

Segmentación de los solicitantes de crédito a un banco usando análisis factorial múltiple

Clustering of the Applicants for Loans to a Bank using Multiple Factor Analysis

JOHANA CARDONA^a, LINA ALEXANDRA OSORIO^b, MABELIN VILLARREAL^c, CAMPO ELÍAS PARDO^d

DEPARTAMENTO DE ESTADÍSTICA, FACULTAD DE CIENCIAS, UNIVERSIDAD NACIONAL DE COLOMBIA, BOGOTÁ, COLOMBIA

Resumen

Se presenta la implementación del análisis factorial múltiple como metodología previa a una clasificación de solicitantes de crédito en una entidad bancaria del país en Bogotá. Se obtienen tres clases utilizando las características sociodemográficas y económicas (cinco variables cuantitativas y seis cualitativas), que las personas suministran al Banco al diligenciar las solicitudes de crédito. En las clases obtenidas se encuentran diferencias en sus porcentajes de clientes y en sus distribuciones de tipos de crédito.

Palabras clave: análisis multivariado de datos, clasificación jerárquica, K -medias, tipo de crédito.

Abstract

We illustrate how to use Multiple Factor Analysis as an initial methodology in order to cluster applicants for loans in a Colombian Financial Entity in Bogota City. We obtain three groups using economic and sociodemographic features (five quantitative and six qualitative variables) given by applicants when they fill out the application forms. We find differences between the obtained groups in their percentage of customers and in their distributions of loan types.

Key words: Multivariate data analysis, Hierarchical clustering, K -means clustering, Segmentation, Credit type.

1. Introducción

La gran diversidad social, económica y cultural propia de la idiosincrasia nacional, genera una clara distinción entre los colombianos que se refleja desde las costumbres del diario vivir, hasta características tan complejas como el comportamiento de pago en una entidad bancaria. Aunque estas diferencias, observables incluso entre personas que residen o son originarias de un mismo lugar, hacen de Colombia un país pluricultural, representan también un problema para el ámbito financiero en cuanto a la otorgamiento de créditos se refiere, pues es difícil determinar si un individuo en particular tendrá o no un buen hábito de pago.

Siendo Bogotá el principal centro financiero nacional y a su vez un gran receptor de población proveniente de otros lugares del país, esta ciudad se convierte en un nicho de diversidad socioeconómica útil para la actualización de los estándares crediticios en el país y la implementación de nuevas metodologías

^aEstudiante de Estadística. E-mail: jcardonah@unal.edu.co

^bEstudiante de Estadística. E-mail: laosorioc@unal.edu.co

^cEstudiante de Estadística. E-mail: mvillarreal@unal.edu.co

^dProfesor asociado. E-mail: cepardot@unal.edu.co

que permitan el aprovechamiento de la infinidad de perfiles existentes mediante la creación de grupos de clientes con características sociodemográficas similares.

Así, en el momento del ingreso de un nuevo cliente solicitante de crédito residente en Bogotá, este podrá ser asignado a un grupo específico. Se supone que la construcción de reglas de discriminación (clasificación supervisada) para cada uno de los grupos es más eficiente que para la población global de clientes.

En los datos disponibles generalmente se encuentran variables cuantitativas y cualitativas. Para utilizarlas conjuntamente se suelen discretizar las variables cuantitativas o cuantificar las variables cualitativas. En este documento se presenta la utilización del análisis factorial múltiple como otra alternativa.

2. Metodología

El procedimiento para la obtención de los grupos es el propuesto por Lebart et al. (1995), utilizando para la clasificación las coordenadas factoriales de un análisis factorial múltiple (Pagés 1992, Pagès 2004), que permite la combinación de variables cuantitativas y cualitativas como activas en el análisis. En la figura 1 se muestra un diagrama de flujo del procedimiento.

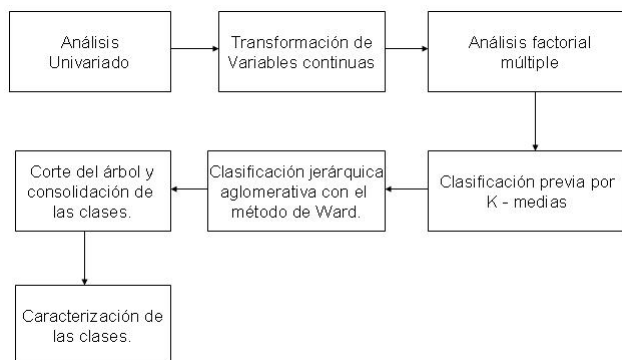


FIGURA 1: Diagrama de flujo del procedimiento

2.1. Los datos

La tabla de datos tiene cinco variables cuantitativas y nueve variables cualitativas de 5860 personas naturales que solicitaron créditos a una entidad financiera del país, en la ciudad de Bogotá. La información se asocia a condiciones sociodemográficas y económicas (ver tabla 1) de los individuos interesados en adquirir un crédito con la entidad bancaria en el año 2007. Las dos últimas variables son *ilustrativas* en el análisis.

2.2. Transformaciones a las variables cuantitativas

Al realizar la exploración de las distribuciones de cada una de las cinco variables (tendencia central, forma y dispersión), se hace evidente la necesidad de una transformación en algunas de ellas, debido a la presencia de colas muy alargadas que puedan influenciar la adecuada descripción de los clientes. En las figuras 2 a 5 se muestran los histogramas de las variables originales y transformadas.

2.3. Variables categóricas

En la tabla 2 para cada variable se presenta la distribución, en porcentaje, de los clientes en sus categorías.

TABLA 1: Descripción de las variables en estudio

Tipo de variable	Variable
Cuantitativas	Edad
	Antigüedad en la vivienda
	Tiempo de empleo
	Antigüedad en el empleo actual
	Ingresos
Cualitativas	Sexo
	Estado civil
	Tipo de vivienda de vivienda
	No de personas a cargo
	Tipo de cargo
Ilustrativas	Cliente
	Tipo de crédito

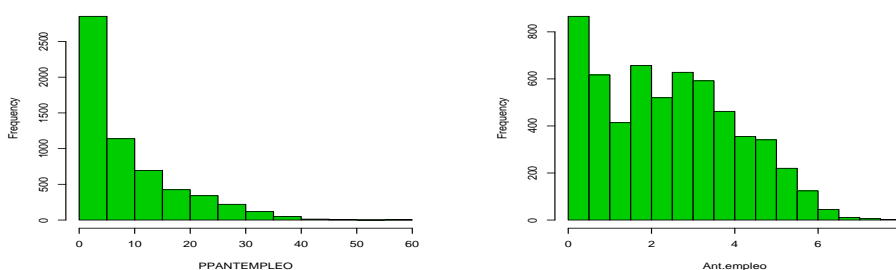


FIGURA 2: Antigüedad en el empleo actual y Raíz cuadrada de antigüedad en el empleo actual

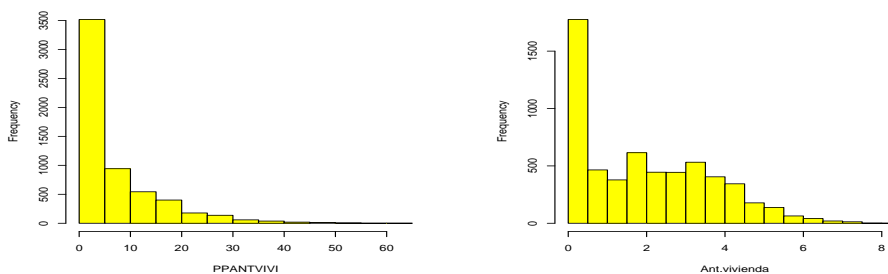


FIGURA 3: Antigüedad en la vivienda y Raíz cuadrada de la antigüedad en la vivienda

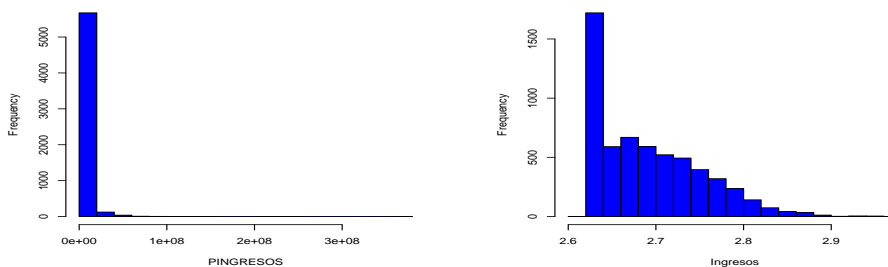


FIGURA 4: Ingresos y Ln(Ln(Ingresos))

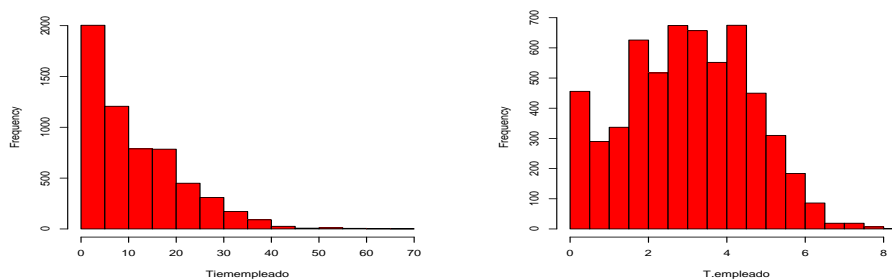


FIGURA 5: Tiempo de empleado y Raíz cuadrada de Tiempo de empleado

TABLA 2: Variables categóricas

Sexo	%	No. personas a cargo	%
Hombre	54.7	Cero	0.1
Mujer	45.3	Una persona	45.7
		2 ó más	54.3
Estado civil	%	Tipo de cargo	%
Separado	9.2	Empleado calificado	34.6
Casado	57.0	Empleado	17.4
Soltero	27.7	Comerciante	12.5
Unión libre	9.3	Otra profesión	13.2
Viudo	2.9	Jubilado	7.4
		Otro cargo	14.8
Tipo de vivienda	%	Tipo de contrato	%
Familiar	17.6	Indefinido	60.2
Propia con hipoteca	20.7	Independiente	29.9
Propia sin hipoteca	49.4	Otra	8.0
Rentada	12.3	Fijo, temporal	1.9

2.4. Análisis factorial múltiple

Para lograr la combinación en el análisis de las cinco variables continuas y las seis categóricas, se realiza un análisis factorial múltiple (AFM), el cual consiste en la ejecución de un análisis en componentes principales (ACP) normado para el grupo de las variables continuas y un análisis de correspondencias múltiples para el grupo de las variables categóricas y la realización posterior de un ACP conjunto, ponderando las variables de cada grupo con el inverso del primer valor propio del análisis precedente (Escofier & Pagès, 1992).

El histograma de valores propios (figura 6) muestra que los primeros cuatro ejes factoriales son suficientes para una lograr representación adecuada de la información.

En la figura 7 se observa que el primer eje recoge la información de los dos grupos de variables, mientras que el segundo esta determinado prácticamente por el grupo de variables categóricas. El círculo de correlaciones muestra que el primer factor esta establecido como un factor tamaño, pues todas las variables cuantitativas se correlacionan positivamente con este eje.

Finalmente, en la figura 8, se muestra la proyección de la nube de puntos de los individuos sobre el primer plano factorial, evidenciando una clara partición de los clientes en dos grupos, uno de ellos cercano al promedio y otro mas alejado de éste.

2.5. Análisis de conglomerados

Con el objetivo de segmentar la población, se realiza un K -medias con 1000 grupos, previo a una clasificación jerárquica aglomerativa mediante el método de Ward. El histograma de índices de nivel

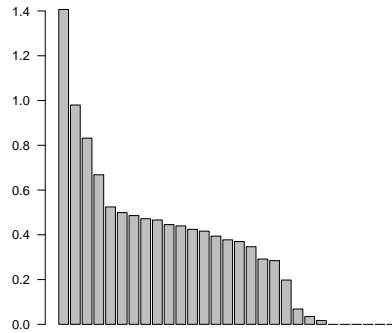


FIGURA 6: Histograma de valores propios

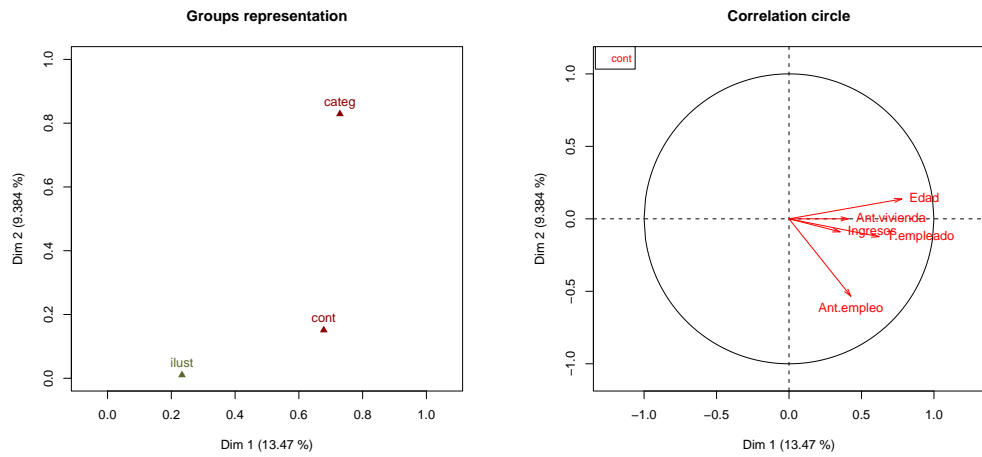


FIGURA 7: Primer plano factorial de los grupos de variables y círculo de correlaciones de las variables continuas

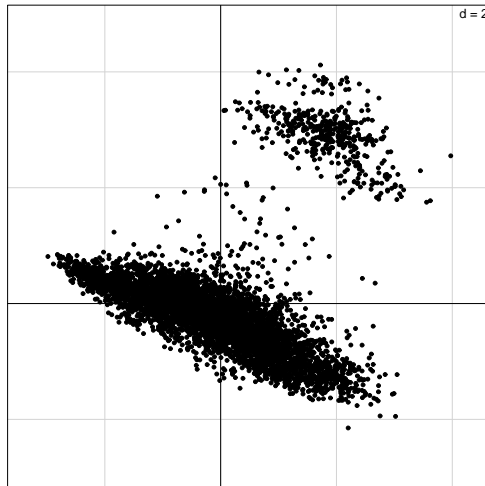


FIGURA 8: Proyección de la nube de puntos de individuos en el primer plano factorial

(figura 9) y el dendrograma (figura 10) evidencian un buen corte en tres clases. Finalmente se realiza la consolidación de la clasificación mediante un proceso de K -medias en 3 clases, tomando como punto iniciales los centros de gravedad de los grupos generados anteriormente. En la figura 11 se muestra la conformación de los tres grupos sobre le primer plano factorial del AFM.

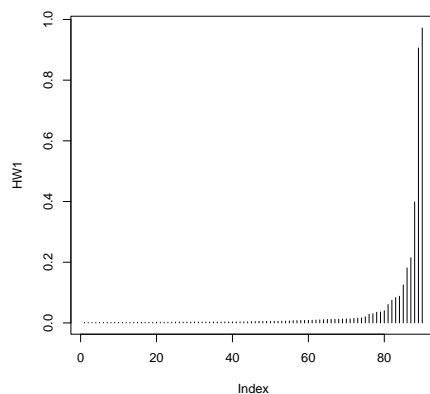


FIGURA 9: Histograma de índices de nivel de la clasificación jerárquica de los 1000 grupos obtenidos con el K -medias inicial

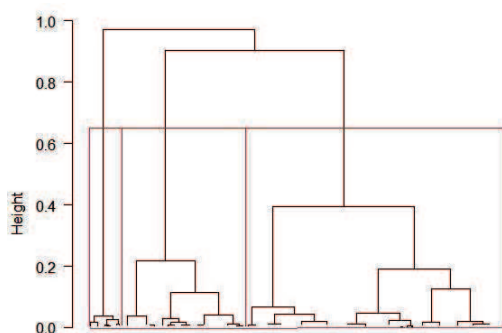


FIGURA 10: Dendrograma de la clasificación jerárquica con el método de Ward

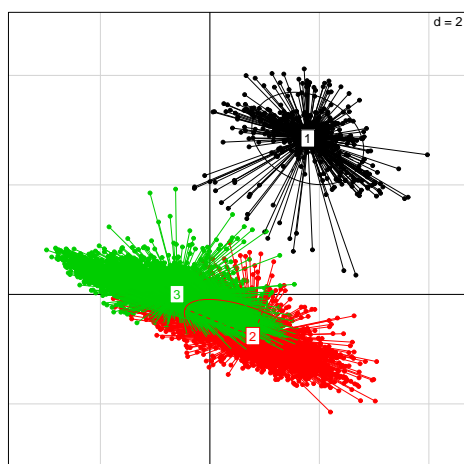


FIGURA 11: Primer plano factorial del AFM con la proyección de las tres clases

2.6. Caracterización de las clases

Para caracterizar a las clases con las variables cuantitativas y cualitativas disponibles se utiliza el mecanismo de los valores *test* (Lebart et al. 1995), utilizando la función *cluster.carac* de *FactoClass* (Pardo & Del Campo 2007).

Con los valores *test* se detectan las categorías de las variables cualitativas que caracterizan a cada una de las clases, en el sentido de que su porcentaje dentro de la clase es razonablemente superior al porcentaje global. En las variables continuas los valores *test* resultan de la comparación del promedio de la variable dentro de la clase con el promedio global. Los valores *test* son cuantiles de la distribución normal estándar.

A continuación se resumen las características de las clases obtenidas.

■ Clase 1 (tabla 3)

Las 461 personas que hacen parte de este grupo, correspondientes al 7.8% de la población, se caracterizan por poseer vivienda sin hipoteca (74.6%); ser personas jubiladas, religiosas, amas de casa o rentistas (94.6%); casadas o viudas (75.7%); y además el 52.9% tienen una persona que depende económicamente de ellos. La edad promedio de este grupo es 60.5 años, los ingresos oscilan alrededor de 2.900 millones, habiendo trabajado en promedio 17 años y tienen una antigüedad en la vivienda en promedio de siete años.

Por las características anteriormente descritas, asignaremos a este grupo el nombre de *Adultos mayores estables*.

■ Clase 2 (tabla 4)

Este grupo lo conforman 1744 individuos, correspondientes al 29.7% de la población, los cuales son en su mayoría hombres (64.2%), casados (64.6%) y poseen vivienda sin hipoteca (61.5%). El 86% son comerciantes, abogados, administradores, psicólogos entre otros, cuyo contrato es de carácter independiente. Los individuos de este grupo han trabajado en promedio 8.4 años en su empleo actual, y están alrededor de los 50.8 años de edad además de residir en la misma vivienda aproximadamente 4.4 años.

De acuerdo con estos resultados, el nombre apropiado para este grupo es: *Independientes estables*.

■ Clase 3 (tabla 5)

Al siguiente segmento identificado pertenecen 3655 clientes que corresponden al 62.3% de la población, conformado por un 50.5% de mujeres, 27.5% de personas solteras y 59.6% de individuos que habitan aproximadamente hace 3.6 años en vivienda familiar, con hipoteca o rentada; son clientes empleados con contrato temporal o a término indefinido. El promedio de edad de este grupo es de 43 años y su antigüedad en el empleo actual es de 8.4 años en promedio.

Teniendo en cuenta estas características, se asigna el nombre *Empleados inestables*.

TABLA 3: Caracterización de la clase 1

	V.test	clas.mod	mod.clas	Global	peso		V.test	mediaClase	mediaGlobal
Tipo.vivienda.PSH	Inf	11.9	74.6	49.4	2895	Edad	27.9	60.5	46.7
Tipo.cargo.Otro.cargo	Inf	100.0	94.6	7.4	436	T.empleado	14.1	4.1	3.1
Tipo.contrato.O	Inf	98.1	100.0	8.0	470	Ant.vivienda	8.2	2.6	2.0
Est.civil.VIU	7.6	27.1	10.0	2.9	170	Ingresos	4.6	2.7	2.7
Est.civil.CAS	3.9	9.1	65.7	57.0	3339				
No.personas.Una persona	3.2	9.1	52.9	45.7	2676				

TABLA 4: Caracterización de la clase 2

	V.test	clas.mod	mod.clas	Global	peso		V.test	mediaClase	mediaGlobal
Sexo.H	Inf	34.9	64.2	54.7	3206	Edad	18.2	50.8	46.7
Tipo.vivienda.PSH	Inf	37.1	61.5	49.4	2895	Ingresos	12.4	2.7	2.7
Tipo.cargo.Comerciante	Inf	100.0	42.1	12.5	735	Ant.empleo	12.4	2.9	2.5
Tipo.cargo.Otra.profes	Inf	99.4	43.9	13.2	771	Ant.vivienda	3.6	2.1	2.0
Tipo.contrato.I	Inf	98.5	98.9	29.9	1750				
Est.civil.CAS	7.7	33.8	64.6	57.0	3339				
No.personas.2 o mas	4.4	32.2	58.7	54.3	3180				

TABLA 5: Caracterización de la clase 3

	V.test	clas.mod	mod.clas	Global	peso		V.test	mediaClase	mediaGlobal
Sexo.M	Inf	69.6	50.5	45.3	2654	Ant.vivienda	-7.9	1.9	2.0
Est.civil.SOL	Inf	79.2	27.5	21.7	1269	T.Empleado	-9.1	2.9	3.1
Tipo.vivienda.FAM	Inf	83.8	23.7	17.6	1033	Ingresos	-14.3	2.7	2.7
Tipo.cargo.Emp.calif	Inf	100.0	55.5	34.6	2030	Edad	-32.7	43.0	46.7
Tipo.cargo.Empleado	Inf	99.9	27.9	17.4	1021				
Tipo.contrato.F	Inf	99.6	96.1	60.2	3526				
Tipo.contrato.T	Inf	95.6	3.0	1.9	114				
Tipo.cargo.Otro	4.6	69.3	16.4	14.8	867				
Tipo.vivienda.PCH	4.0	67.3	22.3	20.7	1211				
Tipo.vivienda.REN	3.8	68.8	13.6	12.3	721				

2.7. Comparación del porcentaje de clientes y de la distribución de tipos de crédito entre las clases

Las variables *cliente* y *tipo de crédito* no participan en la construcción de las clases y por lo tanto la presencia de diferencias se interpretan como asociación entre ellas y las clases.

- Clase 1 (tabla 6)

El 61.4% *Adultos mayores estables* son clientes del banco y suelen solicitar créditos de consumo o tarjeta de crédito (61.1%).

- Clase 2 (tabla 7)

De los *Independientes estables* el 62.3% son clientes del banco y tienen créditos de consumo, hipotecarios y/o tarjeta de crédito.

- Clase 3 (tabla 8)

El 73% de los *Empleados inestables* no son clientes del banco y el producto que solicitan comúnmente es la tarjeta de crédito (70.6%).

TABLA 6: Caracterización de la clase 1 mediante las categorías de las variables ilustrativas

	V.test	clas.mod	mod.clas	Global	peso
Cliente.S	Inf	12.0	61.4	40.2	2357
Tipo.credito.C.consumo	8.1	16.2	26.2	12.8	748
Tipo.credito.C.cons y tarj	5.4	11.4	34.9	24.2	1416

TABLA 7: Caracterización de la clase 2 mediante las categorías de las variables ilustrativas

	V.test	clas.mod	mod.clas	Global	peso
Cliente.S	Inf	46.1	62.3	40.2	2357
Tipo.credito.C.cons y tarj	Inf	45.3	36.8	24.2	1416
Tipo.credito.C.consumo	Inf	49.6	21.3	12.8	748
Tipo.credito.cons,hip,tar	3.6	46.3	2.9	1.8	108
Tipo.credito.C. hip	3.1	53.7	1.3	0.7	41

TABLA 8: Caracterización de la clase 3 mediante las categorías de las variables ilustrativas

	V.test	clas.mod	mod.clas	Global	peso
Cliente.N	Inf	76.1	73.0	59.8	3503
Tipo.credito.tarjeta	Inf	78.6	70.6	56.0	3280

3. Conclusiones

Se muestra en la practica que el procedimiento de realizar un AFM previo a la clasificación es una buena estrategia de combinar variables cuantitativas y cualitativas en los procesos de segmentación de clientes.

Las 5860 personas naturales que solicitaron créditos a la Entidad Bancaria en Bogotá se dividieron en tres clases bien diferenciadas (sección 2.6). Se identifican algunas diferencias, entre las clases, en las distribuciones de los tipos de préstamos:

- La forma más común en la que los *empleados inestables*, los más jóvenes de la población en estudio, inician su historial crediticio con esta entidad, es mediante la solicitud de tarjetas de crédito; mientras que los *adultos mayores estables*, gracias a un vínculo previo a la solicitud del préstamo y a su estabilidad económica, prefieren adquirir además de tarjetas de crédito, créditos de consumo.
- Las personas que pertenecen al grupo de *independientes estables*, suelen pedir créditos de los tres tipos, consumo, hipotecario o tarjeta de crédito, y generalmente adquieren dos o más de ellos a la vez.

Software

En el procesamiento de la información se utilizó el programa estadístico R (R Development Core Team 2008); los paquetes: *ade4* (Chessel et al. 2004), *FactoMineR* (Husson et al. 2007) y *FactoClass* (Pardo & Del Campo 2007); y parte del código en R de Sardinle (2007).

Referencias

- Chessel, D., Dufour, A.-B. & Thioulouse, J. (2004), ‘The ade4 package-I- One-table methods’, *R News* **4**, 5–10.
- Husson, F., Lê, S. & Mazet, J. (2007), *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.05.
*<http://factominer.free.fr>, <http://www.agrocampus-rennes.fr/math/>
- Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- Pagès, J. (2004), ‘Multiple factor analysis: Main features and application to sensory data’, *Revista Colombiana de Estadística* **27**(1), 1–26.
*<http://www.matematicas.unal.edu.co/revcoles>

- Pagés, E. . (1992), *Análisis factoriales simples y múltiples*, Servicio editorial de la Universidad del país Vasco.
- Pardo, C. E. & Del Campo, P. C. (2007), ‘Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass’, *Revista Colombiana de Estadística* **30**(2).
*<http://www.matematicas.unal.edu.co/revcoles/>
- R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*<http://www.R-project.org>
- Sadinle, M. (2007), Anexo técnico. Análisis de clasificación. Código en R, Technical report, Centro de Recursos para el Análisis de Conflictos - CERAC.