UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2017

Método de Cluster Jerárquicos para Datos Funcionales Multivariados: Una Aplicación en Mercadeo

Hierarchical Cluster Method for Multivariate Functional Data: An Application in Marketing

Gustavo Adolfo Venegas.[a] Wilmer Pineda Ríos.[b]
gustavovenegas@usantotomas.edu.co wilmerpineda@usantotomas.edu.co

Resumen

En el presente trabajo se propone un método para la segmentación y caracterización de cluster jerárquicos en datos funcionales multivariados; se explicara paso por paso el algoritmo construido con su método de validación, también se presentara la aplicación en el área de marketing a la segmentación de clientes de un almacén de Retail, en el cual se evaluara tres variables: Recencia, Promedio de Compra Mensual y Monto Vendido en el periodo analizado. Como resultado final será la segmentación de clientes en cluster o grupos cada uno caracterizado.

Palabras clave: Datos funcionales Multivariados, Cluster Jerárquicos, Segmentación de mercado.

Abstract

In the present work a method for the segmentation and characterization of hierarchical clusters in multivariate functional data is proposed; the algorithm constructed with its validation method will be explained step by step, the application in the marketing area will also be presented to the customer segmentation of a Retail warehouse, and three variables will be evaluated: Receipt, Average Monthly Purchase and Amount Sold in the period analyzed. The final result will be the segmentation of clients into clusters or groups each characterized.

Key words: Functional Data Multivariate, Hierarchical Cluster, Market Segmentation.

[*]Estudiante de Estadística Universidad Santo Tomas Bogotá
[†]Docente de Estadística Universidad Santo Tomas Bogotá

1. Introducción

Actualmente el manejo, compresión y análisis, de datos masivos se ha vuelto más complicado, y es habitual encontrar elementos de análisis que son datos funcionales, de manera que la falta de un algoritmo para generar clusters jerárquicos en datos funcionales multivariados se vuelve cada vez más importante. Es así que, en este contexto, se propone la construcción de un método para la caracterización de cluster jerárquicos en datos funcionales multivariados, con una aplicación a la segmentación de clientes. Además, es un aporte importante utilizar la variabilidad que tiene un individuo y que es representada bajo el enfoque de los datos funcionales.

En la estadística clásica multivariada se ha estudiado ampliamente el concepto de métodos jerárquicos que permiten aglomerar o dividir clusters para crear nuevos, que desde el enfoque clásico es indispensable la construcción de la matriz de distancias o similitudes que según Peña (2002),

"Una matriz de distancias o de similitudes se desea clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera que los elementos son sucesivamente asignados a los grupos, pero la asignación es irrevocable, es decir, una vez hecha, no se cuestiona nunca más. Los algoritmos son de dos tipos: De aglomeración y de división."

Con referencia a lo anterior, se resalta que el procedimiento tiene la misma estructura y se diferencia por la manera de calcular las distancias entre grupos siendo la distancia de Ward la más relevante para el presente trabajo ya que integra en su cálculo la variabilidad de los individuos.

El objetivo principal de esta investigación es establecer un método para la caracterización de clusters jerárquicos en datos funcionales multivariados con una aplicación a la segmentación de clientes. Para encaminar la investigación se determinara la distancia más adecuada para construir la matriz de similitudes en datos funcionales multivariados y se definirá los clusters más convenientes para los clientes de un almacén de consumo masivo por medio de su comportamiento de compra.

En los estudios recientes de datos funcionales para el caso multivariado es conocido el avance en las agrupaciones no supervisadas K-means (análisis de componentes principales, funcional (Jacques, Preda 2014; Yamamoto, 2012; Berrendero, Justel, Svarc, 2011)), y en los clusters jerárquicos para un dato funcional (Santiago y otros, 2017). Abriendo el campo para el desarrollo de clusters jerárquicos de dos o más datos funcionales que es la construcción principal del presente trabajo.

2. Marco de Referencia

2.1. Datos Funcionales

2.1.1. ¿Qué es un dato funcional?

Definición 1: Un dato funcional $f(t), t \in T \subset \mathbb{R}$, se representa como un conjunto finito de pares (t_i, x_i), $t_i \in T, i = 1, 2, ..., N$ donde N representa la cantidad de puntos de la variable funcional de interés

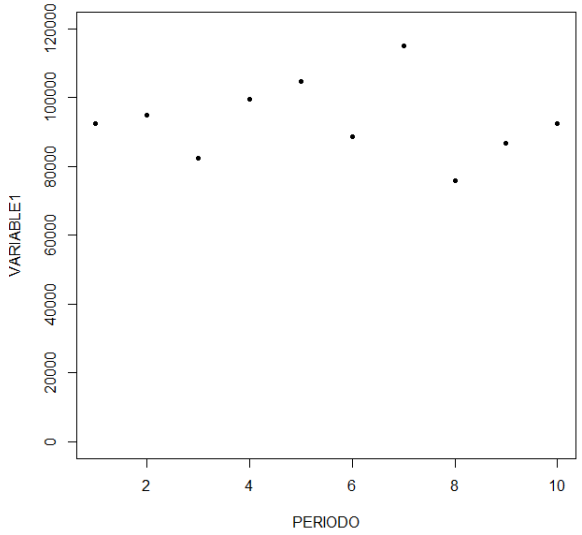

Imagen 1. Fuente: Elaboración Propia.

2.1.2. Definición $L^2(T)$

Sea $L^2(T)$, con $T = [a, b] \subset \mathbb{R}$ el espacio de las funciones cuadrado integrable (Espacio de Hilbert)

$$L^2(T) = \left\{ \ f : \Re \Rightarrow \Re | \int_a^b f(t)^2 dt < \infty \ \right\} \qquad (1)$$

Con producto interno

$$< a, b >= \int_a^b f(t)g(t)dt \qquad (2)$$

2.1.3. B-Splines

Una función spline de grado m en el intervalo $[0, 1]$ es una función tal que, dados n nodos $0 \leq x_1 < x_2 < ... < x_n \leq 1$, está definida como un polinomio de grado m en cada uno de los intervalos $[0, x_1), ..., (x_j, x_{j+1}), ..., (x_n, 1]$

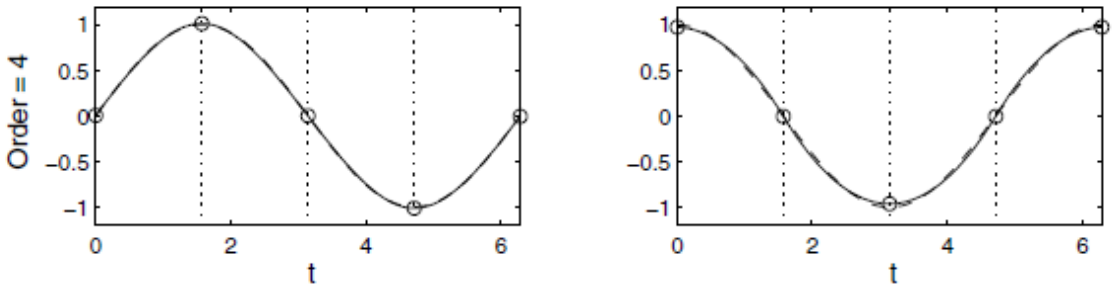

Imagen 2. Fuente: Ramsay, J. O. y Silverman, B.W. (2009)

2.2. Algoritmos Jerárquicos enfoque clásico

Los algoritmos jerárquicos se construyen a partir de una matriz de distancia o similitudes, los algoritmos actuales funcionan según Peña (2002) como elementos o individuos que,

"son sucesivamente asignados a los grupos, pero la asignación es irrevocable, es decir, una vez hecha, no se cuestiona nunca más. Los algoritmos son de dos tipos:

1. De aglomeración. Parten de los elementos individuales y los van agregando en grupos.

2. De división. Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Los algoritmos de aglomeración requieren menos tiempo de cálculo y son los más utilizados" (p. 244)

2.2.1. El método de Ward

Ward indico que la pérdida de información al realizar el proceso de segmentación bajo clusters puede medirse a través de la suma total de los cuadrados de las desviaciones entre cada punto (individuo) y la media del cluster que lo contiene. En cada paso del análisis, considerar la posibilidad de la unión de cada par de grupos y optar por la fusión de aquellos dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse. Este método a diferencia de los demás,

- Encadenamiento simple o vecino más próximo.
- Encadenamiento completo o vecino más alejado.
- Media de grupos.
- Método del centroide

Ya que minimiza la varianza dentro de cada cluster. Según Peña (2002),

"Esta medida suma las distancias de euclídeas al cuadrado entre cada elemento y la media de su grupo:

$$W = \sum_g \sum_{i \in G} (X_i g - \hat{X}_g)'(X_i g - \hat{X}_g) \qquad (3)$$

Donde $\hat{X}_g$ es la media del grupo g. El criterio comienza suponiendo que cada dato forma un grupo, $g = n$ y por tanto W (8.7) es cero. A continuación se unen los elementos que produzcan el incremento mínimo de W. Obviamente esto implica tomar los más próximos con la distancia euclídea. En la siguiente etapa tenemos $n1$ grupos, $n2$ de un elemento y uno de dos elementos. Decidimos de nuevo que dos grupos unir para que W crezca lo menos posible, con lo que pasamos a $n2$ grupos y así sucesivamente hasta tener un único grupo. Los valores de W van indicando el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuantos grupos naturales contienen nuestros datos." (p. 244)

2.2.2. Dendograma

Según Peña (2002), "El dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol. Los criterios para definir distancias que hemos presentado tienen la propiedad de que, si consideramos tres grupos, A, B, C, se verifica que

$$d(A, C) \leq max\{d(A, B), d(B, C)\} \qquad (4)$$

y una medida de distancia que tiene esta propiedad se denomina *ultramétrica*. Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto si $d^2(A, C)$ es menor o igual que el máximo de $d^2(A, B), d^2(B, C)$ forzosamente será menor o igual que la suma $d^2(A, B) + d^2(B, C)$. El dendrograma es la representación de una ultramétrica, y se construye como sigue:

1. En la parte inferior del gráfico se disponen los n elementos iniciales.
2. Las uniones entre elementos se representan por tres líneas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos están conectados por líneas rectas.

Si cortamos el dendrograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

El dendrograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se aplica ciegamente, ya que dos puntos pueden parecer próximos cuando no lo están, y pueden aparecer alejados cuando están próximos.(p. 247)

3. Método de Cluster Jerárquicos para Datos Funcionales Multivariados:

La metodología se planteó en cuatro fases principales:

Fase 1: Construcción de la base de datos. En primer lugar se realiza la construcción de la base de datos a través de la simulación de 30 individuos con comportamientos previamente definidos con el objetivo de conocer a priori el comportamiento de los individuos en 3 grupos (cada uno de 10 individuos) y con cada grupo de individuos se tendrá en cuenta 3 variables funcionales.

Fase 2: Suavizamiento y Estandarización En segundo lugar se realiza el suavizamiento con funciones de B-Splines para después estandarizar las variables funcionales.

Fase 3: Matriz de similitudes. Para la construcción de la matriz de Similitudes o de distancias, se halla la distancia entre los individuos para cada variable funcional y después se suman matricialmente apara obtener la matriz de Similitudes definitiva.

Fase 4: Segmentación de Cluster Jerárquicos, tabla de validación y criterio de Exactitud. En cuarto lugar se realiza el método clásico para cluster jerárquicos bajo la matriz de Similitudes definitiva y se construye tabla de validación cruzada de los grupos construidos a priori vs los cluster definidos con la aplicación del método. Por último se construye el criterio de exactitud.

3.1. Construcción de la base de datos.

Inicialmente se realiza la simulación para la primera variable de los 30 individuos. Para los primeros 10 individuos se distribuye con la ecuación (1), para los siguientes 10 individuos se distribuye con la ecuación (2) y para los últimos 10 individuos se distribuye con la ecuación (3), así:

$$f(x) = x^2 \text{ con } 2 \leq x \leq 11 \quad (5)$$
$$f(x) = 10x \text{ con } 1 \leq x \leq 10 \quad (6)$$
$$f(x) = \frac{x^3}{6} \text{ con } 1 \leq x \leq 10 \quad (7)$$

El cuadro (1) indica la media y desviación para la simulación de cada individuo. Se evidencia que la media de los 30 individuos es similar.

Cuadro 1: *Simulación variable 1.*

Indiv	**Media**	**Desv**	**Indiv**	**Media**	**Desv**	**Indiv**	**Media**	**Desv**
1	59.8	45.5	**11**	60.7	29.8	**21**	58.5	63.6
2	58.0	43.7	**12**	58.2	30.3	**22**	59.0	66.0
3	58.3	45.8	**13**	59.9	29.4	**23**	64.0	74.0
4	55.8	42.4	**14**	59.0	29.7	**24**	59.3	63.4
5	57.3	40.9	**15**	60.5	30.7	**25**	62.2	68.5
6	58.3	43.3	**16**	60.0	30.7	**26**	61.9	67.6
7	56.9	42.0	**17**	59.7	30.1	**27**	59.4	68.7
8	56.7	41.0	**18**	59.7	31.8	**28**	61.2	68.3
9	57.1	42.4	**19**	59.9	30.1	**29**	59.8	64.7
10	59.5	45.8	**20**	61.2	31.6	**30**	58.6	64.9

Para la segunda variable igualmente se realiza la simulación de 30 individuos. Para los primeros 10 individuos se distribuye con la ecuación (4), para los siguientes 10 individuos se distribuye con la ecuación (5) y para los últimos 10 individuos se distribuye con la ecuación (6), así:

$$f(x) = cos(x) \text{ con } 1 \leq x \leq 10 \quad (8)$$
$$f(x) = sen(x) \text{ con } 1 \leq x \leq 10 \quad (9)$$
$$f(x) = tan(\tfrac{x}{10}) \text{ con } 1 \leq x \leq 10 \quad (10)$$

El cuadro (2) indica la media y desviación para la simulación de cada individuo para la segunda variable. Se evidencia que la media de los 30 individuos es similares.

Cuadro 2: *Simulación variable 2.*

Indiv	**Media**	**Desv**	**Indiv**	**Media**	**Desv**	**Indiv**	**Media**	**Desv**
1	- 0.22	0.66	**11**	0.08	0.74	**21**	- 0.23	0.50
2	- 0.14	0.70	**12**	0.02	0.81	**22**	- 0.22	0.54
3	- 0.17	0.67	**13**	0.13	0.72	**23**	- 0.22	0.52
4	- 0.18	0.61	**14**	0.04	0.83	**24**	- 0.22	0.55
5	- 0.07	0.69	**15**	0.13	0.71	**25**	- 0.22	0.51
6	- 0.18	0.62	**16**	0.03	0.81	**26**	- 0.21	0.52
7	- 0.19	0.58	**17**	0.06	0.72	**27**	- 0.24	0.50
8	- 0.16	0.72	**18**	0.02	0.78	**28**	- 0.22	0.54
9	- 0.17	0.69	**19**	0.04	0.74	**29**	- 0.23	0.50
10	- 0.15	0.71	**20**	0.15	0.77	**30**	- 0.21	0.54

Para la segunda variable igualmente se realiza la simulación de 30 individuos. Para los primeros 10 individuos se distribuye con la ecuación (7), para los siguientes 10 individuos se distribuye con la ecuación (8) y para los últimos 10 individuos se distribuye con la ecuación (9), así:

$$f(x) = log(x) \text{ con } 1 \leq x \leq 10 \tag{11}$$
$$f(x) = e^{\frac{x}{10}} \text{ con } 1 \leq x \leq 10 \tag{12}$$
$$f(x) = x^{\frac{1}{3}} \text{ con } 1 \leq x \leq 10 \tag{13}$$

El cuadro (3) indica la media y desviación para la simulación de cada individuo para la tercera variable. Se evidencia que la media de los 30 individuos es similares.

Cuadro 3: *Simulación variable 3.*

Indiv	**Media**	**Desv**	**Indiv**	**Media**	**Desv**	**Indiv**	**Media**	**Desv**
1	1.60	0.70	**11**	1.90	0.51	**21**	1.77	0.34
2	1.64	0.67	**12**	1.91	0.55	**22**	1.77	0.32
3	1.63	0.69	**13**	1.91	0.59	**23**	1.77	0.34
4	1.62	0.67	**14**	1.87	0.55	**24**	1.77	0.33
5	1.63	0.68	**15**	1.90	0.57	**25**	1.76	0.33
6	1.61	0.66	**16**	1.90	0.55	**26**	1.77	0.33
7	1.60	0.68	**17**	1.89	0.52	**27**	1.76	0.34
8	1.64	0.66	**18**	1.88	0.55	**28**	1.77	0.32
9	1.61	0.69	**19**	1.88	0.54	**29**	1.77	0.34
10	1.62	0.67	**20**	1.87	0.52	**30**	1.76	0.32

3.2. Suavizamiento y Estandarización

Con el objetivo de generar una función bajo los datos discretizados generados en el paso anterior, se realiza una aproximación bajo una curva $f(t)$ por medio de la combinación lineal de funciones a través de:

$$X_n(t) = \sum_{m=1}^{M} c_{nm} B_m(t), 1 \leq n \leq N \qquad (14)$$

Para suavizar las variables se utiliza una base B-Spline de grado 4, como los muestra las siguientes figuras

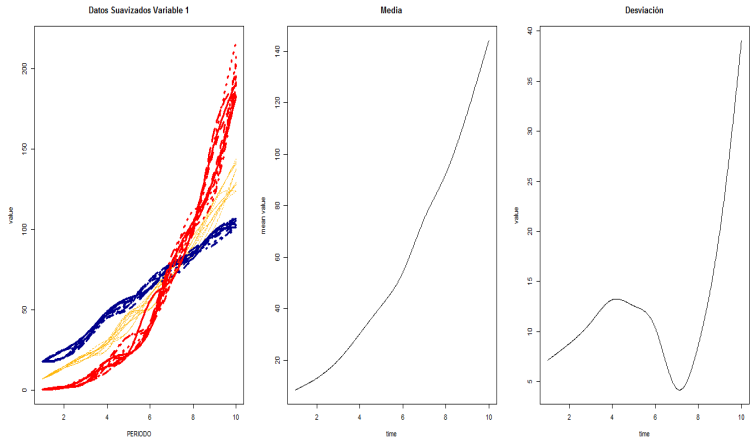

Imagen 3. Suavizamiento variable 1. Fuente: Elaboración Propia.

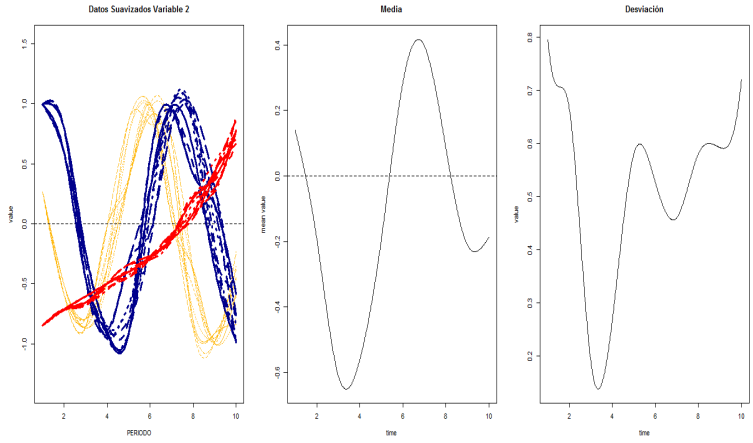

Imagen 4. Suavizamiento variable 2. Fuente: Elaboración Propia.

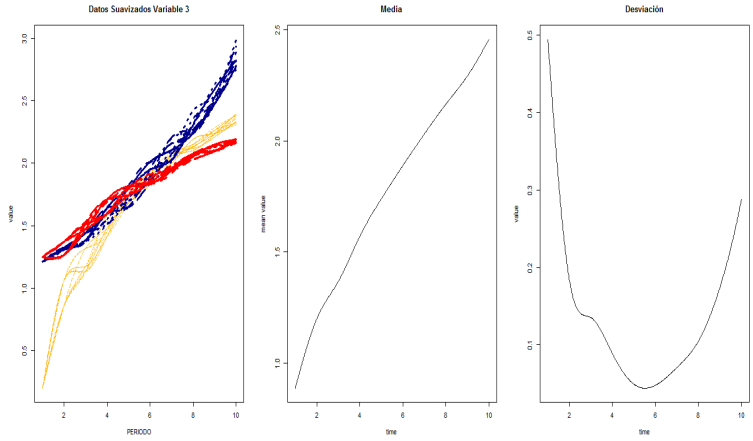

Imagen 5. Suavizamiento variable 3. Fuente: Elaboración Propia.

Para la estandarización de las variables suavizadas anteriormente, definimos la Media y Desviación Estándar que según Pineda (2017),

"Sea el conjunto de datos funcionales $f_1, f_2, ..., f_n$ definidos en $t \in [a, b]$. Las funciones descriptivas están dadas por las expresiones.(Ramsay, 2005)

- Media: $\hat{f(t)} = \frac{1}{n} \sum_{i=1}^{n} f_i(t)$ (15)
- Varianza: $s(t) = \frac{1}{n-1} \sum_{i=1}^{n} (f_i(t) - \hat{f(t)})^2$ (16)
- Desviación Estándar: $\sigma(t) = \sqrt{s(t)}$ (17)

Con el fin de poder estandarizar las funciones suavizadas, Ladino (2017) en el trabajo " Ajuste de un modelo funcional para medir el efecto de la tasa de cambio sobre el precio del Petróleo WTI en el periodo (2000-2015) " genera la función llamada exponencial.fd con la cual se puede lograr $\sigma(t)^{(} - 1)$ en datos funcionales.

Se estandariza las funciones suavizadas bajo la siguiente formula:

$$(f(t) - \hat{f(t)}) * \sigma(t)^{(-1)} \quad (18)$$

Quedando las 3 variables suavizadas y estandarizadas de media $f(t) \approx 0$ y desviación estándar $f(t) \approx 1$, así:

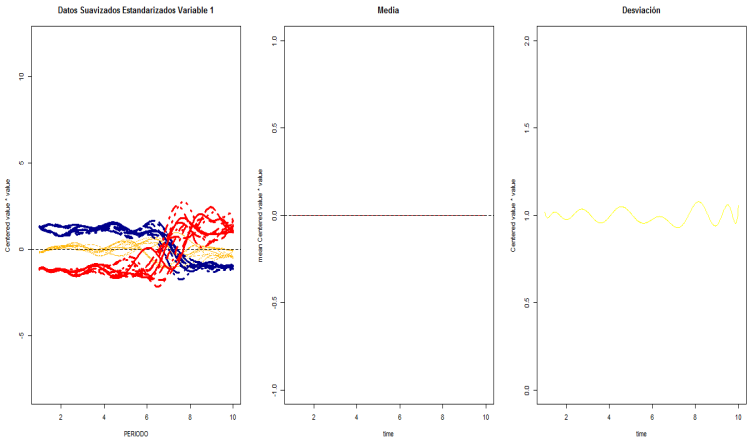

Imagen 6. Variable 1 Suavizada Estandarizada. Fuente: Elaboración Propia

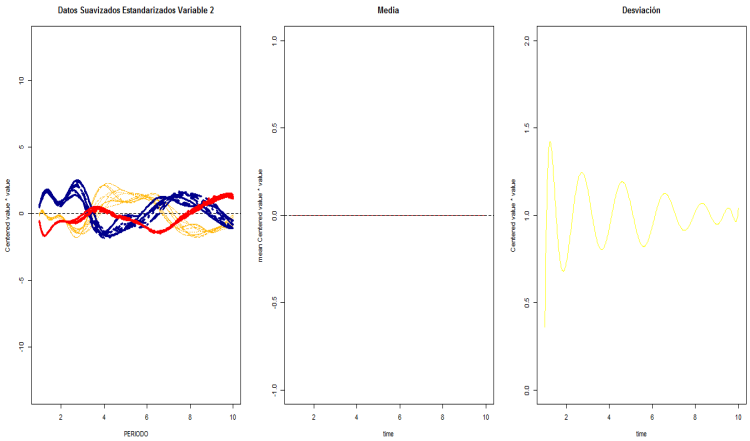

Imagen 7. Variable 2 Suavizada Estandarizada. Fuente: Elaboración Propia

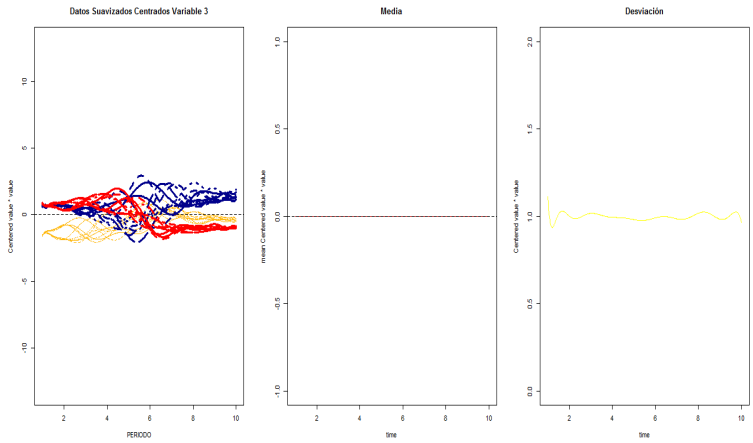

Imagen 8. Variable 3 Suavizada Estandarizada. Fuente: Elaboración Propia

3.3. Matriz de similitudes.

Para la construcción de la matriz de similitudes se define la distancia para datos funcionales como;

Sea $f(t) = \sum_{i=1}^{k} a_i\phi_i(t)$ y $g(t) = \sum_{i=1}^{k} b_i\phi_i(t)$ datos funcionales en $t \in [a, b]$, en un intervalo, entonces la distancia entre dos funciones $D^2(f, g)$, es igual a

$$D^2(f,g) = \int (f(t) - g(t))^2 dt$$

$$D^2(f,g) = \int (\sum (a_i - b_i) * \phi_i(t))^2 dt$$

$$D^2(f,g) = \sum_i (\sum_j (a_i - b_i) * (a_j - b_j) \int \phi_i(t) * \phi_j(t) dt$$

$$D^2(f,g) = \sum_i (a_i - b_i)^2 \qquad (19)$$

Para cada variable i se construye una matriz de similitud denotada como S_i con la ecuación (10), distancia entre datos funcionales. Por último definimos la matriz de similitud definitiva como la suma de las matrices de similitud de cada variable, así:

$$S = \sum_i S_i \qquad (20)$$

3.4. Segmentación de Cluster Jerárquicos, tabla de validación y criterio de Exactitud.

Apartir de la matriz de similitud S se aplica el algoritmo para métodos de aglomeración:

1. Comenzar con tantas clases como elementos, n. Las distancias entre clases son las distancias entre elementos originales.
2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.
3. Sustituir los dos elementos utilizados en 2. para definir la clase en 2, por un nuevo elemento que la represente.
4. Volver a 2 y repetir 2 y 3 hasta que tengamos todos los elementos queden agrupados en una clase única.

Criterios para definir distancias entre grupos: Sean A y B grupos con η_a y η_b elemento que se fusionan en un grupo (AB) con $\eta_a + \eta_b$ elementos.

Utilizamos el método de Ward para construir el dendograma,

$$W = \sum_g \sum_{i \in G} (X_i g - \hat{X}_g)'(X_i g - \hat{X}_g) \qquad (21)$$

El propósito es maximizar la homogeneidad dentro (mínima variabilidad dentro del conglomerado).

1. Suponer que cada dato forma un grupo de manera que $G = n$ y por tanto $W = 0$.
2. Unir los elementos que aumenten lo mínimo posible el valor de W.
3. Sustituir los dos elementos utilizados en 2 para definir la clase en 2., por un nuevo elemento que la represente.

Dado la construcción obtenemos el siguiente dendograma.

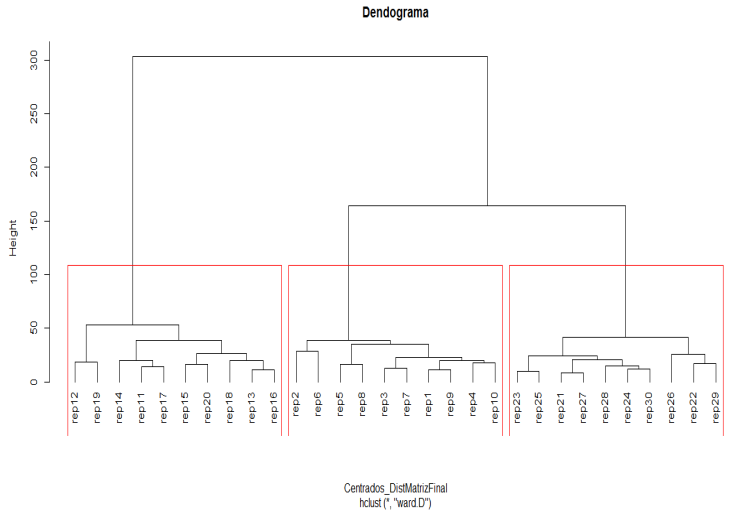

Imagen 9. Dendograma. Fuente: Elaboración Propia

Se evidencia claramente los tres grupos formados, se hace el corte en 3 grupos y se construye la tabla de validación cruzada de los grupos construidos a priori vs los cluster clasificados a partir del método, así:

Cuadro 4: *Matriz de validación*

	Cluster 1	**Cluster 2**	**Cluster 3**
Grupo_1	**10**	0	0
Grupo_2	0	**10**	0
Grupo_3	0	0	**10**

Se construye el criterio de exactitud ya que se requiere verificar el porcentaje que clasificó correctamente, como la sumatoria de la diagonal de la matriz de validación.

$$Exactitud = \frac{\sum a_i i}{n} = \frac{10+10+10}{30} = 100\,\% \qquad (22)$$

4. Aplicación Método de Cluster Jerárquicos para Datos Funcionales Multivariados en segmentación de clientes.

Cada vez se vuelve más importante las aplicaciones estadísticas en Marketing, ya que siempre se requiere mayor conocimientos de los clientes con el fin de poderle ofrecer las mejores opciones en productos y servicios. Una de estas aplicaciones es la segmentación de clientes ya que todos tienen un comportamiento diferente entre sí y por consiguiente sus necesidades.

4.1. Construcción de la base de datos.

Para muchas empresas y entre ellas del sector Retail es muy importante catalogar a los clientes en tres variables:

1. Recencia. Días transcurridos desde su última compra.
2. Compras Promedio Mensual. Número de compras promedio mensual en el último año.
3. Monto. Valor de las compras totales realizadas por el cliente en el último año.

La base para la aplicación se construyó a partir de las anteriores 3 variables para 1.000 clientes de un almacén de Retail durante un periodo de 12 meses de Enero a Diciembre de 2017.

4.2. Suavizamiento y Estandarización

Para los 1.000 individuos que se observaron en un periodo de 12 meses se realizó el suavizamiento con bases B-Spline, generando los siguientes comportamientos.

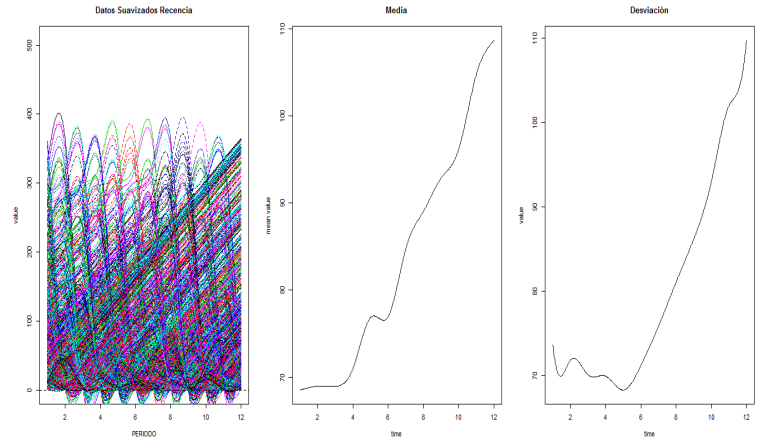

Imagen 10. Grafica Variable Recencia Suavizada. Fuente: Elaboración Propia

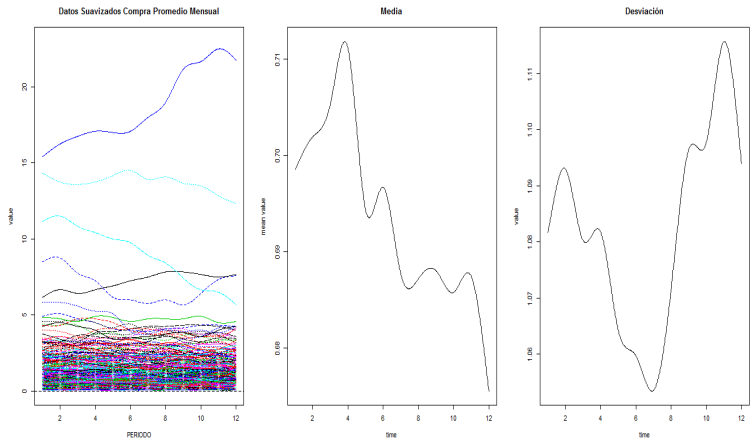

Imagen 11. Grafica Variable Compras Promedio Mensual Suavizada. Fuente: Elaboración Propia

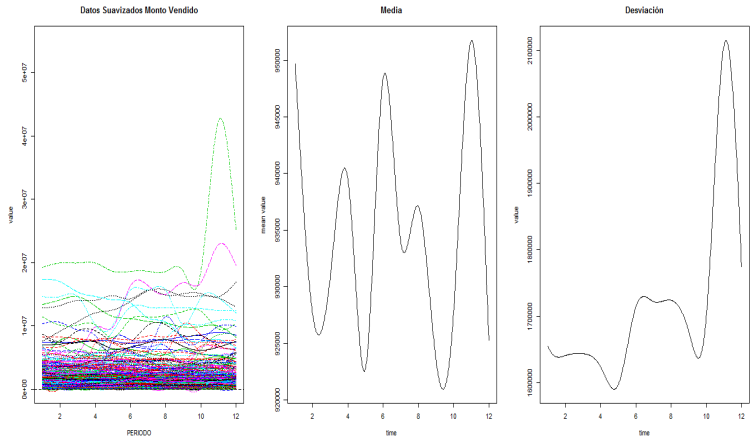

Imagen 12. Grafica Variable Monto Vendido Suavizada. Fuente: Elaboración Propia

Estandarizamos las tres variables funcionales generando variables funcionales de Media $s(t) \approx 0$, Desviación Estándar $\sigma(t) \approx 1$ y que en análisis del estudio no se vea afectada por las unidades de medida.

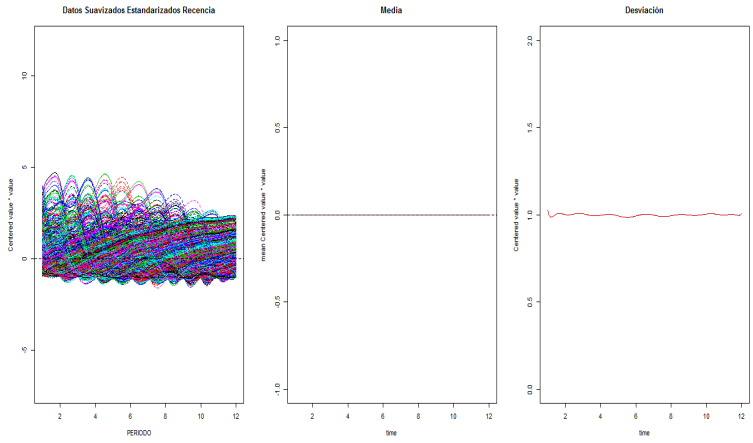

Imagen 13. Grafica Variable Recencia Suavizada Estandarizada. Fuente: Elaboración Propia

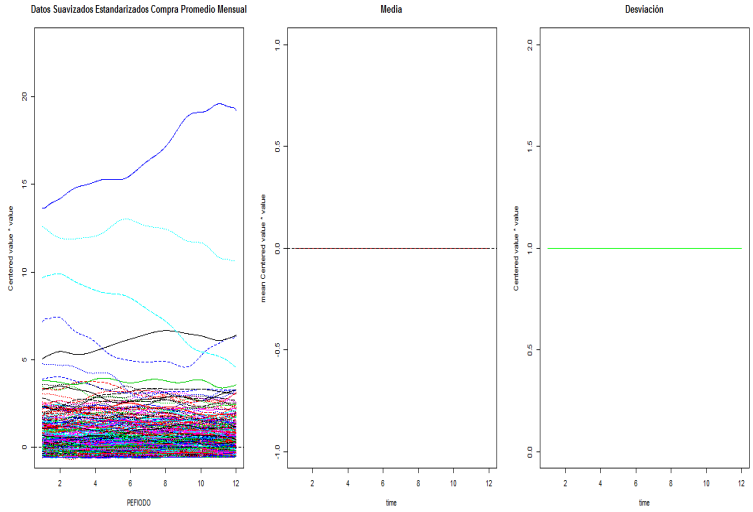

Imagen 14. Grafica Variable Compras Mensual Promedio Suavizada Estandarizada.
Fuente: Elaboración Propia

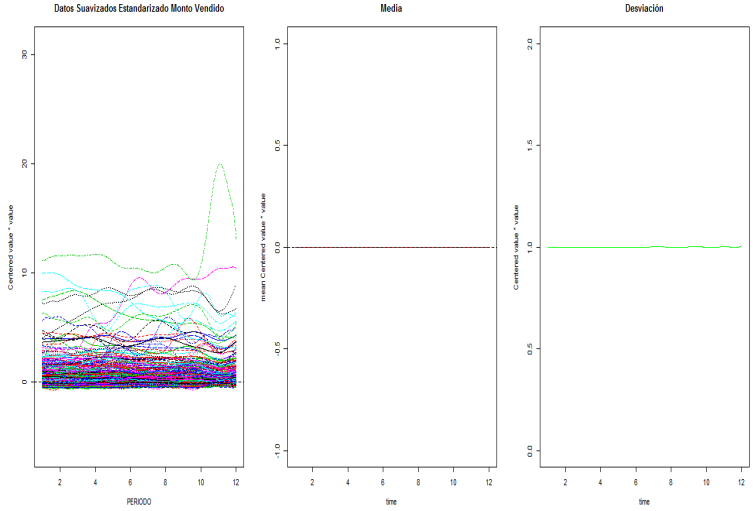

Imagen 15. Grafica Variable Monto Vendido Suavizada Estandarizada. Fuente: Elaboración Propia

4.3. Matriz de similitudes y dendograma.

Se construye la matriz de similitud y su respectivo dendograma. Se realiza el corte en 3 grupos ya que cada cluster sería una estrategia no es recomendable generar muchos cluster y además se distinguen claramente, aunque cabe aclarar que la escogencia del número de cluster depende del investigador.

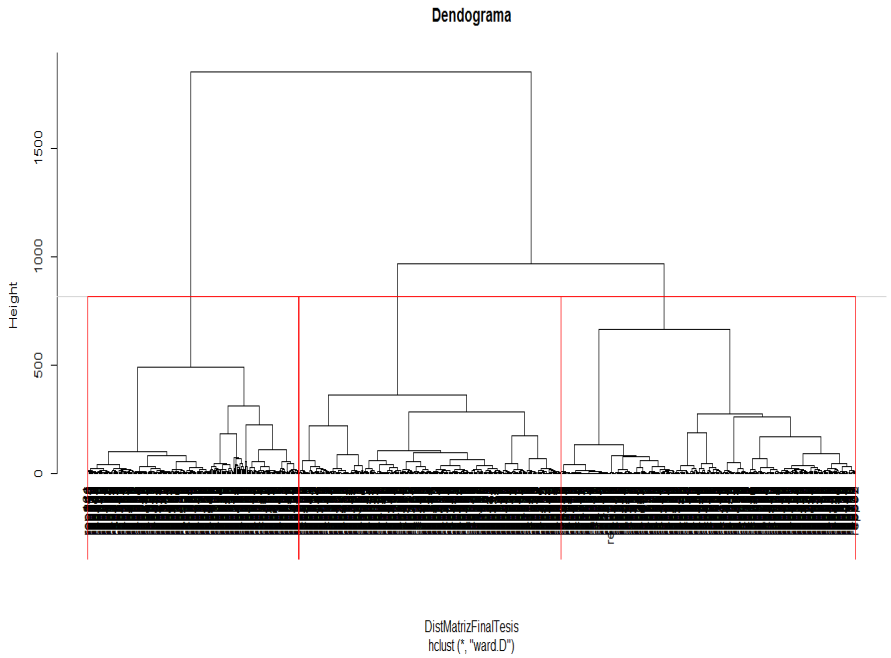

Imagen 16. Dendograma aplicación. Fuente: Elaboración Propia

Dado los cluster generados bajos el método, a continuación vemos el cluster 1 de color amarillo, el cluster 2 de color azul y el cluster 3 de color rojo. Donde a primera vista vemos que el cluster 1 son individuos de mayor Recencia y de menor promedio de compras mensual y montos vendidos. Del Cluster 2 encontramos que son clientes de Recencia media y de menor promedio de compras mensual y montos vendidos y por último el cluster 3 son clientes de menor Recencia y mayor promedio de compras mensual y montos vendidos, así:

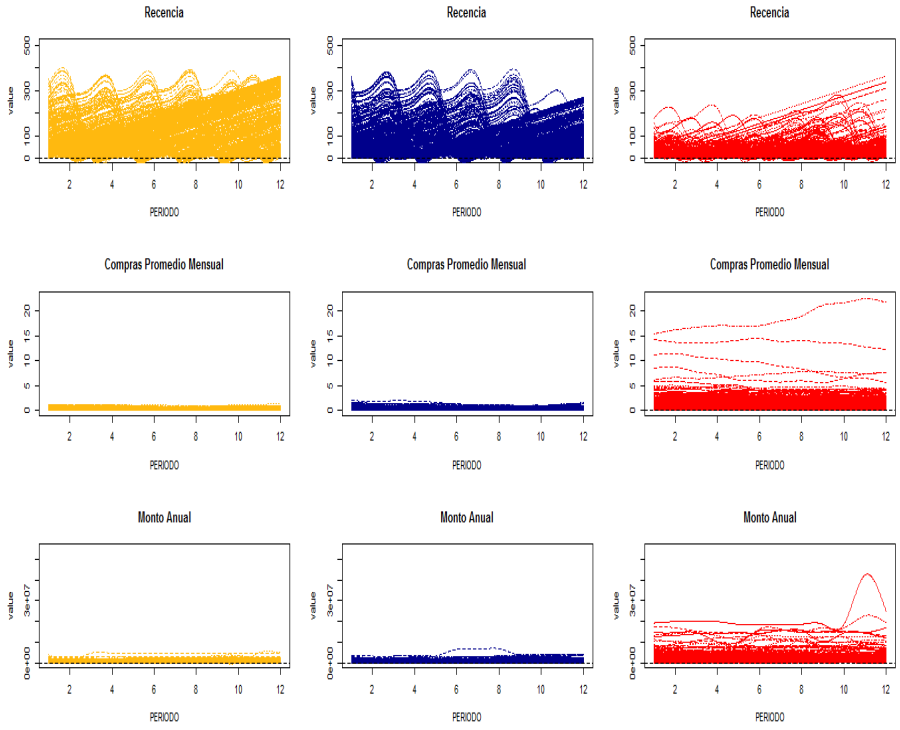

Imagen 17. Grafica de las 3 Variables por Cluster. Fuente: Elaboración Propia

Para tener mayor claridad se construye el vector de medias y desviación estándar por cluster, así:

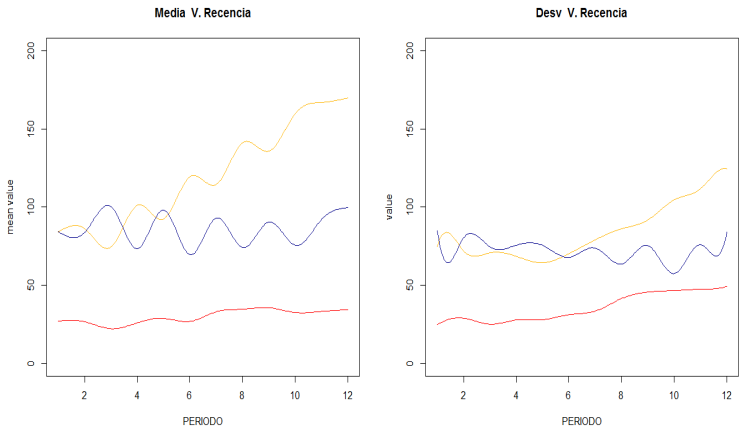

Imagen 18. Vector de Medias y Desviación Estándar Variable de Recencia. Fuente: Elaboración Propia

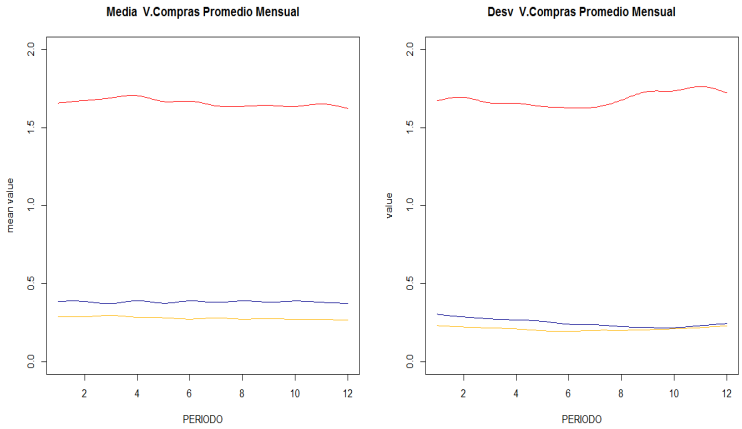

Imagen 19. Vector de Medias y Desviación Estándar Variable Compras Promedio Mensual. Fuente: Elaboración Propia

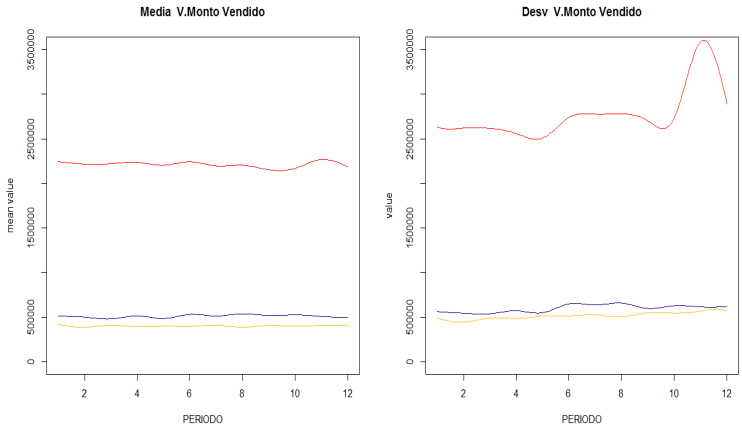

Imagen 20. Vector de Medias y Desviación Estándar Variable Monto Vendido. Fuente: Elaboración Propia

Caracterización de los cluster.

- **Cluster 1. Nombre: Clientes No Fidelizados.** Son Clientes que generalmente su Recencia es baja al comenzar el año y a medida que pasa los meses su Recencia incrementa es decir que no se generan visitas a la tienda. Además son clientes cuyo promedio de ventas en el último año son de 0.27 es decir 3 transacciones o tickets al año y por ultimo su Mon to vendido en el último año es de aproximadamente $400.000.
- **Cluster 2. Nombre: Clientes Recientes de Bajo Promedio de Ventas Mensuales y de Bajos Montos Vendidos.** Son Clientes que generalmente su Recencia es baja durante todo el año es decir que compran cada 3 meses. Además son clientes cuyo promedio de ventas en el último año son de 0.38 es decir 4.5 transacciones o tickets al año y por ultimo su Monto vendido en el último año es de aproximadamente $510.000.
- **Cluster 3. Nombre: Clientes Premium.** Son Clientes que generalmente su Recencia es baja cercana a los días es decir que nos visitan cada mes, en términos de Recencia su comportamiento es parecido al comienzo del año pero al terminar el año son muchos más dispersos. Además son clientes cuyo promedio de ventas mensual fue constante en el último año de 1.6 es decir 20 transacciones o tickets al año y por ultimo su Monto de Ventas en el último año es de aproximadamente $2.200.000, generalmente dispersos en el mes de Noviembre.

5. Discusión

1. El método de Cluster Jerárquico para datos funcionales multivariados evidencio que agrupa correctamente el comportamiento de las variables funcionales como se evidencio en el criterio de exactitud del 100/
2. La caracterización de los cluster o grupos bajo el método de Cluster Jerárquico para datos funcionales multivariados presentado en el actual trabajo aporta a que el investigador pueda generar características sobre los individuos respecto al tiempo o la variable de dominio.
3. Se evidencia que el método de Cluster Jerárquico para datos funcionales multivariados es facil de aplicabilidad y utilidad ya que se aporta el script del código en el software R.
4. La aplicación permite segmentar los clientes mediante sus tres variables del comportamiento de compra: Recencia, Promedio de Compra Mensual y Monto Vendido, para así mismo determinar estrategias claras y más productivas.

6. Bibliografía

[1] Berrendero, J. R., Justel, A., Svarc, M. (2011). Principal components for multivariate functional data. Computational Statistics Data Analysis. Recuperado de https://repositorio.uam.es/bitstream/handle/10486/13879 /65054$_{P}$*rincipal*$_{c}$*omponents*$_{f}$*or*$_{m}$*ultivariate*2011.*pdf*

[2]Julien Jacques, Cristian Preda (2014). Functional data clustering: a survey. Advances in Data Analysis and Classification, Springer Verlag. Recuperado de https://hal.inria.fr /file/index/docid/771030/filename/RR-8198.pdf

[3] Ladino, S. J. Pineda, R. W. (2017). Ajuste de un modelo funcional para medir el efecto de la tasa de cambio sobre el precio del Petróleo WTI en el periodo (2000-2015). Recuperado de http://repository.usta.edu.co/handle/11634/10376

[4] Peña, D. (2002). Análisis de datos multivariante. Mc Graw Hill.

[5] Pineda, R. W., Carrillo, R. A., Garatejo. E. O.. (2017). Análisis multivariado de datos funcionales aplicado a curvas de encefalogramas. Recuperado de http://revistas.usantotomas.edu.co/index.php/estadistica/article/viewFile/3200/3445

[6] Ramsay, J. O. y Silverman, B.W. (2002). Applied Functional Data Analysis. Segunda edición, Springer-Verlag.

[7] Ramsay, J. O. y Silverman, B.W. (2005). Functional Data Analysis. Segunda edición, Springer-Verlag.

[8] Ramsay, J. O. y Silverman, B.W. (2009). Functional Data Analysis with R and Matlab. Segunda edición, Springer-Verlag.

[9] Santiago E, Chincangana F, Olaya J. (2017). Análisis de conglomerados de datos funcionales que representan el comportamiento diario de partículas en el aire en una zona urbana de Cali en 2015. XXVII Simposio Internacional de Estadística – 5th International Workshop on Applied Statistics. Recuperado de http://simposioestadistica.unal.edu.co /fileadmin/content/eventos/simposioestadistica/documentos/memorias/Memorias$_{2}$017 /*Comunicaciones*/*Estadistica*$_{A}$*mbiental*/4.$_{C}$*omunicacion*$_{1}$25.*pdf*

[10]Yamamoto, M. (2012). Clustering of functional data in a low-dimensional subspace. Advances in Data Analysis and Classification. Recuperado de https://link.springer.com/article/10.1007/s11634-012-0113-3

7. Anexo: Código fuente en el programa R.

```
############################################################
#******Simulación Variable 1, Familia Polinomica*********#
############################################################

f1=matrix(,ncol=10, nrow=10)
for(i in 1:10) {
r=matrix(runif(i, min=i+1, max=i+2)^(2),ncol = 10)
f1[i,]=(r)
}
f1

f2=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
r=matrix(10*runif(i, min=i, max=i+1),ncol = 10)
f2[i,]=(r)
}
f2

f3=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
r=matrix((runif(i, min=i, max=i+1)^3)/6,ncol = 10)
f3[i,]=(r)
}
f3

  #*********************************************
  #Tabla Variable 1
  #*********************************************

  f_final=as.matrix(cbind(f1,f2,f3), nrow=30, 10)
  fmediai=apply(f_final, 2,mean)
  fdesvi=apply(f_final, 2,sd)
  resumenf=cbind(fmediai,fdesvi)
  resumenf

  #*********************************************
  #Grafico de Puntos
  #*********************************************

  windows()
  plot(Ffinal_Bs, main="Datos Suavizados Variable 1")
  plot(1:10, f_final[,1], col=j, main="Datos discretizados VARIABLE 1",pch=20,
       ylab="VARIABLE1", xlab="PERIODO",ylim = c(0,max(f_final)))
  for (j in 1:ncol(f_final))
    points(1:10, f_final[,j], col=j,pch=20)

  #*************************************************************************
```

```
#Crea las bases de Funciones usando Series de Fourier y B-splines
#*************************************************************************

BSpl=create.bspline.basis(norder=4, breaks=seq(1:10))

#*****************************************************
#Suavisar datos, se genera la Media y la Desviación
#*****************************************************

Ffinal_Bs<-Data2fd(1:10,f_final, basisobj=BSpl)
FMedia=mean.fd(Ffinal_Bs)
FDesv=sd.fd(Ffinal_Bs)

windows()
op<-par(mfrow = c(1, 3))
plot(Ffinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="Datos Suavi
lines(Ffinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Ffinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(FMedia, main="Media")
plot(FDesv, main="Desviación")

  #*********************************************
#Estandarización de datos
#*********************************************

Centrados_Ffinal_Bs=times.fd(center.fd(Ffinal_Bs),exponencial.fd(FDesv,-1))
CenFMedia=mean.fd(Centrados_Ffinal_Bs)
CenFDesv=sd.fd(Centrados_Ffinal_Bs)

windows()
op<-par(mfrow = c(1, 3))
plot(Centrados_Ffinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="D
lines(Centrados_Ffinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Centrados_Ffinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(CenFMedia, main="Media",col = "red3", ylim = c(-1,1))
plot(CenFDesv, main="Desviación",col = "yellow", ylim = c(0,2))

#########################################################
#*******Simulación Variable 2 Trigonometricas**********#
#########################################################

g1=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
  r=matrix(cos(runif(i, min=i, max=i+1)),ncol = 10)
  g1[i,]=(r)
}
g1

g2=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
```

```
  r=matrix(sin(runif(i, min=i, max=i+1)),ncol = 10)
  g2[i,]=(r)
}
g2

g3=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
  r=matrix(tan(runif(i, min=i, max=i+1)/10)-1,ncol = 10)
  g3[i,]=(r)
}
g3

#*********************************************
#Tabla Variable 2
#*********************************************

g_final=as.matrix(cbind(g1,g2,g3), nrow=30, 10)
gmediai=apply(g_final, 2,mean)
gdesvi=apply(g_final, 2,sd)
resumeng=cbind(gmediai,gdesvi)
resumeng

#*********************************************
#Grafico de Puntos
#*********************************************

windows()
plot(1:10, g_final[,11], col=1, main="Datos discretizados VARIABLE 2",pch=20,
     ylab="VARIABLE2", xlab="PERIODO",ylim = c(0,max(g_final)))
for (j in 1:ncol(g_final))
  points(1:10, g_final[,j], col=j,pch=20)


#*****************************************************************************
#Crea las bases de Funciones usando Series de Fourier y B-splines
#*****************************************************************************

BSpl=create.bspline.basis(norder=4, breaks=seq(1:10))

#*****************************************************
#Suavisar datos, se genera la Media y la Desviación
#*****************************************************

Gfinal_Bs<-Data2fd(1:10,g_final, basisobj=BSpl)
windows()
plot(Gfinal_Bs, main="Datos Suavizados Variable 2")
GMedia=mean.fd(Gfinal_Bs)
GDesv=sd.fd(Gfinal_Bs)

windows()
```

```
op<-par(mfrow = c(1, 3))
plot(Gfinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="Datos Suavi
lines(Gfinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Gfinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(GMedia, main="Media")
plot(GDesv, main="Desviación")


#*********************************************
#Estandarización de datos
#*********************************************

Centrados_Gfinal_Bs=times.fd(center.fd(Gfinal_Bs),exponencial.fd(GDesv,-1))
CenGMedia=mean.fd(Centrados_Gfinal_Bs)
CenGDesv=sd.fd(Centrados_Gfinal_Bs)

windows()
op<-par(mfrow = c(1, 3))
plot(Centrados_Gfinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="D
lines(Centrados_Gfinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Centrados_Gfinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(CenGMedia, main="Media",col = "red3", ylim = c(-1,1))
plot(CenGDesv, main="Desviación",col = "yellow", ylim = c(0,2))

############################################################
#*******Simulación Variable 3 Exp,Log y Raiz***********#
############################################################

h1=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
  r=matrix(log(runif(i, min=i, max=i+1)),ncol = 10)
  h1[i,]=(r)
}
h1

h2=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
  r=matrix(exp(runif(i, min=i, max=i+1)/10),ncol = 10)
  h2[i,]=(r)
}
h2

h3=matrix(,ncol=10, nrow=10)
for (i in 1:10) {
  r=matrix((runif(i, min=i, max=i+1)^(1/3)),ncol = 10)
  h3[i,]=(r)
}
h3

#*********************************************
```

```
#Tabla Variable 3
#*******************************************

h_final=as.matrix(cbind(h1,h2,h3), nrow=30, 10)
hmediai=apply(h_final, 2,mean)
hdesvi=apply(h_final, 2,sd)
resumenh=cbind(hmediai,hdesvi)
resumenh

#*******************************************
#Grafico de Puntos
#*******************************************

windows()
plot(1:10, h_final[,1], col=1, main="Datos discretizados VARIABLE 3",pch=20,
     ylab="VARIABLE3", xlab="PERIODO",ylim = c(0,max(h_final)))
for (j in 1:ncol(h_final))
  points(1:10, h_final[,j], col=j,pch=20)

#*************************************************************************
#Crea las bases de Funciones usando Series de Fourier y B-splines
#*************************************************************************

BSpl=create.bspline.basis(norder=4, breaks=seq(1:10))

Hfinal_Bs<-Data2fd(1:10,h_final, basisobj=BSpl)
windows()
plot(Hfinal_Bs, main="Datos Suavizados Variable 3")
HMedia=mean.fd(Hfinal_Bs)
HDesv=sd.fd(Hfinal_Bs)

windows()
op<-par(mfrow = c(1, 3))
plot(Hfinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="Datos Suavi
lines(Hfinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Hfinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(HMedia, main="Media")
plot(HDesv, main="Desviación")

#*******************************************
#Estandarización de datos
#*******************************************

Centrados_Hfinal_Bs=times.fd(center.fd(Hfinal_Bs),exponencial.fd(HDesv,-1))
CenHMedia=mean.fd(Centrados_Hfinal_Bs)
CenHDesv=sd.fd(Centrados_Hfinal_Bs)
windows()
op<-par(mfrow = c(1, 3))
plot(Centrados_Hfinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="D
lines(Centrados_Hfinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
```

```
lines(Centrados_Hfinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(CenHMedia, main="Media",col = "red3", ylim = c(-1,1))
plot(CenHDesv, main="Desviación",col = "yellow", ylim = c(0,2))

#*******************************************
#***************CLUSTER*********************#
#*******************************************

###Graficos 3 Variables

windows()
op<-par(mfrow = c(1, 3))
plot(Ffinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="Datos Suavi
lines(Ffinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Ffinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(Gfinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="Datos Suavi
lines(Gfinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Gfinal_Bs[seq(21,30)],col = "red1",lwd=3)
plot(Hfinal_Bs[seq(1,10)],col = "darkgoldenrod1", xlab="PERIODO", main="Datos Suavi
lines(Hfinal_Bs[seq(11,20)],col = "darkblue",lwd=3)
lines(Hfinal_Bs[seq(21,30)],col = "red1",lwd=3)

#*******************************************
#*****Matriz de Distancias o Similitudes*****#
#*******************************************

Centrados_MatrizF<-Centrados_Ffinal_Bs$coefs
dim(Centrados_MatrizF)
Centrados_DistMatrizF<-dist(t(Centrados_MatrizF),diag=T,upper=T)

Centrados_MatrizH<-Centrados_Hfinal_Bs$coefs
dim(Centrados_MatrizH)
Centrados_DistMatrizH<-dist(t(Centrados_MatrizH),diag=T,upper=T)

Centrados_MatrizG<-Centrados_Gfinal_Bs$coefs
dim(Centrados_MatrizG)
Centrados_DistMatrizG<-dist(t(Centrados_MatrizG),diag=T,upper=T)

Centrados_DistMatrizFinal<-Centrados_DistMatrizH+Centrados_DistMatrizG+Centrados_Di

#*******************************************
#****Cluster Jerarquicos (Metodo de Ward)*****#
#*******************************************

cj=hclust(Centrados_DistMatrizFinal, method = "ward.D")
cj$merge
windows()
plot(cj,main="Dendograma",labels=row.names(Centrados_DistMatrizFinal),hang=-1)
abline(h=1000000,col="lightgrey")
rect.hclust(cj, k=3, border="red")
```

```
ColumnaCluster<-c((cutree(cj, k=3)))

#*****************************
#****Tabla de Validación*****#
#*****************************

Elemento1<-as.data.frame(rep(c("Grupo_1"),10))
Elemento2<-as.data.frame(rep(c("Grupo_2"),10))
Elemento3<-as.data.frame(rep(c("Grupo_3"),10))

MatrixInicial<-sqldf("Select *
                     From Elemento1
                     Union all
                     Select *
                     From Elemento2
                     Union all
                     Select *
                     From Elemento3
                     ")

str(MatrixInicial)

Centrados_MatrixCluster<-cbind(MatrixInicial,ColumnaCluster)
names(Centrados_MatrixCluster)=c("Grupo","Cluster")

TablaConfusionCentrada<-table(Centrados_MatrixCluster$Grupo,
                      Centrados_MatrixCluster$Cluster)

TablaConfusionCentrada
```