Cuadernos del CIMBAGE Nº8 (2006) 97-128

ANÁLISIS DE COMPORTAMIENTO DE CUENTAS CORRIENTES EN ENTIDADES BANCARIAS MEDIANTE EL USO DE FUZZY CLUSTERING Y ANÁLISIS DISCRIMINANTE PARA LA ADMINISTRACIÓN DE RIESGO CREDITICIO[1]

María T. Casparri*, Federico A. Alcalde Bessia**, Julio Fabris***
Centro de Métodos Cuantitativos Aplicados a la Economía y la Gestión
Facultad de Ciencias Económicas - Universidad de Buenos Aires
Av. Córdoba 2122 - Ciudad de Buenos Aires – C1120AQ - Argentina
*casparri@econ.uba.ar,**falcalde@econ.uba.ar,***jfabris88@hotmail.com

Recibido 12 de febrero de 2006, aceptado 18 de abril 2006

Resumen

Cada entidad bancaria tiene sus propios parámetros de evaluación de clientes y aplica sus propios métodos para hacerlo. Esto forma parte de su política de administración de riesgos. El análisis del comportamiento de la cuenta corriente de cada cliente es de suma importancia en este caso, ya que describe la conducta del cliente en relación a sus deudas y ayuda a evaluar los riesgos que el banco asume. A su vez, este análisis permite la descripción de la evolución de los riesgos mediante el hallazgo de un patrón de conducta de cada cliente. Una vez descripta esta evolución, la entidad podrá definir su política crediticia de corto plazo en cuenta corriente, pudiendo hacer un seguimiento de las cuentas que entran en zonas que el banco evaluaría como indeseables.

En el presente trabajo se desarrolla, mediante un modelo simple con datos generados en forma aleatoria, una aplicación referida a estos métodos de evaluación. Se presenta, fundamentalmente, el método de *Fuzzy Clustering*, utilizando los programas SPSS® y R para desarrollar los cálculos. Además, se hace un análisis discriminante canónico para la asignación de nuevos individuos a los grupos definidos y la reasignación en caso de cambio de las características.

Palabras Clave: Fuzzy clustering, validación, Clustering, análisis discriminante.

[1] Presentado en XII Congreso Internacional de la Sociedad de Gestión y Economía Fuzzy (SIGEF). 26-28 de Octubre 2005, Bahía Blanca, Argentina.

BEHAVIOURAL ANALYSIS OF CURRENT ACCOUNTS IN BANKING INSTITUTIONS THROUGH THE USE OF FUZZY CLUSTERING AND DISCRIMINANT ANALYSIS FOR CREDIT RISK MANAGEMENT[2]

María T. Casparri*, Federico A. Alcalde Bessia**, Julio Fabris***
Centro de Métodos Cuantitativos Aplicados a la Economía y la Gestión
Facultad de Ciencias Económicas - Universidad de Buenos Aires
Av. Córdoba 2122 - Ciudad de Buenos Aires – C1120AQ - Argentina
*casparri@econ.uba.ar,**falcalde@econ.uba.ar,***jfabris88@hotmail.com

Received 12 February 2006, accepted 18 April 2006

Abstract

Each bank has its own parameters to evaluate its clients and applies its own methods to do that. This belongs to its management risk policy. The behavioural analysis of the current accounts of each client is of great importance in this case, since it describes the client's behaviour related to his debts and helps to evaluate the risks that the bank deals with. In the same way, this analysis permit to describe the risk's evolution through the gathering of behavioural patterns. Once one has described this evolution, the institution can define its short term credit policy in current account, being able to follow the ones that enter the zones which the bank has defined as unwishable.

Over the present paper, an application referred to those methods of evaluation is developed with a simple model using simulated data. Fundamentally, the method of Fuzzy Clustering is shown using SPSS and R software to make the calculations. Additionally, we make a canonical discriminant analysis to assign new individuals to the defined groups and the reorganization in the case of characteristic's changes.

Keywords: Fuzzy clustering, Validation, Clustering, Discriminant analysis, Risk management, Fuzzy theory.

[2] Presented in XII Congreso Internacional de la Sociedad de Gestión y Economía Fuzzy (SIGEF). 26-28 October 2005, Bahía Blanca, Argentina.

1. ANÁLISIS DE COMPORTAMIENTO DE CUENTAS CORRIENTES

1.1. Introducción al problema

Las entidades bancarias asumen riesgos de corto plazo con sus clientes debido al otorgamiento de montos máximos de giro en descubierto en cuenta corriente. Estas cuentas tienen la particularidad de que el cliente puede realizar extracciones en descubierto hasta un monto predefinido por la entidad. Una vez alcanzado ese monto, el cliente ya no dispone de crédito y debe saldar su deuda para poder continuar operando.

Las ganancias para el banco respecto de ese tipo de cuenta surgen de la cantidad de días que el cliente permanece con saldo deudor y de las tasas diarias vigentes. Una vez que el cliente alcanza el máximo asignado, el banco ya no le otorga crédito. En general, las entidades bancarias están interesadas en mantener carteras con movimientos dinámicos y que no alcancen el monto máximo. Esto es así, dado que si el cliente lo hubiese alcanzado por problemas de insolvencia, el banco lo perdería y aparecerían los problemas de cobro con sus consecuentes gastos asociados.

Cada entidad bancaria tiene sus propios parámetros de evaluación de clientes y aplica sus propias reglas para clasificarlos como parte de su política de administración de riesgos. El análisis del comportamiento de la cuenta corriente de cada cliente es de suma importancia para esa política, ya que describe la conducta del cliente del banco en relación a sus deudas y ayuda a evaluar los riesgos que éste asume. A su vez, este análisis permite la descripción de la evolución de los riesgos mediante el hallazgo de un patrón de conducta de cada cliente. Una vez descripta esta evolución, la entidad podrá definir su política crediticia de corto plazo en cuenta corriente, pudiendo hacer un seguimiento

generalizado de las cuentas que entran en zonas que el banco evaluaría como indeseables.

El trabajo se enfocará en el análisis de un segmento específico de clientes de una entidad bancaria supuesta que llamaremos A1 donde ubicaremos a todos los clientes que cumplen con:

- Límite máximo de descubierto en cuenta corriente de $1.000
- La antigüedad de las deudas es menor a 120 días.

Dentro de esta categoría intentaremos reconocer grupos o patrones de comportamiento de los clientes y asignaremos cada cliente a cada grupo hallado.

Para llevar a cabo este análisis se debe poder reconocer cuáles son las variables destacadas que ayudarán a comprender mejor el perfil de movimientos de la cuenta y su relación con las características del cliente. No es lo mismo para una entidad bancaria un cliente con alto endeudamiento pero de larga antigüedad en el banco y de gran importancia institucional que un cliente de alto endeudamiento pero de incorporación reciente y sin importancia estratégica para el banco. Llegado a este punto, se debe pensar cómo se llevará a cabo la selección de las variables relevantes para la descripción suficiente de cada individuo. Esto forma parte del análisis previo de los datos disponibles que podría realizarse mediante análisis estadístico o, simplemente, utilizar el criterio del investigador. En nuestro caso, supondremos realizado un análisis estadístico que ha seleccionado las siguientes variables:

- Monto total de deuda acumulada.
- Días promedio de retraso antes de pago de deuda.

- Monto actual de la deuda.
- Días de retraso de la deuda actual.
- Años de pertenencia al banco.
- Índice de la entidad.

El índice de la entidad se refiere a la consideración que tiene el banco hacia el cliente, resumida en un número que representa su importancia relativa. Esto puede deberse a una estrategia de posicionamiento del banco en determinados sectores u otras causas.
Supondremos, además, que los clientes no están atentos a las variaciones de las tasas en el tiempo (por ser estas constantes, por ejemplo) lo cual nos autorizará a hacer una comparación estática de los niveles de endeudamiento absoluto de cada uno. Tampoco nos interesará la frecuencia de los movimientos porque aceptaremos que a la entidad sólo le interesa la cancelación de las deudas. Esto justifica hacer un análisis transversal, es decir no se intentará hacer un análisis en el tiempo de cada cuenta, sino que se estudiará el estado de todas las cuentas en un momento dado considerando que dicho estado es representativo de la dinámica.
Por otra parte, se han elegido algunas variables discretas y otras continuas, dependiendo de la característica que describen. Por ejemplo, "monto actual de la deuda" es continua y "días promedio de retraso" es discreta. Si bien la variable "días promedio de retraso" asume valores enteros del intervalo $[0,120]$, se considera continua con la finalidad de simplificar su tratamiento, aunque el método tiene la posibilidad del tratamiento de variables discretas[3].

[3] Respecto a esto, además de poder considerar variables discretas, el método puede aplicarse a variables *categóricas*. Éstas son variables que pueden asumir un número finito de valores, vg. la variable *género* es masculino ó femenino.

1.2. Análisis previo y desarrollo

El comportamiento de cada cuenta individual será descripto por comparación. El primer problema que se presenta es la gran cantidad de datos, en nuestro ejemplo son 1404 individuos. Para ello, aplicaremos el método de *Fuzzy Clustering*, el cual se describe en el Anexo I. Para implementar el método se deben previamente definir: la cantidad de conjuntos en los que se quiere agrupar los datos, la función que comparará dato por dato (función de distancia o similitud) y un criterio de detención de la iteración. El resultado que se obtiene depende fuertemente de la distribución de los datos, del análisis previo que lleva a la definición de la cantidad de grupos y de la definición de la función de distancia.

Respecto de la cantidad de grupos a considerar, en nuestro caso, realizaremos el análisis para 2, 3, 4 y 5 grupos y evaluaremos las medidas de validación para determinar la cantidad de grupos más adecuada. También haremos un reconocimiento gráfico de la cantidad de grupos posibles. Con los métodos más avanzados de *clustering* puede optarse, para la definición de grupos, por el Clustering Jerárquico o métodos que hagan uso de criterios de información cuando se hacen enfoques probabilísticos. Para una descripción de estos métodos puede verse Fraley y Raftery (1998) y Fraley y Raftery (2002). Las respuestas de estos métodos no siempre son satisfactorias dado que, a pesar de que son sistemáticos, la elección de la cantidad de grupos termina siendo una decisión puramente subjetiva.

1.2.1. Análisis gráfico

Para comenzar realizamos un análisis informal de los siguientes gráficos con el objetivo de reconocer los grupos. Se puede apreciar la distribución de los datos de las variables tomadas de dos en dos

(*scatter plot*). Es de esperar que en los siguientes gráficos no haya relaciones visibles entre las variables que, en teoría, no deben estar asociadas:

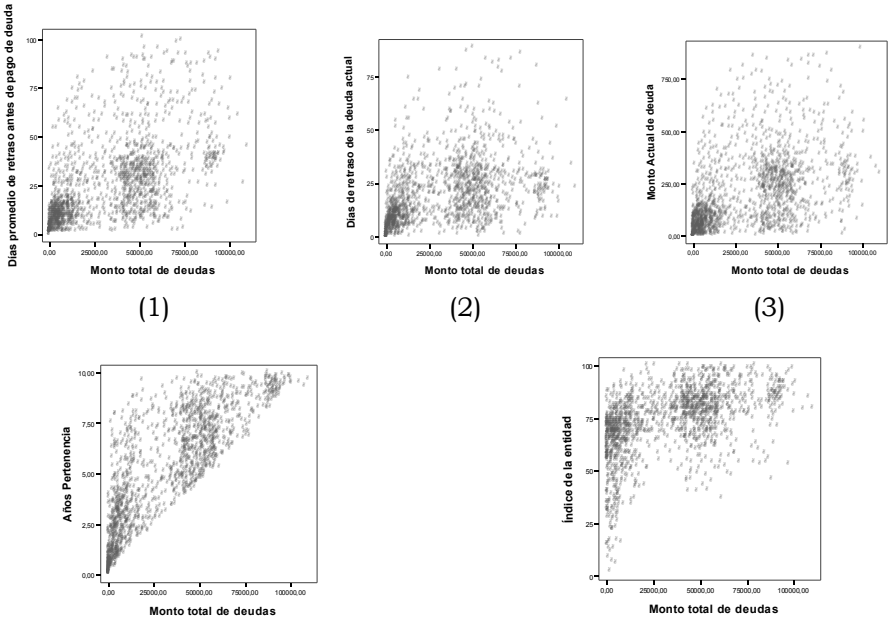


Figura 1. Monto total de deudas contra las demás variables

Aquí pueden reconocerse tres grupos asociados a la variable "Monto total de deudas". Seleccionar tres grupos a esta altura no implicará necesariamente que el gráfico será dividido en tres partes y luego se agruparan los datos de esa forma, sino que se pueden reconocer tres comportamientos distintos respecto de la variable mencionada y las demás. Por ello, habrá que realizar el mismo análisis con las demás variables.

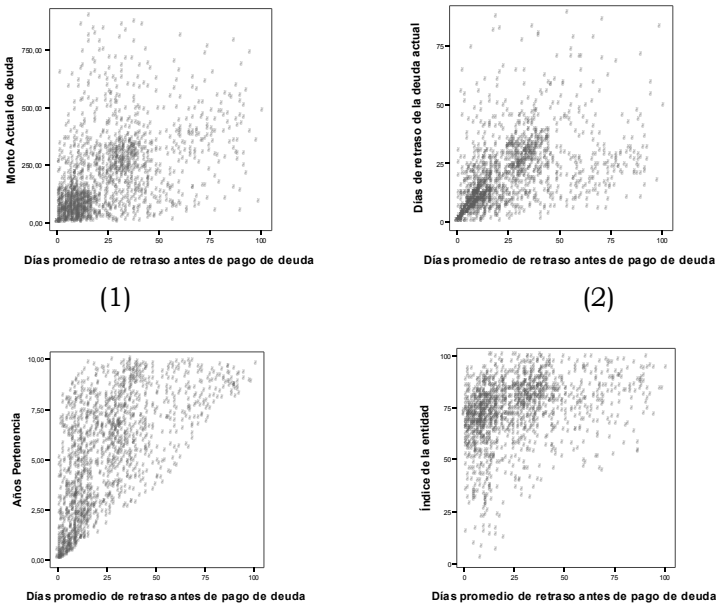


Figura 2. "Días promedio de retraso antes del pago" contra las demás variables

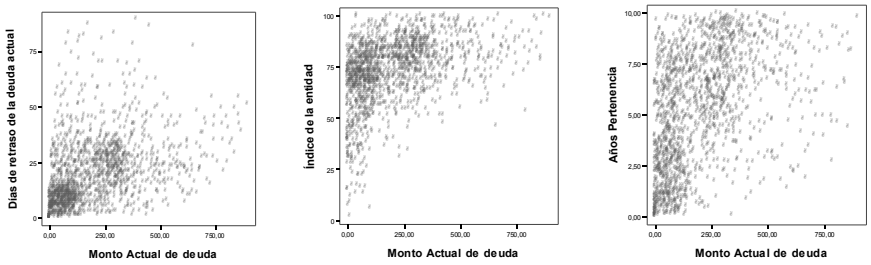
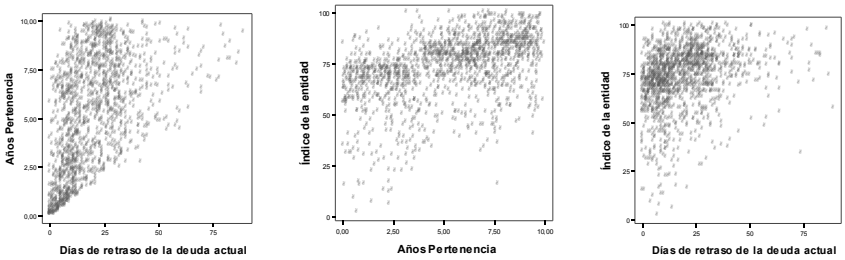


Figura 3. "Monto actual de deuda" contra las demás variables



(1) (2) (3)

Figura 4. (1),(2)"Días de retraso de la deuda actual" contra las demás variables y (3)"Años de pertenencia" contra "Índice de la entidad"

Por ejemplo, en el gráfico 2 de la figura 4, los "Días de retraso de la deuda actual" parecen no formar grupos con "Índice de la entidad" como es de esperarse dada la hipótesis sobre el significado del índice. Vemos además que, cuando las variables pueden ser relacionadas en teoría, las observaciones verifican esa suposición.

En todos los gráficos pueden verse al menos dos zonas de alta densidad de puntos y luego una nube alrededor de ellas hacia el extremo derecho. Por lo anterior, en principio, seleccionaremos 3 grupos.

En cuanto al análisis formal, veremos qué sucede con los índices si se seleccionan 2, 4 ó 5 grupos.

Para este primer análisis elegiremos m = 2 (coeficiente de borrosidad igual a dos), la distancia euclidea y un criterio de detención, TOL, de 1.10^{-16} de cambio en la función objetivo.

Definidos estos parámetros de entrada se procede a realizar el *Fuzzy Clustering* mediante la función *cmeans* en el software R. Los resultados para 2, 3, 4 y 5 grupos se evalúan mediante los índices de validación (ver Anexo I para una descripción de los mismos):

Cantidad de Grupos	Coeficiente de Partición	Coeficiente de Entropía	Índice de Xie-Beni	Índice de Fukuyama-Sugeno
2	0,891152 *	0,209313 *	0,059358	-5,37232E+11
3	0,866102	0,257038	0,042818 *	-7,68962E+11
4	0,82407	0,3443192	0,102443	-8,11228E+11 *
5	0,790254	0,409584	12,372499	2,0114E+12

Cuadro 1. Índices de Validación

Se ha indicado en el cuadro con un asterisco la cantidad de grupos seleccionados por cada índice. Vemos que los índices se contradicen. En el caso del Xie-Beni se favorece la elección de 3 grupos. Para el Fukuyama-Sugeno se considera mejor la elección de 4 grupos. Mientras que en el caso del coeficiente de partición, se favorece la elección de 2 grupos y, por construcción, el coeficiente de entropía coincide. Se verifica en este caso que los índices no nos ayudan a seleccionar la cantidad de grupos. Tomar en cuenta sus resultados al momento de tomar una decisión dependerá de la interpretación que se haga de la construcción del índice y cómo utiliza las variables de entrada.

El siguiente problema a enfrentar es la selección definitiva del coeficiente de borrosidad (m). Asignaremos el valor dos para este coeficiente como sugieren Pal y Bezdek (1995). Cuanto más cercano a uno sea m, más fuerte será la asignación de la pertenencia a cada grupo por lo que la matriz de pertenencias se parecerá mucho más a la matriz asociada al método de *Clustering* estándar. Inversamente, cuanto más alto sea m, más suave será la asignación a cada grupo.

Veamos cómo varían los centros y la pertenencia de cada cliente al variar m numéricamente. Por ejemplo, tomemos, de los datos con los que contamos, el cliente 1098.

Clientes	Monto total de deudas	Días promedio de retraso antes de pago de deuda	Monto actual de deuda	Días de retraso de la deuda actual	Años Pertenencia	Índice de la entidad
cliente1098	63733,82	42	129,70	38	9,32	85

Cuadro 2. Detalle de los datos del Cliente1098

Se ha aplicado el *Fuzzy Clustering* para 3 grupos con un criterio de detención de 1.10^{-16} de cambio en la función objetivo utilizando la función *cmeans* en el software R, obteniendo los siguientes resultados. Las columnas mantienen el orden del presentado en el cuadro 2:

Centros con m = 1,05 luego de 26 iteraciones

CENTRO 1	81624,31	47	336,4	27	8,67	83
CENTRO 2	48424,887	33	263,51	25	6,7	79
CENTRO 3	8516,76	16	128,17	14	3,03	65

Centros con m = 1,5 luego de 19 iteraciones

CENTRO 1	82502,45	47	335,21	27	8,73	79
CENTRO 2	48373,14	32	263,31	25	6,7	79
CENTRO 3	8176,48	15	124,93	14	2,99	64

Centros con m = 2 luego de 27 iteraciones

CENTRO 1	83003,35	47	333,92	27	8,75	83
CENTRO 2	48284,25	32	262,19	25	6,7	79
CENTRO 3	7621,46	15	119,17	13	2,93	64

Centros con m = 5 luego de 100 iteraciones y Pertenencias del Cliente 1098

CENTRO 1	74739,384	44	349,50	29	8,23	83
CENTRO 2	46531,22	32	260,66	26	6,59	78
CENTRO 3	5785,20	13	101,03	12	2,70	62

Centros con m = 12 luego de 68 iteraciones

CENTRO 1	64523,575	32	281,53	23	8,14	76
CENTRO 2	44949,09	30	248,67	25	6,51	76
CENTRO 3	5347,63	13	88,40	12	2,53	62

Pertenencias para distintos valores de m del Cliente 1098

M	Grupo 1	Grupo 2	Grupo 3
1,05	0,0019568	0,9980431	0,0000000
1,5	0,3084410	0,6875407	0,0040182
2	0,373997	0,581888	0,044115
5	0,447275	0,357784	0,194940
12	0,494371	0,278785	0,226843

Cuadro 3. Centros y pertenencias del cliente1098

Dada la naturaleza heurística del método, el cambio en alguno de sus parámetros produce cambios en la estructura del resultado. Al haber elegido m = 5, por ejemplo, cambiaron los centros. Por lo tanto, también cambiaron las distancias a los centros. A su vez, esto provoca el cambio en la totalidad del resultado y, por ende, de su interpretación.

Como se dijo anteriormente, a medida que el nivel de borrosidad aumenta, las pertenencias comienzan a estar menos diferenciadas entre sí como puede apreciarse en "Pertenencias para distintos valores de m del Cliente1098".

1.3 Reconocimiento de los grupos con Fuzzy C-Means Clustering

Se reportan entonces los resultados obtenidos.

Luego de 27 iteraciones, utilizando m = 2, c = 3 y TOL = 1.10^{-16}, se obtienen los siguientes centros de cada grupo:

CENTRO 1	83003,35	47	333,92	27	8,75	83
CENTRO 2	48284,25	32	262,19	25	6,7	79
CENTRO 3	7621,46	15	119,17	13	2,93	64

Cuadro 4.Centros para los tres grupos

Cada cliente recibe su coeficiente de pertenencia a cada grupo. Podemos ver algunos de ellos:

Cliente	Grupo 1	Grupo 2	Grupo 3	-
cliente368	0,053048	0,806397	0,140556	2
cliente369	0,001800	0,006663	0,991537	3
cliente370	0,565217	0,394981	0,039802	1
cliente371	0,002112	0,006717	0,991171	3
cliente372	0,984560	0,012995	0,002444	1
cliente373	0,931038	0,053872	0,015090	1
cliente374	0,056616	0,923984	0,019400	2
cliente375	0,033846	0,906259	0,059895	2

Cuadro 5. Pertenencias a cada grupo de los clientes 368 a 375

La última columna indica a qué grupo pertenece el dato según el mayor valor de pertenencia. Si se considera la pertenencia a un grupo como el máximo del valor de pertenencia, en el grupo 1 se ubican 165 clientes, en el grupo 2 se ubican 576, y en el grupo 3 quedan 663. Si se considera la suma de los coeficientes de pertenencia por grupo como la cantidad de individuos en un grupo: el grupo 1 contiene 186,6 datos, el grupo 2 tiene 560,2 y el grupo 3 termina con 657,2.

El resultado nos habla de tres grupos con las siguientes características:

Grupo 1

> Alto nivel de uso de la cuenta corriente – Alto endeudamiento. Altos montos de endeudamiento. El período de saldo de la situación deudora es de 47 días en promedio, haciendo gastos cercanos a los $300 pesos. Estos clientes llevan mucho tiempo trabajando con la entidad.

Grupo 2

Nivel medio de uso de la cuenta corriente – Endeudamiento medio. Tienen niveles de endeudamiento acorde a lo esperado por el tiempo que llevan con la entidad y, también, menores. El período de repago, en promedio, es de 32 días.

Grupo 3

Nivel bajo de uso de la cuenta corriente – Endeudamiento bajo. Niveles de endeudamiento muy bajos. Este grupo está totalmente diferenciado de los demás. Aquí se agrupan clientes nuevos y de corta pertenencia al banco. El índice de calificación de la entidad los evalúa con 64 en promedio.

La entidad clasificará a los clientes de acuerdo con los resultados que se obtuvieron por el método de agrupamiento utilizado. Evidentemente, deberá hacer un seguimiento más cuidadoso de los clientes del grupo 1 que se manifiestan como más riesgosos, dado su mayor nivel de endeudamiento.

1.4. Análisis discriminante canónico de los resultados obtenidos

Como se mencionó al inicio, el análisis realizado es estático y lo que a la entidad le interesa es poder clasificar a sus clientes en estos grupos a medida que se vayan incorporando y reclasificar a los existentes si se modifican sus características. Para ello se utilizará el análisis discriminante canónico.

Para llevar a cabo la construcción de las funciones canónicas discriminantes se han eliminado de la lista todos los clientes que presentaban valores de la función de pertenencia cercanos a dos o tres grupos. Por ello, dado que existen tres grupos, la máxima borrosidad

ocurre cuando la pertenencia es 1/3 para cada grupo. Entonces, se han eliminado los datos con pertenencias mayores a 1/3 y menores a 2/3 quedando 86 datos fuera y 1318 dentro del análisis.

Utilizando el SPSS, se obtienen dos funciones canónicas discriminantes. La primera tiene un alto poder discriminatorio por corresponder al autovalor mayor de la matriz de varianzas. En el siguiente cuadro se presentan los resultados:

Función	Autovalor	% de Varianza	Acumulativo	Correlación Canónica
1	13,298	99,5	99,5	0,964
2	0,065	0,5	100,0	0,246

Cuadro 6. Eigenvalores de las funciones canónicas discriminantes

La primera función obtenida explica el 99,5% de la variabilidad total. La segunda función aporta poco al análisis por lo que no la consideraremos. La función discriminante es la siguiente:

$$D_1 = -4,214426 + 0,000143.x_1 + 0,003289.x_2 + 0,000183.x_3 - \\ - 0,004255.x_4 - 0,060383.x_5 - 0,002469.x_6$$

donde,

- x_1 : Monto total de deuda acumulada.
- x_2 : Días promedio de retraso antes de pago de deuda.
- x_3 : Monto actual de la deuda.
- x_4 : Días de retraso de la deuda actual.
- x_5 : Años de pertenencia a la compañía.
- x_6 : Índice de la entidad.

En el gráfico que se presenta a continuación, se pueden observar las puntuaciones de cada individuo según las dos funciones obtenidas por el análisis discriminante.

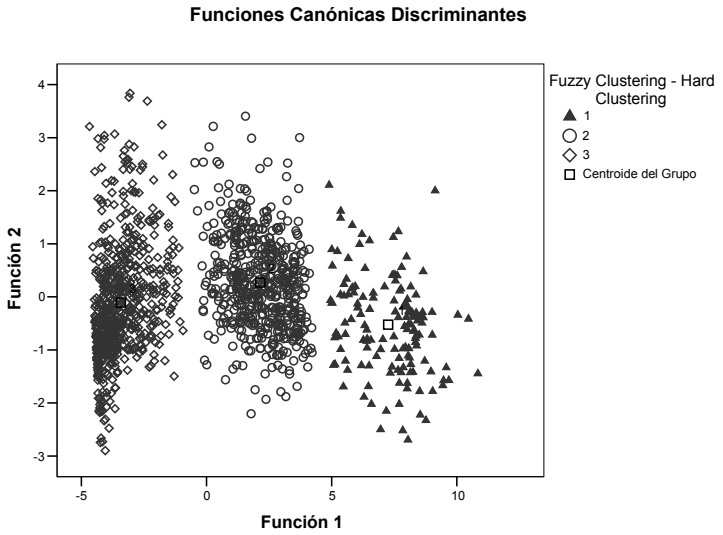


Figura 5. Puntuaciones de las funciones discriminantes 1 y 2

Los intervalos definidos para la asignación según el resultado que surge de la discriminación son:

$$D \geq 4,5 \;\Rightarrow\; cliente\, x \in G1$$
$$-0,5 \leq D < 4,5 \;\Rightarrow\; cliente\; x \in G2$$
$$D < -0,5 \;\Rightarrow\; cliente\; x \in G3$$

Con esta regla de decisión se han recalificado correctamente, o sea según la clasificación del *Fuzzy Clustering*, 74 de los 86 casos de borde anteriormente excluidos. Por lo tanto, aún en el caso de los individuos

difíciles de clasificar, la función discriminante permitió hacerlo con una efectividad del 86%. Esta herramienta más expeditiva permitirá a la entidad manejarse en forma sencilla espaciando la reclasificación completa de la cartera de clientes.

2. CONCLUSIÓN

La utilización del *Fuzzy Clustering* para el análisis de datos presenta información que no puede proporcionar el *Clustering* tradicional. El método ha progresado de tal manera que existe una amplia gama de variaciones que lo hacen más efectivo al momento de la utilización. Además, se han incorporado medidas de validación que ayudan al criterio del usuario.

En el campo de los negocios, este tipo de herramientas ayuda a hacer un análisis completo de grandes cantidades de datos. En el caso de nuestra entidad bancaria modelo y bajo los supuestos utilizados, el *Fuzzy Clustering* sirvió para poder reconocer los datos que presentaban ambigüedad respecto de los demás. Además, su resultado fue utilizado para crear una herramienta de asignación de nuevos clientes a los grupos existentes.

REFERENCIAS

[1] _____ (2003). *The economic and social structure of London and the South East. An analysis of the economic and social structure of London, the South East Region and the Eastern Region using the technique of Fuzzy Clustering*. Volterra Consulting Ltd, Reino Unido.

[2] Fraley, C.; Raftery, A. (2002). "Model-Based clustering, discriminant analysis and density estimation". *Journal of the American Statistical Association.* pp.611-631

[3] Fraley, C.: Raftery, A. (1998). "How many clusters? Which clustering method? Answers via model-based cluster analysis". *The computer Journal*. pp.578-588.

[4] Gath, I.; Geva, A. B. (1989). "Unsupervised optimal fuzzy clustering". *IEEE Transactions On Pattern Analysis and Machine Intelligence.* Vol. II, No. 7, pp.773–781.

[5] Hammah, R.; Curran, J. (2000). "Validity measures for the fuzzy cluster analysis of orientations". *IEEE Transactions On Pattern Analysis and Machine Intelligence*. Vol. XXII, No. 12, pp.1467–1472.

[6] Hsu, T.-H. (2000). "An application of fuzzy clustering in group-positioning analysis". *Proc. Natl. Sci. Counc.* Vol. X, No. 2, pp.157–167.

[7] Klawonn, F.; Poner, F. (2000). *What is fuzzy about fuzzy clustering? Understanding and improving the concept of fuzzified*. Department of Computer Science, University of Applied Sciences Braunschweig/Wolfenbuttel, Alemania.

[8] Klir, G.; Yuan, B. (1995). *Fuzzy sets and fuzzy logic. Theory and applications*. Prentice Hall, Estados Unidos.

[9] Koutsoupias, K.; Papadimitiou, I. (2000). "Utilizing hierarchical classification in life insurance data analysis". *Actas VII Congreso de SIGEF*, Grecia.

[10] Mendes, M.; Sacks, L. (1999). *Evaluating fuzzy clustering for relevance-based information access*. Dept. of E&EE, University College London, Reino Unido.

[11] Pal, N. R.; Bezdek, J. C. (1995). "On cluster validity for the Fuzzy C-Means Model". *IEEE Transactions On Fuzzy Systems*. Vol. III, No. 3, pp.370–379.

[12] Pedrycz, W. (2005). *Knowledge-based clustering. From data to information granules*, Wiley, Estados Unidos.

[13] Pérez, C. (2004). *Técnicas de análisis multivariante de datos. Aplicaciones con SPSS*. Pearson Educación S.A., Madrid.

[14] Xie, X.; Beni, G. (1991). "A validity measure for fuzzy clustering". *IEEE Transactions On Pattern Analysis and Machine Intelligence*. Vol. XIII, No. 8, pp.841–847.

[15] Zimmermann, H.J. (1991). *Fuzzy set theory and its applications. Second, revised edition*. Kluwer Academic Publishers, Estados Unidos.

ANEXO I: INTRODUCCIÓN AL CLUSTERING Y FUZZY CLUSTERING

El *clustering* es una técnica de *data mining* (exploración de datos) que consiste en agrupar datos según similitud o distancia. Esa técnica asigna con certeza cada dato al grupo al que se encuentra mejor emparentado. Por ejemplo, veamos la siguiente figura a título ilustrativo haciendo un análisis gráfico de los resultados.

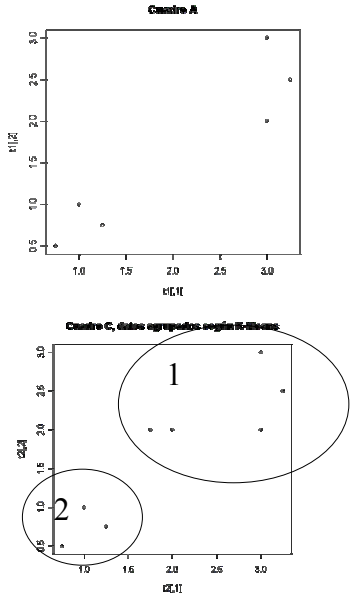


Figura 1

Cuadro D - Coeficientes fuzzy de pertenencia

	[,1]	[,2]
1	0,97194173	0,028058268
2	0,99468365	0,005316347
3	0,99145526	0,008544737
4	0,03459092	0,965409081
5	0,04935013	0,950649867
6	0,02884443	0,971155567
7	0,40757656	0,592423439
8	0,26023142	0,739768581

En el cuadro A, donde se ha trabajado con 6 datos, son fácilmente definibles dos grupos de datos y cualquier algoritmo que se elija los reconocerá rápidamente. En el cuadro B, se han agregado dos datos que se ubican en el medio entre los dos grupos. ¿A qué grupo pertenece cada dato? Si se hace el agrupamiento por el método de *clustering* tradicional utilizando el método *K-Means*, el resultado es el que se

expone en el cuadro C. Véase que el algoritmo ha asignado ambos datos al grupo 1 (de la derecha superior). Sin embargo, no necesariamente este es el caso con ambos datos. El cuadro D muestra la asignación que resulta de utilizar el *Fuzzy Clustering.* Este método asigna un grado de pertenencia a cada grupo para cada dato. Los datos iniciales tienen asignados coeficientes que implican casi certeza de pertenencia. Mientras que los nuevos datos muestran en sus coeficientes la ambigüedad de su pertenencia. Sería necesario ahora definir el umbral a partir del cual se define la pertenencia a un grupo específico.

El método de *Fuzzy Clustering* presenta la posibilidad de utilizarlo como apoyo cuando se debe trabajar con gran cantidad de datos, debido a su flexibilidad y su fácil interpretación. El problema de reconocer grupos, dado un conjunto grande de individuos (1000) para cada uno de los cuales hay más de una variable que se pretende comparar (3 variables por lo menos), nos presenta la necesidad de trabajar con matrices grandes (1000 x 3 = 3000 datos). Por otra parte, dado que la asignación categórica de un individuo a un grupo puede ocasionar la pérdida de la visión general del problema, sería mejor asignar, como se quiso ejemplificar antes, un nivel de pertenencia a cada grupo y, de esa forma, ver la posibilidad tanto de fusionar grupos como de separarlos o de considerar cierto rango de datos como pertenecientes a dos grupos simultáneamente.

Fuzzy Clustering. Fuzzy C-Means Clustering

Como vimos en la introducción, dado el conjunto de datos a agrupar y encontrados los patrones que agrupan los datos, parece conveniente asignar grados de pertenencia a ciertos datos cuando su distribución no es perfectamente divisible. Algunos de los datos con que contamos

pueden ser casos de borde de dos grupos al mismo tiempo como los que vimos en el ejemplo gracias a agrupar según la técnica *Fuzzy*.

El método de *Fuzzy Clustering* parte del siguiente problema de optimización[4]:

$$Min\,Q = Min_{u,\mathbf{v}}\left[\sum_{i=1}^{c}\sum_{k=1}^{N} g(u_{ik}).d(\mathbf{x}_k,\mathbf{v}_i)\right] \quad (1)$$

$$s.\,a\ \sum_{i=1}^{c} u_{ik} = 1 \quad (2)$$

donde Q es la función objetivo, c es el número de grupos que se quieren formar, N es el número de individuos con los que se cuenta, $\mathbf{v}_i$ es un vector prototipo o centro que se define más adelante, $\mathbf{x}_k$ es un vector correspondiente a cada dato, $g(u_{ik})$ es una función que depende de los factores de pertenencia u_{ik}, y $d(\ ,\ )$ representa la distancia pudiéndose elegir entre distintos tipos de definiciones de distancias que mejor describan el espacio en donde se trabaja.

La función g es tal que el cambio en el valor de la función objetivo sea mayor que el que el cambio en los valores de los factores de pertenencia. Entonces, g debe cumplir las siguientes restricciones:

---

[4] El problema surge de considerar, para el agrupamiento de datos, la optimización de una función objetivo del tipo:

$f = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}.d_{ik} \quad s.a \quad \sum_{i=1}^{c} u_{ik} = 1$

donde los u_{ij} , entre cero y uno, son ponderadores de cada distancia d_{ij}. Pero por ser todas las distancias positivas, la resolución es simple y se encontrará un óptimo asignando 1 a cada u_{ij} que acompañe a la menor de las distancias por cada *i*. Por ello, se incorpora la función ***g***.

$$g \in C^2[0;1]$$
$$g(0) = 0 \ \wedge \ g(1) = 1$$
$$g'(x) > 0 \quad \forall x \in [0;1]$$
$$g''(x) > 0 \quad \forall x \in [0;1]$$

La matriz de partición, que es la matriz cuyas filas representan a los individuos y en cuyas columnas se asignan los coeficientes de pertenencia para cada grupo, ya no será compuesta por unos y ceros según la pertenencia o no a un determinado grupo como en el *clustering* tradicional, sino que se representa por un grado de pertenencia a cada grupo. La matriz U que describe la pertenencia será:

$$\begin{bmatrix} u_{11} & u_{21} & \dots & u_{c1} \\ u_{12} & u_{22} & & u_{c2} \\ & & \ddots & \\ u_{1N} & u_{2N} & & u_{cN} \end{bmatrix}$$

Esta matriz debe cumplir dos condiciones:

- Los cluster formados no deben ser triviales. Entonces, para todo i:

$$0 < \sum_{k=1}^{N} u_{ik} < N$$

- La suma de los factores de pertenencia debe ser igual a 1. Entonces, para todo k:

$$\sum_{i=1}^{c} u_{ik} = 1$$

Uno de los algoritmos de *clustering* que contempla los casos de pertenencia parcial se llama *Fuzzy C-Means Clustering* (FCM). Este

algoritmo fue desarrollado por Dunn (1974)[5] y generalizado por Bezdek (1981)[6] y consiste en optimizar la siguiente función objetivo:

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} . d(\mathbf{x}_k, \mathbf{v}_i)$$

En este caso, la función *g* toma la forma: $g(u_{ik}) = u_{ik}^{m}$ y cumple las restricciones antes impuestas.

La solución de la minimización de la función objetivo, de forma de hallar los valores de la matriz U, se completa en dos pasos. Para desarrollar el método utilizaremos la distancia euclidea que es la empleada a lo largo del presente trabajo. Primero se seleccionan, en forma aleatoria o mediante algún criterio particular, los vectores que serán utilizados como prototipos o centros y que serán los vectores representativos de los datos de un grupo. Estos vectores se utilizarán en el primer paso y luego el método irá modificándolos a los efectos de buscar su configuración óptima. En el primer paso, se debe optimizar cada término de la suma de los clusters respecto de los valores de u_{ik}^{m} utilizando las condiciones. Con lo que se plantea:

$$V_s = \sum_{i=1}^{c} u_{is}^{m} . \left\| \mathbf{x}_s - \mathbf{v}_i \right\|^2 - \lambda \left(\sum_{i=1}^{c} u_{is}^{m} - 1 \right)$$

donde λ es el multiplicador de Lagrange. Las condiciones de primer orden para cada t son:

$$\frac{\partial V_s}{\partial u_{rs}} = m u_{rs}^{m-1} . \left\| \mathbf{x}_s - \mathbf{v}_r \right\|^2 - \lambda = 0$$

[5] J. C. Dunn (1974): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". *Journal of Cybernetics* 3, pp.32-57

[6] J. C. Bezdek (1981): *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York

despejando se llega a:

$$u_{rs.} = \left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} \frac{1}{\left\|\mathbf{x}_s - \mathbf{v}_r\right\|^{\frac{2}{m-1}}}$$

sumando miembro a miembro sobre *r* y recordando la restricción sobre la suma de los factores de pertenencia:

$$\sum_{i=1}^{c} u_{is} = \left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} \sum_{i=1}^{c} \frac{1}{\left\|\mathbf{x}_s - \mathbf{v}_i\right\|^{\frac{2}{m-1}}} = 1$$

con lo cual, el multiplicador de Lagrange será:

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{i=1}^{c} \frac{1}{\left\|\mathbf{x}_s - \mathbf{v}_i\right\|^{\frac{2}{m-1}}}}$$

entonces, se llega a:

$$u_{rs.} = \frac{1}{\sum_{i=1}^{c} \left(\frac{\left\|\mathbf{x}_s - \mathbf{v}_r\right\|}{\left\|\mathbf{x}_s - \mathbf{v}_i\right\|}\right)^{\frac{2}{m-1}}} \qquad (3)$$

De esta forma se obtienen los valores de las u_{ik} que minimizan cada uno de los términos de la función Q. Una vez hallados éstos, para obtener lo valores de los vectores $\mathbf{v}_i$ se debe buscar el mínimo de la función objetivo respecto de cada uno de esos vectores. Por ello, la

minimización será tal que las condiciones de primer orden, siempre pensando en la distancia euclidea, son:

$$\frac{\partial Q}{\partial \mathbf{v}_i} = -2\sum_{k=1}^{N} u_{ik}^m . \left(\mathbf{x}_k - \mathbf{v}_i\right) = 0$$

De lo que se desprende que:

$$\sum_{k=1}^{N} u_{ik}^m . \left(\mathbf{x}_k - \mathbf{v}_i\right) = 0 \;\Rightarrow\; \mathbf{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^m . \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^m} \qquad (4)$$

De haberse adoptado otras formas de distancias, por ejemplo la distancia de Hamming o norma Taxi, no se hubiese arribado a una solución tan sencilla y se hubiera requerido un mayor esfuerzo en la optimización. También se utiliza, como una herramienta más poderosa de descripción, la distancia exponencial o la distancia de Mahalanobis. La elección de la distancia dependerá del tipo de espacio de datos al que se esté enfrentando.

Resumiendo, el FCM es un proceso iterativo que consiste en la optimización de la función Q y sus pasos son los siguientes:

1) Se seleccionan los valores de c, m, un criterio de detención y una función de distancia apropiada.
2) Se definen los vectores que harán de prototipos.
3) Se repite:

   a) Cálculo de la matriz de pertenencia
   b) Cálculo de los vectores prototipo.
   c) Se evalúa el criterio de detención. Si se verifica se detiene el método, si no se verifica, se comienza desde (a) con los nuevos valores de u_{ik} y de. $\mathbf{v}_i$

El número de grupos elegido refleja el nivel de generalidad con el que evaluamos los datos. Al inicio del problema que se plantea, la cantidad de grupos a elegir depende estrictamente del conocimiento que se tenga de la información disponible. Existen métodos, como el Clustering Jerárquico (*Hierarquical Clustering*), para intentar encontrar la cantidad de grupos de que se debe disponer. Sin embargo, dicha metodología no evita la necesidad de decisiones tomadas por el usuario. Por supuesto, el caso trivial en el que se elije armar sólo un grupo no describe los datos con los que nos enfrentamos. Tampoco lo hace el otro caso trivial en el que se cuenta con n datos y se elige construir n grupos. Además, se observa que la función objetivo vista como función de la cantidad de grupos, decrece cuando la cantidad de grupos aumenta[7].

Medidas de validación

Para evaluar el resultado que se obtiene de aplicar el método existen tres medidas importantes: la primera está constituida por el coeficiente

[7] Pedrycz, W. *Knowledge-based clustering. From data to information granules*. Wiley, 2005

de partición y la entropía de la partición y las otras son los índices de Fukuyama-Sugeno 1989) y de Xie-Beni (1991).

Coeficiente de partición y entropía de la partición

El coeficiente de partición se define como:

$$CP = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^2}{N} \tag{5}$$

y el Coeficiente de entropía de la partición se define:

$$CE = -\frac{1}{n}.\left[\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}.\log_a(u_{ik})\right] \tag{6}$$

Las propiedades de estos índices como función de los coeficientes de pertenencia y la cantidad de grupos son las siguientes:

$CP = 1 \Leftrightarrow CE = 0 \Leftrightarrow U$ es matriz de partición del *clustering* tradicional

$$CP = \frac{1}{c} \Leftrightarrow CE = \log_a(c) \Leftrightarrow u_{ij} = \frac{1}{c}$$

Entonces, cuando se obtiene una partición con pertenencias unitarias correspondiente al *clustering* tradicional, se obtiene el máximo valor del coeficiente de partición y el mínimo valor del coeficiente de entropía. Cuando la borrosidad es máxima, o sea que todos los datos tienen igual pertenencia a cada grupo, el coeficiente de partición asume su mínimo valor y el de entropía su máximo valor.

Estos índices no hacen uso del total de información que proporciona el algoritmo y el conjunto de datos utilizado. Por el contrario, tienen en

cuenta la información proporcionada por los datos indirectamente a través de los coeficientes de borrosidad.

Índice de Fukuyama-Sugeno

El índice de Fukuyama-Sugeno se define:

$$IFS = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m}.\left(\left\|\mathbf{x}_k - \mathbf{v}_i\right\|^2 - \left\|\mathbf{v}_i - \overline{\mathbf{v}}\right\|^2\right) \qquad (7)$$

donde $\overline{\mathbf{v}}$ es el vector de medias correspondiente a todos los datos, también llamado gran media.

Este índice compara la distribución de los datos en sus grupos con la distribución de los grupos respecto de la totalidad de los datos, con lo cual, indica, a mayor valor, una peor descripción de los centros de cada grupo. Cuando se alcanza el mínimo de este índice, se está frente a un buen agrupamiento de los datos.

Se puede ver que:

$$IFS = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m}.\left(\left\|\mathbf{x}_k - \mathbf{v}_i\right\|^2 - \left\|\mathbf{v}_i - \overline{\mathbf{v}}\right\|^2\right) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m}.\left\|\mathbf{x}_k - \mathbf{v}_i\right\|^2 - \sum_{i=1}^{c}\left(\sum_{k=1}^{N} u_{ik}^{m}\right).\left\|\mathbf{v}_i - \overline{\mathbf{v}}\right\|^2$$

Si se recuerda que:

$$Q = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m}.d(\mathbf{x}_k, \mathbf{v}_i)^2$$

Y también se define:

$$\sum_{k=1}^{N} u_{ik}^{m} = N_i \quad \text{tal que} \quad \sum_{i=1}^{c} N_i = N$$

Se puede escribir el índice de la siguiente manera:

$$IFS = Q - \sum_{i=1}^{c} N_i . \left\| \mathbf{v}_i - \overline{\mathbf{v}} \right\|^2$$

En esta formulación del índice salta a la vista que se compara el valor de la función objetivo con la situación ideal en la que cada dato es prototipo de su grupo y, por lo tanto, todos los datos de un grupo están equidistantes al vector representativo de todo el conjunto de datos.

Índice de Xie-Beni

El índice de Xie-Beni se define:

$$IFS = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} . \left\| \mathbf{x}_k - \mathbf{v}_i \right\|^2}{n . \left(\min \left\| \mathbf{v}_i - \mathbf{v}_j \right\|^2 \right)} \qquad (8)$$

El denominador se refiere a la separación mínima entre los centros de los datos agrupados. Tomando esos centros como representativos de lo que sucede con todos los datos, la mínima distancia entre ellos representa la situación ideal de la separación de todos los datos. Por lo tanto, ponderar el valor de la función objetivo por este valor mínimo, equivaldrá a señalar qué tan separados se encuentran los grupos. Entonces, si IFS1<IFS2, se tiene que IFS1 es una mejor partición que IFS2.

Estos dos últimos índices, a diferencia de los primeros dos, hacen uso del total de la información suministrada por el algoritmo. Emplean la cantidad de grupos, la función de distancia utilizada, los centros establecidos y los coeficientes de borrosidad. Los resultados a los que se llegan con el empleo de estos índices no son necesariamente los

mismos. Mientras que el coeficiente de partición y el de entropía deberían coincidir, no necesariamente sucederá ello con el índice Xie-Beni y con el Fukuyama-Sugeno. Para ver una ilustración de estas conclusiones puede verse Pal y Bezdek (1995).