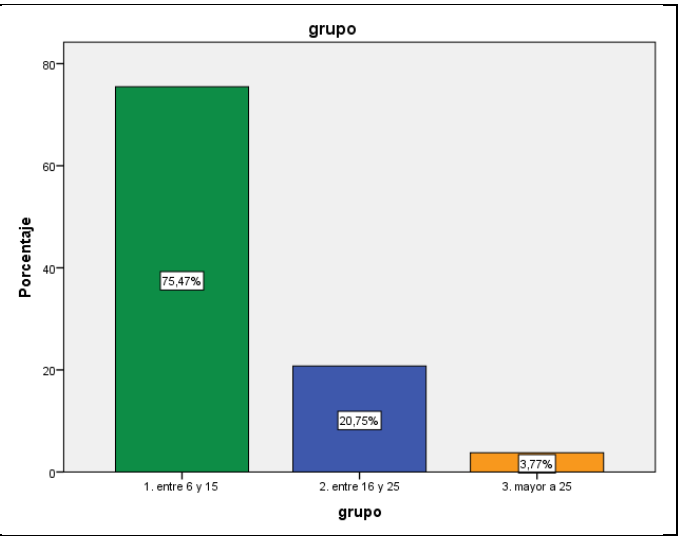
Nota: cooperativa, base de datos con corte a diciembre de 2019.

La variable antigüedad (en años) de las agencias corresponde al número de años en los cuales la agencia ha estado activa, tiene un promedio de 22 años. En la actualidad el 43% de las agencias de la cooperativa tienen una antigüedad considerable, entre 16 y 40 años, hay un porcentaje similar de agencias nuevas o con menor tradición, tan solo un 15% corresponde a agencias antiguas con más de 40 años de antigüedad en el mercado.

6.3.2 Número de empleados por agencia

Figura 10
Tabla de frecuencia y diagrama de barras. Variable cantidad empleados

Frecuencia		Porcentaje	% acum
Entre 6 y 15	40	75,5	75,5
Entre 16 y 25	11	20,8	96,2
Mayor a 25	2	3,8	100,0



Nota: cooperativa, base de datos con corte a diciembre de 2019.

La variable cantidad empleados corresponde al número de empleados que tienen asignados a cada agencia de la cooperativa, el promedio es de 14 empleados, se observa que la mayoría son agencias pequeñas en cuanto nómina, el 75% de las agencias de la cooperativa tiene entre 6 y 15 empleados y solo un 3% de las agencias tienen más de 25 empleados en su planta.

6.4 Análisis de factorial

El análisis factorial es una técnica multivariada que tiene como objetivo el resumen y la reducción de datos (variables), analizando la estructura de correlaciones entre las variables definiendo una serie de dimensiones subyacentes comunes conocidas como factores (Hair, Anderson, Tatham & Black, 1999). El principal objetivo del análisis factorial exploratorio es definir el número de variables latentes o factores que explican la varianza común entre un conjunto de variables observadas. El modelo general del análisis factorial está definido por la siguiente expresión:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{1m}F_m + e_i \qquad (1)$$

Donde:

X_i: Es la i-ésima puntuación para la variable X.

a_{i1}, a_{i2}, …, a_{1m}: Son las cargas o pesos de los factores para la i-ésima variable.

F_1, F_2, F_m: Son los m factores comunes no correlacionados, cada uno con media cero y varianza uno.

ei: es el único factor específico a la i-ésima variable que no está correlacionado con ningún factor común y tiene media cero y varianza uno.

Existen una serie de pruebas estadísticas que apoyan la decisión de realizar un análisis de factores. Entre los procedimientos de uso común está el cálculo de la matriz de correlación de los datos que permiten contrastar la hipótesis de que existe un número adecuado de correlaciones significativas entre las variables. Si se presenta una correlación baja entre las variables lo más indicado es no realizar un análisis factorial. Otros supuestos que deben abordarse se hacen a través de la prueba de esfericidad de Bartlett y el Índice KMO (Kaiser-Meyer-Olkin) de adecuación de la muestra que se describen a continuación.

La prueba de Bartlett (1950) permite verificar la hipótesis acerca de que si la matriz de correlación es equivalente a una matriz de identidad (incorrelación entre variables). El estadístico de Bartlett para el contraste de esfericidad es

$$\chi^2_{0,5(p^2-p)} = -\left[n - 1 - \frac{1}{6}(2p+5)\right] \ln|\hat{\rho}|,$$

donde p es el número de variables,

la matriz de correlación muestral (Uriel y Aldas, 2015).

La medida de adecuación muestral de Kaiser-Meyer-Olkin se puede considerar como un índice de dependencia, otros la describen como un recurso para evaluar el grado en que cada una de las variables es predecible a partir de las otras, se calcula mediante esta fórmula (Hair, Anderson, Tatham & Black, 1999):

$$KMO = \frac{\sum_{j \neq i} \sum_{i \neq j} r_{ij}^2}{\sum_{j \neq i} \sum_{i \neq j} r_{ij}^2 + \sum_{j \neq i} \sum_{i \neq j} r_{ij(p)}^2}$$

Donde r_{ij} son los coeficientes de correlación entre cada par de variables observadas, por otro lado, $r_{ij(p)}$ son los coeficientes de correlación parcial entre variables originales. Cuanto más alto sea el valor más dependencia lineal entre el conjunto de variables. Un valor de medida alrededor de 0.9 es muy bueno, mientras que un valor por debajo de 0.5 es demasiado pobre para llevar a cabo un análisis de factores.

A continuación, se procedió a llevar a cabo un análisis de factores para las 60 agencias, seleccionando las variables saldo de cartera, saldo de Pap, saldo de aportes, saldo de ahorros y saldo de Cdat como variables originales. En la tabla 3 se muestran las pruebas KMO y Bartlett, donde aparece la salida arrojada por SPSS; cómo se puede observar la prueba de esfericidad de

Bartlett (Valor P = 0.00) y el KMO permiten concluir que los datos son apropiados para implementar el análisis.

Tabla 3
Prueba de KMO y Bartlett

Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,757
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	292,800
	gl	10
	Sig.	,000

En la figura 11 se muestra el grafico de sedimentación la cual muestra el valor de los eingevalues (valores propios); se ha seleccionado un factor usando el criterio de los valores propios de la matriz de correlación sean mayor a 1. La varianza que recoge de las variables originales es igual a 70% de la variabilidad total. La matriz de componentes recoge las cargas o pesos del factor en cada una de las variables originales que deben servir de base para identificar la información que lleva el factor, en este caso vemos que son muy similares destacándose sólo el peso de la variable Aportes con un peso levemente superior al resto. A manera de conclusión, podemos decir que el análisis factorial nos sugiere que un solo factor puede reunir a las variables relacionadas con saldos: saldo de Pap, saldo de aportes, saldo de ahorros y saldo de Cdat el cual se llamará Total del negocio acorde a la terminología adoptada en la Cooperativa.

Figura 11
Sedimentación de las variables

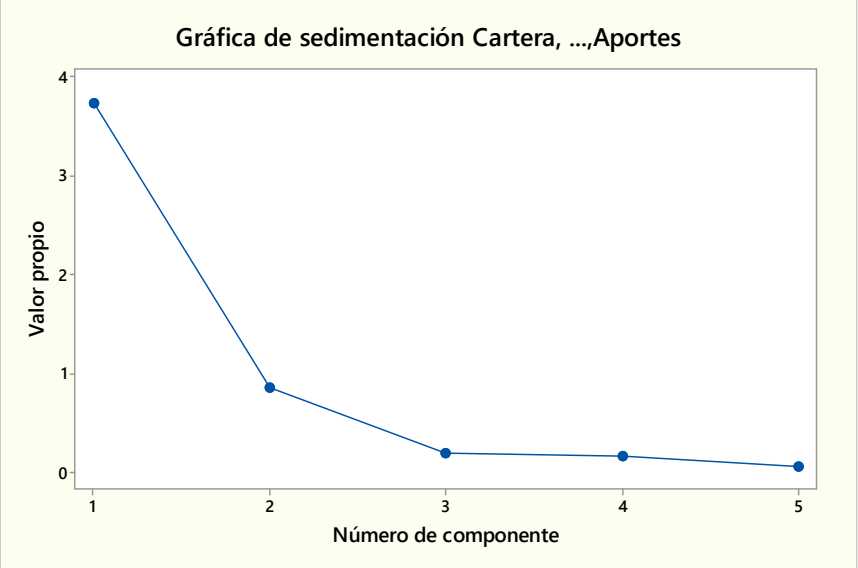


Tabla 4

Resultados del análisis factorial para las variables relacionadas con saldos

Varianza total explicada

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado	
	Total	% de varianza	% acumulado	Total	% de varianza
1	3,514	70,281	70,281	3,514	70,281
2	,843	16,863	87,143		
3	,299	5,972	93,115		
4	,191	3,820	96,936		
5	,153	3,064	100,000		

Matriz de componente[a]

	Componente
	1

CARTERA	,837
AHORROS	,834
CDAT	,803
PAP	,787
APORTES	,924

Método de extracción: análisis de componentes principales.
a. 1 componentes extraídos.

De igual forma se llevó a cabo un análisis factorial para las variables relacionadas con excedentes mensuales, los resultados del análisis se muestran en las tablas 5 y 6:

Tabla 5

Resultado del análisis factorial para las variables relacionadas con excedentes

Varianza total explicada

	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	7,340	61,168	61,168	7,340	61,168	61,168
2	1,176	9,800	70,969	1,176	9,800	70,969
3	,775	6,460	77,429			

Método de extracción: análisis de componentes principales.

Matriz de componente[a]

	Componente	
	1	2
ene_2019	,831	-,055
feb_2019	,861	-,148
mar_2019	,774	,187
abr_2019	,818	-,147
may_2019	,848	,194
jun_2019	,742	-,284

jul_2019	,827	-,179
ago_2019	,784	-,152
sep_2019	,844	,097
oct_2019	,367	,851
nov_2019	,791	-,255
dic_2019	,773	,349

Tabla 6

Prueba de KMO y Bartlett para excedentes por años

Prueba de KMO y Bartlett para excedentes por año

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,876
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	560,103
	gl	66
	Sig.	,000

Los criterios para determinar la retención del número de factores a extraer son la magnitud del valor propio (deber ser mayor a 1) o el porcentaje de varianza acumulada, en este caso a pesar de que el porcentaje de varianza explicado con un solo factor es sólo del 61%, se decide conservar una sola dimensión dado que la segunda dimensión estaría influenciada por una única variable, que son los excedentes del mes de octubre (ver matriz de componentes en Tabla 5), para esta variable se tiene que el año anterior presentó cierta anormalidad comparado con lo que históricamente ocurre. La siguiente es la matriz de pesos del factor eliminando dicho mes, advirtiendo que este cambio elevo a 65 % la varianza asociado con el uso de un solo factor. La conclusión es la viabilidad de utilizar un único factor que reúne el comportamiento de las variables en consideración el cual será referido como Excedentes total

Matriz de componente[a]

	Componente

	1
ene_2019	,831
feb_2019	,869
mar_2019	,768
abr_2019	,820
may_2019	,842
jun_2019	,748
jul_2019	,832
ago_2019	,787
sep_2019	,842
nov_2019	,803
dic_2019	,761

6.4.1 Análisis de clúster

Se realizan los cálculos para la obtención de las dos nuevas variables llamadas Total del negocio y Excedentes total por ser dos variables que describen en un alto grado la variación de las variables originales, que se presentan en la tabla 7.

En la siguiente sección se presenta el análisis de clúster con el objetivo de construir una clasificación de las 60 agencias que actualmente cuenta la Cooperativa. En la tabla 8 se muestran las variables originales y las nuevas variables sintéticas construidas a partir del análisis de factores que se realizó con anterioridad.

Tabla 7.
Variables utilizadas en el estudio de conglomerados

Variables analizadas	Variables utilizadas en conglomerados	Variables usadas en el perfilamiento
Saldo de Crédito	Saldo de Crédito	Total del negocio:
Saldo de ahorros	Saldo de ahorros	- Saldo de Crédito
Saldo de Cdat	Saldo de Cdat	- Saldo de ahorros
Saldo de Pap	Saldo de Pap	- Saldo de Cdat
Saldo de aportes	Saldo de aportes	- Saldo de Pap
Excedentes enero	Excedentes enero	- Saldo de aportes
Excedentes febrero	Excedentes febrero	Excedentes Total:
Excedentes marzo	Excedentes marzo	- Excedentes enero
Excedentes abril	Excedentes abril	- Excedentes febrero
Excedentes mayo	Excedentes mayo	- Excedentes marzo
Excedentes junio	Excedentes junio	- Excedentes abril
Excedentes julio	Excedentes julio	- Excedentes mayo
Excedentes agosto	Excedentes agosto	- Excedentes junio
Excedentes septiembre	Excedentes septiembre	- Excedentes julio
Excedentes octubre	Excedentes octubre	- Excedentes agosto
Excedentes noviembre	Excedentes noviembre	- Excedentes septiembre
Excedentes diciembre	Excedentes diciembre	- Excedentes octubre
Antigüedad agencias		- Excedentes noviembre
Cantidad empleados		- Excedentes diciembre

6.5 Selección y análisis de la solución clústeres (conglomerados)

La selección del número de conglomerados se basa en el método de Elbow (Elbow Method); una vez obtenidos los valores de Inercia después de aplicar el método de *k-means*, el valor donde se observó un cambio grande se ubica en 4 clústeres tal como se muestra a continuación:

Figura 12
Numero óptimo de clústeres

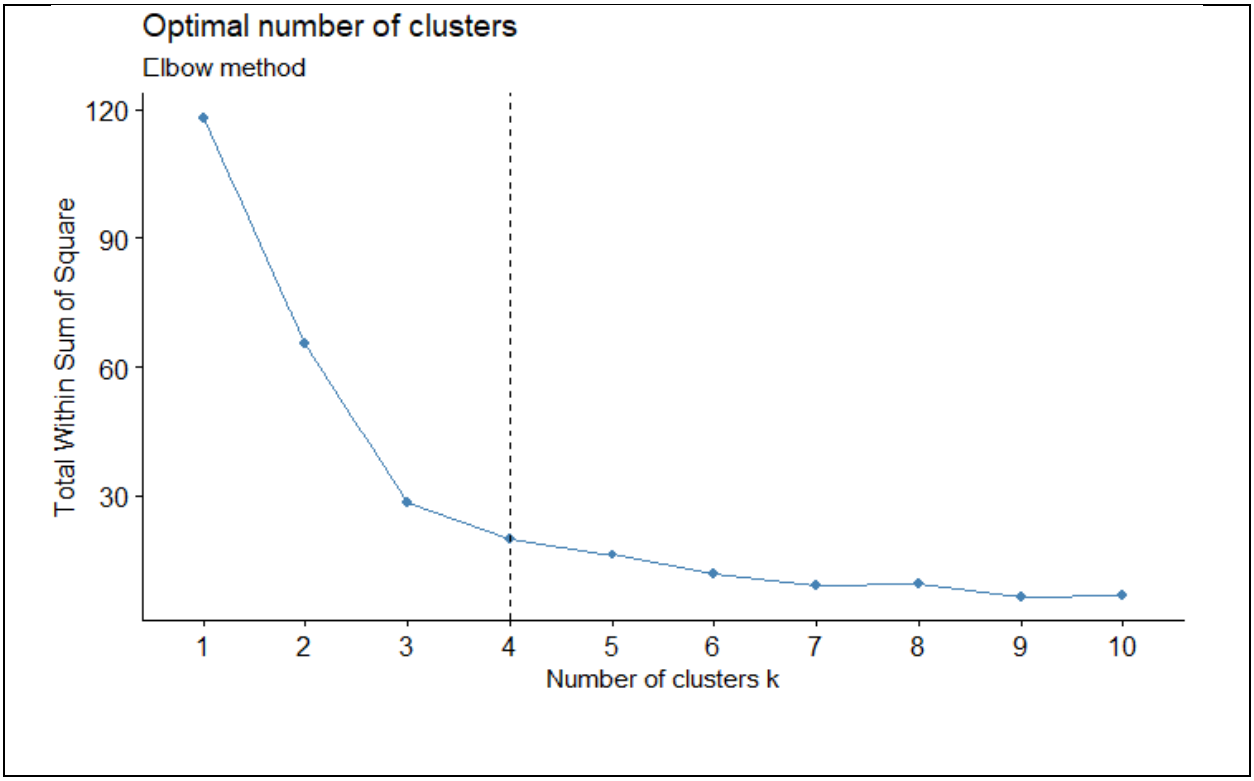


Figura 12 Método elbow, se logra identificar que con 4 clústeres se pueden clasificar las agencias.

Tabla 8

Distribución de agencias en los conglomerados

Agencias											
grupo 1			**grupo 2**				**grupo 3**			**grupo 4**	
11	19	56	74	36	25	62	86				
80	24	57	75		39	64	90				
	38	61	76		46	67	95				
	40	63	77		47	68	140				
	49	65	79		48	69	164				
	50	66	82		51	71	169				
	53	70	83		58	78	171				
	54	72	85		59	81	172				
	55	73	88		60	84	195				
			96				264				
			361								

Nota: Basada en datos de la Cooperativa con corte a diciembre de 2019.

En la Tabla 8 se muestra un agrupamiento para cada uno de los grupos como resultado de aplicar el método de k-medias tomando como solución k = 4 grupos, y como variables de segmentación el Total de negocio y Excedentes. Se puede observar que el grupo 3 consta de solo una agencia, que corresponde a la agencia de código 36 la cual es la agencia principal de la cooperativa, la agencia actualmente es la que tiene el mayor número de empleados y fue la primera que se fundó en Bucaramanga, además, maneja la mayor cantidad de excedentes y total negocio por tanto es de esperar que ninguna otra agencia exhiba su mismo comportamiento.

De otro lado, el grupo 4 consta de 28 agencias que tienen el menor valor de total negocio y aun no generan excedentes, estas agencias corresponden a las que tienen una menor antigüedad y menor número de empleados. El grupo 2 consta de 29 agencias que corresponden a las agencias que tienen valores similares en total negocio y excedentes, tienen una antigüedad superior a 10 años y un máximo de 30 empleados. Las 2 agencias que conforman el grupo 1 hacen parte de las agencias con mayor antigüedad, pero cuyo valor de total negocio y excedentes no superan los de la agencia del grupo 3. La Tabla 9 presenta un resumen descriptivo básico por clúster.

Tabla 9.
Solución total negocio y excedentes

solución con total negocio y excedentes		
	EXCEDENTES	TOTAL_NEGOCIO
Media	1,649E+09	6,224E+10

Tamaño cluster	7	7
Des_estándar	433075413,3	18610611535
Media	3,722E+09	2,217E+11
N	1	1
Des_estándar	.	.
Media	7,248E+08	4,317E+10
Tamaño cluster	24	24
Des_estándar	284460370,5	13948947567
Media	-5,503E+07	1,682E+10
Tamaño cluster	28	28
Des_estándar	261661266,5	12627257462

La Tabla 9 permite comparar la categorización que hoy maneja la Cooperativa con la solución clúster descrita anteriormente. La nueva propuesta sugiere clasificar las agencias en 4 categorías en lugar de las 6 categorías que actualmente se manejan.

Esto comparado con la categorización de 8 niveles o categorías que tiene la cooperativa actualmente y tomando las variables total negocio y excedentes, permite identificar que en el nivel o categoría 6 se ubica la agencia de código 36 como la única agencia que pertenece a este grupo el cual corresponde al grupo que maneja los mayores valores para estas variables, adicional es la que tiene el mayor número de empleados y una mayor antigüedad. Para el nivel o categoría 1, se ubican un total de 23 agencias las cuales corresponden aquellas cuyo valor de total negocio son

inferiores a los $ 23.000 millones de pesos y los excedentes se encuentran por debajo de los $466millones de pesos cuya antigüedad no es mayor a los 5 años.

6.6 Resultados Análisis algoritmo *Fuzzy C-Means*

La información con la cual se aplica el algoritmo *fuzzy c-means* consta de 6 variables saldo de crédito, saldo de ahorros, saldo de Cdat, saldo de Pap, saldo de aportes y excedentes Total, con fecha a diciembre de 2019 para un total de 60 agencias, aplicando para ello la técnica de análisis de datos *C-means*. Se determina el número de clúster (en total son 4) para el conjunto de datos y así obtener una mejor visualización.

6.6.1 Análisis general

Las observaciones están representadas por puntos en el gráfico (Ver figura 13) donde se visualiza una figura alrededor de cada grupo. La distribución de las agencias en cada uno de los clústeres es como se muestra en la tabla a continuación, posteriormente en la Tabla 11 se presenta una descripción que resume la variable valores (niveles) de membresía en cada cluster.

Tabla 10.
Distribución de agencias en cada clúster

Clúster	Cantidad
1	23
2	1

3	16
4	19

Nota: Basada en datos de la Cooperativa con corte a diciembre de 2019.

Tabla 11

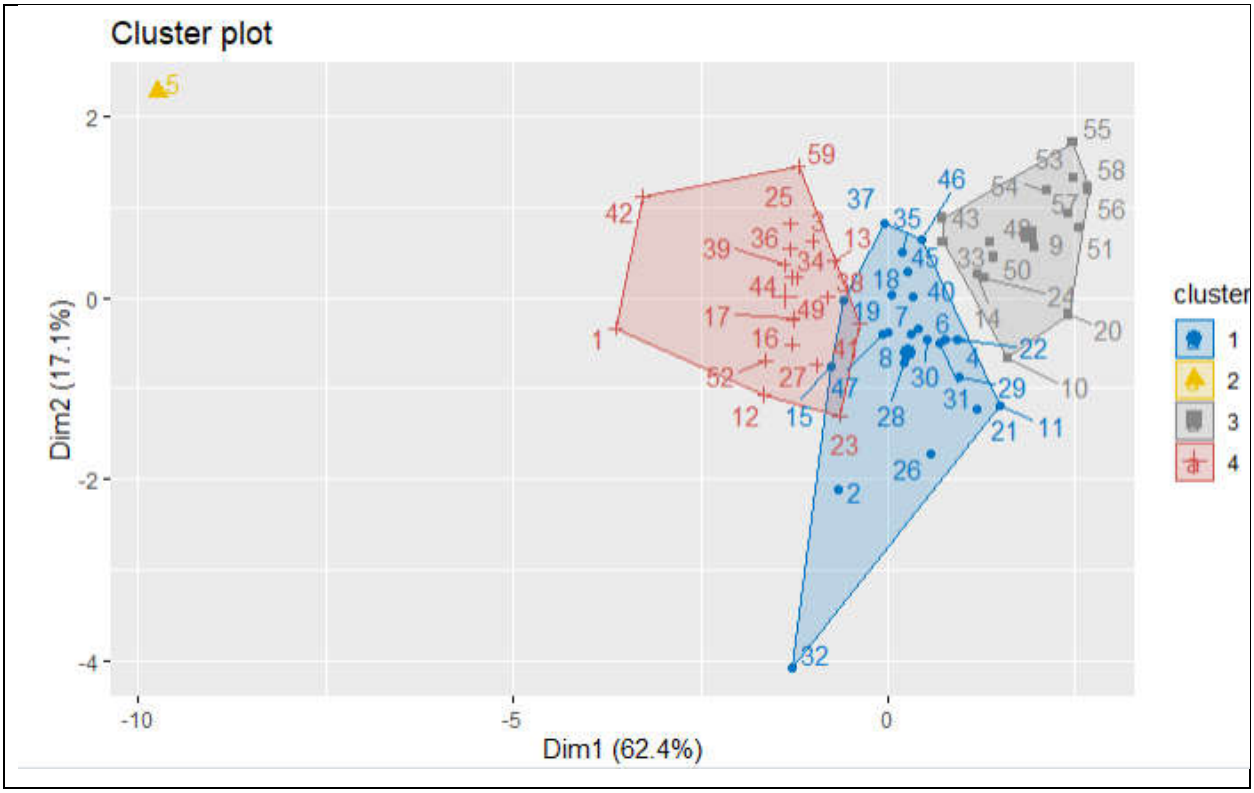
Estadística descriptiva de los valores de pertenencia o membresía de los clústeres

Estadísticas descriptivas de los grados de membresía por conglomerados							
	Size	Min	Q1	Mean	Median	Q3	Max
Cluster 1	23	0,3945933	0,5161545	0,6132411	0,5816723	0,713203	0,8699858
Cluster 2	1	0,9989333	0,9989333	0,9989333	0,9989333	0,9989333	0,9989333
Cluster 3	16	0,3921485	0,618326	0,7423285	0,7902206	0,8793588	0,9923201
Cluster 4	19	0,4290023	0,5387045	0,6382897	0,6530704	0,8249078	0,8852886

En la Figura 13 se puede observar que el clúster 1 conformado por 23 agencias, corresponden a todas aquellas que poseen una media inferior a los demás clústeres, lo que significa que los valores de las variables para estas agencias son muy bajas, este hecho que puede atribuirse a su corto tiempo de vigencia en la cooperativa. El clúster 2 que lo conforma una única agencia corresponde a aquella que dentro del manejo para cada una de las variables posee los mayores valores y la ubican como la única en su rango.

Figura 13

Visualización de métodos de partición.



Nota: Basada en datos de la cooperativa con corte a diciembre de 2019.

Tabla 12.

Matriz de adhesión- Grado de pertenencia de cada agencia a un clúster

Grado de pertenencia

Membership degree matrix:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4		Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.18	0.10	0.37	0.35	31	0.76	0.08	0.09	0.07
2	0.33	0.12	0.35	0.20	32	0.28	0.14	0.34	0.24
3	0.24	0.10	0.39	0.28	33	0.47	0.30	0.13	0.10
4	0.89	0.05	0.04	0.03	34	0.11	0.04	0.19	0.66
5	0.23	0.18	0.29	0.30	35	0.45	0.13	0.21	0.21
6	0.73	0.06	0.12	0.08	36	0.11	0.04	0.30	0.54
7	0.49	0.14	0.20	0.17	37	0.32	0.14	0.25	0.29
8	0.56	0.08	0.22	0.14	38	0.14	0.04	0.23	0.59
9	0.03	0.96	0.01	0.01	39	0.13	0.06	0.23	0.58
10	0.48	0.34	0.10	0.08	40	0.53	0.11	0.18	0.19
11	0.50	0.26	0.14	0.10	41	0.26	0.07	0.26	0.41
12	0.15	0.06	0.34	0.45	42	0.16	0.09	0.30	0.44
13	0.16	0.04	0.39	0.41	43	0.38	0.30	0.16	0.16
14	0.41	0.44	0.09	0.06	44	0.03	0.01	0.10	0.86
15	0.28	0.06	0.51	0.16	45	0.49	0.13	0.20	0.18
16	0.04	0.01	0.88	0.08	46	0.44	0.19	0.19	0.18
17	0.11	0.03	0.55	0.31	47	0.50	0.06	0.26	0.18
18	0.59	0.06	0.20	0.14	48	0.27	0.59	0.07	0.07
19	0.32	0.08	0.42	0.19	49	0.13	0.05	0.34	0.48
20	0.21	0.64	0.08	0.06	50	0.25	0.63	0.07	0.05
21	0.48	0.24	0.17	0.11	51	0.05	0.91	0.02	0.02
22	0.67	0.16	0.11	0.07	52	0.18	0.09	0.29	0.44
23	0.26	0.07	0.44	0.23	53	0.05	0.90	0.02	0.02
24	0.43	0.42	0.08	0.07	54	0.07	0.88	0.03	0.03
25	0.08	0.03	0.20	0.69	55	0.10	0.80	0.05	0.05
26	0.49	0.14	0.22	0.15	56	0.06	0.88	0.03	0.03
27	0.10	0.02	0.66	0.22	57	0.02	0.96	0.01	0.01
28	0.61	0.07	0.20	0.12	58	0.06	0.88	0.03	0.03
29	0.68	0.14	0.11	0.07	59	0.21	0.12	0.31	0.35
30	0.86	0.04	0.06	0.04					

En la Tabla 12 se identifica el grado de pertenencia para cada agencia con relación al clúster, donde el grado más alto, significa que la agencia tiene una mayor representación hacia ese clúster en particular.

Como se puede observar el clúster 1 representa a las agencias con menor antigüedad en la cooperativa por lo que sus saldos son menores. En una menor proporción el clúster 2 corresponde a la agencia con un mayor periodo de antigüedad y sus saldos son los mayores en la cooperativa. Para los clústeres 3 y 4 se observa que corresponde a aquellas agencias que sus saldos están en

proceso de incremento. Al utilizar el algoritmo *fuzzy c-means* se puede visualizar el grado de membresía, y para el análisis es ideal, ya que una agencia puede tener similares grados de membresía en dos clústeres y permite no excluirla con relación al clúster elegido.

Tabla 13

Categorización actual de las agencias frente a la nueva segmentación

CODIGO AGENCIA	CATEGORIZ. ACTUAL AGENCIA	ANALISIS DE CLÚSTER AGENCIA
11	4	1
80	5	1
36	6	3
19	3	2
24	3	2
38	2	2
40	3	2
49	3	2
50	3	2
53	3	2
54	3	2
55	3	2
56	2	2
57	3	2
61	2	2
63	4	2
65	3	2
66	2	2
70	2	2
72	3	2
73	2	2
25	2	4
39	2	4
46	1	4
47	1	4
48	1	4
51	1	4
58	1	4
59	1	4
60	1	4
62	1	4
64	1	4
67	1	4
68	2	4
69	1	4
71	2	4
78	1	4
81	1	4
84	2	4

74	4	2	86	1	4
75	2	2	90	1	4
76	4	2	95	1	4
77	3	2	140	1	4
79	2	2	164	1	4
82	4	2	169	1	4
83	2	2	171	1	4
85	2	2	172	1	4
88	3	2	195	1	4
96	3	2	264	1	4
361	3	2			

Al aplicar el análisis de clúster se puede visualizar que las agencias 11 y 80, que actualmente en la cooperativa están clasificadas en las categorías 4 y 5 respectivamente, se puede visualizar que se encuentran ubicadas en el grupo 1, esto debido a que los valores que manejan de total negocio y excedentes si bien se pueden encontrar en el mismo rango también se encuentran dentro de los límites entre una y otra categoría.

En el grupo 2 del análisis de clúster existen 29 agencias que en la clasificación actual se encuentran clasificadas en las categorías 2, 3 y 4, lo que permite identificar que los valores de total negocio se encuentran entre los $23.700 millones y los $74.600 millones aproximadamente y los excedentes se encuentran entre los $466 millones y los $1767 millones por lo que se podría implementar como una sola categoría permitiendo una clasificación más homogénea para las agencias y posiblemente otorgar un ajuste salarial equitativo para ellas.

El grupo 3 del análisis de clúster al igual que la categoría 6 en la clasificación actual de las agencias de la cooperativa coinciden en la agrupación de la misma agencia de Código 36 por ser la que cuenta con el máximo valor para total negocio y excedentes.

Algo similar ocurre con el grupo 4 del análisis de clúster el cual reúne las agencias que en la clasificación actual de la cooperativa se encuentran en la categoría 1 y 2 las cuales corresponden a los niveles de clasificación más bajos, esto debido a que los valores de las variables de total negocio y excedentes de la actual categoría son 2, se encuentran en los límites entre estas dos categorías, son los más bajos debido a que tienen un corto tiempo de antigüedad en la cooperativa, o a la poca demanda del mercado que pueda existir en algunas de las áreas geográficas y esto permite visualizar lo bajo de sus resultados.

Con el análisis de clúster se propone una nueva alternativa de segmentación con un numero óptimo de 4 categorías, y no de 6 como se maneja actualmente, que clasifican las 60 agencias de la cooperativa con un mayor grado de homogeneidad entre los elementos de cada grupo, a su vez brinda una clasificación más equitativa basada en los resultados obtenidos por cada una de las agencias, y que posiblemente permita a mediano plazo brindar estímulos motivacionales para los empleados de las agencias como la opción de realizar un ajuste salarial de acuerdo a la segmentación propuesta.

7. Conclusiones

Para el proceso de investigación de la segmentación de las agencias de la cooperativa de ahorro y crédito, se utilizó los métodos clustering hard y *fuzzy c-means*. Se desarrolla el marco teórico estudiando los algoritmos *K-means* y *C-means* y donde se logra diferenciar que este último permite identificar que una agencia pueda pertenecer a varios clústeres, lo que nos da la posibilidad de poder mover una agencia de una categoría a otra mediante ciertas estrategias que induzcan a la agencia a trabajar para lograr esta movilidad. El análisis de factores permitió comprobar que es adecuado seguir con los mismos criterios actuales, esto es, clasificando las agencias por medio de las dos variables total negocio y excedentes total.

Durante el desarrollo del proyecto, se realizó el proceso de extracción, recolección y limpieza de los datos de las agencias para los meses del año 2019. Un conjunto de estas variables se utilizó en el análisis de clúster y en el clúster *fuzzy*, las otras variables se utilizaron para la descripción de los grupos. La clasificación nítida permitió la realización de una nueva categorización de las agencias que se considera más transparente y equitativa, la nueva propuesta considera 4 categorías en lugar de las 6 que actualmente se tienen.

Con la clasificación difusa se consiguió la misma cantidad de categorías, lo que nos permite sugerirla a la dirección como una buena herramienta para determinar cuál agencia podría encontrase bien sea en peligro (disminuir su categoría), o cual puede aumentar su categoría de acuerdo a los grados de pertenencia a los grupos construidos por el algoritmo *c-means*.

En aras de la equidad se recomienda la posibilidad de incluir nuevas variables para la construcción de las categorías donde se tenga en cuenta, por ejemplo, la antigüedad de la agencia, el número de empleados, tecnología y un presupuesto adecuado para el crecimiento. Para la implementación de la propuesta se recomienda automatizarla en un software adecuado.

De esta forma, se da por cumplido los objetivos propuestos en el proyecto de investigación y se puede determinar que con un método de conglomerados se puede clasificar de una forma más completa las agencias de la cooperativa y se puede llegar a contemplar que ellas pertenezcan a dos grupos o que se acerque a un nivel más alto en el que se encuentre actualmente.

Referencias bibliográficas

Amat, J. (2017). *Clustering y heatmaps - aprendizaje no supervisado*. Recuperado el 30 de marzo de 2020 de RPubs https://rpubs.com/Joaquin_AR/310338

Artola, M., Morettini, M. y Blanco, G. (2014). *Financiamiento y tamaño de las universidades publicas argentinas: un análisis de clustering hard y difuso*. Recuperado en mayo de 2020 de Repositorio Institucional UFSC https://repositorio.ufsc.br/xmlui/handle/123456789/132235

Azar, A., El-Said, S., & Hassanien, A. (2013). *Fuzzy and hard clustering analysis for thyroid disease. Computer methods and Programs in Biomedicine,* 111(1), 1 – 16. Recuperado en mayo de 2020 de ScienceDirect https://doi.org/10.1016/j.cmpb.2013.01.002

Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: *The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2–3), 191–203.* Recuperado en 2020 de ScienceDirect https://doi.org/10.1016/0098-3004(84)90020-7

Campello, R.JGB., & Hruschka, E.R., (2006). *A fuzzy extensión of the silhouette width criterion for cluster analysis. Fuzzy Sets and Systems*, 157(21), 2858 – 2875. Recuperado en mayo de 2020 de ScienceDirect https://doi.org/10.1016/j.fss.2006.07.006

Döring, C., Lesot, M.-J., & Kruse, R. (2006). *Data analysis with fuzzy clustering methods. Computational Statistics & data analysis,* 51, 192 – 214. Recuperado en mayo de 2020 de ScienceDirect https://www.sciencedirect.com/science/article/pii/S0167947306001307

Fuzzy clustering. (2019). En Wikipedia. Recuperado en mayo de 2020 de https://es.wikipedia.org/wiki/Fuzzy_clustering

George, J., Klir y Bo, Y. (1995). *From ordinary (crisp) sets to fuzzy sets: a grand paradigm shift. Fuzzy sets and fuzzy logic: Theory and applications*. [Archivo PDF]. Recuperado de http://www.b-farhadinia.ir/bfarhadiadmin/file/stdfile/Klir.pdf

Grekousis & Thomas, (2012). *Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods*. Recuperado en mayo de 2020 de ScienceDirect https://doi.org/10.1016/j.apgeog.2011.11.004

Hair, Anderson, Tatham & Black. (1999). *Multivariate Data Analisys. Pearson Education Limited.*

Kaufman y Rousseeuw. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Recuperado en mayo de 2020 de https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

Kuncheva, L. I. (Ludmila I. (2004). *Combining pattern classifiers : methods and algorithms.* Recuperado en mayo de 2020 de Wiley https://www.wiley.com/en-us/Combining+Pattern+Classifiers%3A+Methods+and+Algorithms%2C+2nd+Edition-p-9781118315231

Lara, A., (2020). *Métodos de análisis multivariante análisis cluster (practica 8).* Recuperado en marzo de 2020 de Estadística Universidad de Granada http://wpd.ugr.es/~bioestad/guia-spss/practica-8/

Morissette & Chartier., (2013). *The k-means clustering technique: General considerations and implementation in Mathematica. 15-24.* Recuperado en 2020 de Tutorials in Quantitative Methods for Psichology http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.667.159&rep=rep1&type=pdf

Ponce, P. (2010). *Inteligencia artificial con aplicaciones a la ingeniería*. [Archivo PDF]. Recuperado de https://lelinopontes.files.wordpress.com/2014/09/inteligencia-artificial-con-aplicaciones-a-la-ingenierc3ada.pdf

Reina, D. (2008). *Fundamentos de matemática difusa. 6 – 12.* [Archivo PDF]. Recuperado de http://www.konradlorenz.edu.co/images/stories/suma_digital_matematicas/EDICION_09_0I/trabajo_de_grado_daniel_reina.pdf

Santiago de la Fuente Fernandez. (2011). *Análisis conglomerados.* [Archivo PDF]. Recuperado de http://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf

Uriel E, Aldás J. (2005). *Análisis Multivariante Aplicado. Thomson*.

Apéndices

Apéndice A Código en R

```
install.packages("ppclust")

library(ppclust)

install.packages("dplyr")
library(dplyr)
library(tidyverse)
library(factoextra)
library(cluster)
library(fclust)


Cargar el archivo de datos

library(readxl)
fuzzybase_proyecto <-
read_excel("C:/Users/trabajofc/Documents/Proyecto/base_proyecto.xlsx")
View(fuzzybase_proyecto)

x=fuzzybase_proyecto[,-1:-6]
x

pairs(x)


Empezamos hacer el clustering difuso, y se debe dar los centros

res.fcm <- fcm(x, centers=4)
v0 <- matrix(nrow=4, ncol=6,
    c(1.5239164,         1.2919351,1.3037109,0.6328540,2.8742195,1.89288365,
      0.4621021,0.0733253,0.0358708,0.5388258,0.2414156,0.2441960,
      3.1658277,6.3478668,6.5249239, 1.8150637, 3.8236991, 4.2023094,
      -0.9532666,-0.4967303,-0.4500848,-0.9304717, -0.7730551,-0.7082930),
      byrow=TRUE)
print(v0)


la data y los centros
```

```
res.fcm <- fcm(x, centers=v0)
```

Veamos otra inicialización de los centros. v, la matriz de prototipo de clúster se inicializa usando K-means mediante el algoritmo (kmpp) en el paquete R inaparc

```
v0 <- inaparc::kmpp(x, k=3)$v
print(v0)

res.fcm <- fcm(x, centers=v0)
```

La matriz de grado de pertenencia es una salida de la función fcm.

```
res.fcm <- fcm(x, centers=4)
as.data.frame(res.fcm$u)
```

Las instrucciones dan los centroides iniciales y finales de la soluciÃ³n clustering

```
res.fcm$v0
res.fcm$v
```

Podemos tener un resumen general del procedimiento realizado por R

```
summary(res.fcm)
```

Visualización de los resultados. Veamos algunos

```
plotcluster(res.fcm, cp=1, trans=TRUE)
```

Algunas versiones nuevas se tienen en R, por ejemplo la función fviz_cluster del paquete factoextra (Kassambara & Mundt, 2017).

Paso:
1. Se hace un clustering difuso (ppclust)
2. luego se convierte el objeto kmeans usando ppclust2 del paquete ppclust como se muestra en la primera línea:

```
res.fcm2 <- ppclust2(res.fcm, "kmeans")
```

```
factoextra::fviz_cluster(res.fcm2, data = x,
  ellipse.type = "convex",
  palette = "jco",
  repel = TRUE)
```

La Cluster validation es un proceso del buen ajuste de los resultados del procedimiento. Existe varios índices:

Observación, clustering es un anàlisis no supervisado, así que no se usa informacion externa, por consiguiente, indices internos son usados para validar los resultados del clustering.

En el entorno de R difuso, fclust package (Ferraro & Giordani, 2015) se usan, Partition Entropy (PE), Partition Coefficient (PC) y Modified Partition Coefficient (MPC), y Fuzzy Silhouette.

```
res.fcm4 <- ppclust2(res.fcm, "fclust")
idxsf <- SIL.F(res.fcm4$Xca, res.fcm4$U, alpha=1)
idxpe <- PE(res.fcm4$U)
idxpc <- PC(res.fcm4$U)
idxmpc <- MPC(res.fcm4$U)
```

Índices

```
cat("Partition Entropy: ", idxpe)
cat("Partition Coefficient: ", idxpc)
cat("Modified Partition Coefficient: ", idxmpc)
cat("Fuzzy Silhouette Index: ", idxsf)
```

Otra manera de visualizar la soluciÃ³n cluster

```
res.fcm3 <- ppclust2(res.fcm, "fanny")

cluster::clusplot(scale(x), res.fcm3$cluster,
  main = "Cluster plot of Iris data set",
  color=TRUE, labels = 2, lines = 2, cex=1)
```