

XXVIII Simposio Internacional de Estadística 2018.

Una adaptación del método STATIS para el análisis no simétrico de bloques de variables cualitativas 2018.

JENNYFER COMBARIZA^{1,a}, GUILLERMO RAMIREZ^{1,b}, MAURA VASQUEZ^{1,c}

¹POSTGRADO EN ESTADÍSTICA, FACULTAD DE CIENCIAS ECONÓMICAS Y SOCIALES, UNIVERSIDAD CENTRAL DE VENEZUELA, CARACAS, VENEZUELA

Resumen

Se hace una adaptación del método STATIS, creado para el análisis simétrico de datos cuantitativos de tipo cúbico, a datos no simétrico catégoricos de tipo cúbico. Realizando una aplicación a un conjunto de datos del sector financiero colombiano.

Palabras clave: STATIS, τ de Goodman-Kruskal, Análisis de correspondencia no simétrico .

Abstract

We adapted the STATIS method, developed for the simmetrical analysis of three-way quantitative data, to non symmetrical of three-way categorical data. We Carry out an application to set of data of the Colombian financial sector.

Key words: STATIS, Goodman-Kruskal τ , Non-Symmetric Correspondence Analysis .

1. Introducción

Predicción es el término usualmente utilizado para referirse a la estimación de una variable criterio y como función de una o más variables independientes x_1, x_2, \dots, x_p , en un modelo lineal o no lineal, lo que implica explicar la influencia que las últimas ejercen como determinantes del comportamiento de la primera. En el caso particular de una única variable explicativa x , y que tanto ésta como la y sean catégoricas, D'Ambra y Lauro (1989) han propuesto el análisis de correspondencias no simétrico (ACNS) del bloque apareado de información (x, y) . Esta técnica, basada en la descomposición del estadístico τ de Goodman-Kruskal, plantea obtener estimaciones mínimo cuadrática de y , obtenidas al proyectarla ortogonalmente sobre el espacio generado por las catégorías de x . Lo anterior consiste esencialmente en proyectar el vector de probabilidades condicionales de y , fijadas cada una de las catégorías de x , sobre las direcciones principales de la matriz de varianzas y covarianzas de las estimaciones de y dado x .

En este trabajo se propone una adaptación del método STATIS al caso de la aplicación de un ACNS en K oportunidades, a una estructura de datos correspondiente a la caracterización de los mismos individuos mediante dos variables catégoricas, organizada sobre bloques apareados no simétricos (x_k, y_k) , $k = 1, 2, \dots, K$.

^aTesista doctoral. E-mail: jennyfer.combariza@gmail.com

^bProfesor titular. E-mail: guillermo.ramirez.ucv@gmail.com

^cProfesor titular. E-mail: mauralvasquez@gmail.com

2. STATIS-CGK

La propuesta que se presenta tiene como propósito fundamental analizar los cambios que se producen en la influencia que tiene una variable explicativa categórica \mathbf{x} sobre una variable categórica a explicar \mathbf{y} , a lo largo de \mathbf{K} ocasiones. La conocida metodología STATIS fue inicialmente diseñada para el tratamiento de información generada al caracterizar \mathbf{n} individuos según \mathbf{p} variables continuas en \mathbf{K} oportunidades diferentes; en este trabajo se adapta esa metodología al caso de bloques apareados no simétricos de dos variables cualitativas, del tipo $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, 2, \dots, K$.

2.1. Análisis interno de cada bloque con un enfoque ACNS

El tratamiento analítico de la información dentro del bloque \mathbf{k} , plantea efectuar un ACNS, lo que implica analizar la variabilidad de \mathbf{y}_k explicada por \mathbf{x}_k , mediante la descomposición del estadístico $\tau_{\mathbf{y}_k \cdot \mathbf{x}_k}$ de Goodman-Kruskal (1954), de acuerdo con las siguientes etapas:

2.1.1. Etapa 1

Se obtienen las estimaciones mínimo cuadrática de \mathbf{y}_k , sobre el espacio generado por las categorías de la variable \mathbf{x}_k :

$$\tilde{Y}_k = X_k(X_k^t X_k)^{-1} X_k^t Y_k^1. \quad (1)$$

En este arreglo, la estimación mínimo cuadrática que corresponde a la variable \mathbf{y}_k en un individuo en el que está presente la categoría \mathbf{i} de la variable \mathbf{x}_k queda definida por el vector de probabilidades condicionales estimadas:

$$\left(\tilde{P}(y_k = 1 | x_k = i) \quad \cdots \quad \tilde{P}(y_k = j | x_k = i) \quad \cdots \quad \tilde{P}(y_k = q | x_k = i) \right) \quad (2)$$

Para la exposición se asume de aquí en adelante, que tanto \mathbf{Y}_k como $\tilde{\mathbf{Y}}_k$ están centradas.

2.1.2. Etapa 2

Se construye el cociente entre variabilidad de las estimaciones (variabilidad explicada) y la variabilidad total de la variable \mathbf{y} ,

1. Variabilidad total: $VT(Y_k) = \text{traza}(Y_{ck}^t Y_{ck})$.
2. Variabilidad explicada: $VE(Y_k) = \text{traza}(Y_k^t P_{X_k} Y_k)$.

encontrándose que este cociente coincide con el estadístico de Goodman-Kruskal:

Definición 1. El $\tau_{y_k \cdot x_k}$ de Goodman-Kruskal se define como

$$\tau_{y_k \cdot x_k} = \frac{VE(Y_k)}{VT(Y_k)}. \quad (3)$$

2.1.3. Etapa 3

Se determinan las direcciones principales de un ACNS, basado en la descomposición de $\tau_{y_k \cdot x_k}$ en sumandos, cada uno de los cuales mide la porción de la variabilidad de \mathbf{y}_k explicada por \mathbf{x}_k , captada por la correspondiente dirección principal. Con este fin se obtienen los autovalores y autovectores de la matriz $\tilde{\mathbf{Y}}_k^t \tilde{\mathbf{Y}}_k$:

$$\tilde{Y}_k^t \tilde{Y}_k v^{\alpha_k} = \lambda_{\alpha_k} v^{\alpha_k}, \quad \alpha_k = 1, \dots, q, \quad k = 1, \dots, K. \quad (4)$$

calculando además para cada dirección el cociente $\frac{\lambda_{\alpha_k}}{\tau_{\mathbf{y}_k \cdot \mathbf{x}_k}}$, que mide la proporción de la variabilidad de \mathbf{y}_k explicada por \mathbf{x}_k , captada por la α_k -ésima dirección principal.

¹ \mathbf{X}_k e \mathbf{Y}_k denotan la matrices disyuntivas asociadas a las variables categóricas \mathbf{x}_k e \mathbf{y}_k , respectivamente.

2.1.4. Etapa 4

Se construye la representación biplot de las probabilidades condicionadas estimadas de \mathbf{y}_k dado \mathbf{x}_k , y de las categorías de la variable \mathbf{x}_k , donde el vector de coordenadas de proyección de los \mathbf{n} individuos sobre el eje α_k :

$$\psi_{\alpha_k} = \tilde{Y}_k v^{\alpha_k}, \quad \alpha_k = 1, \dots, q, \quad k = 1, \dots, K. \quad (5)$$

así que la proyección de un individuo \mathbf{s}_k que posea la categoría i de \mathbf{x}_k (proyección de la i -ésima categoría de \mathbf{x}_k) viene dada por:

$$\psi_{s\alpha_k} = \sum_{j=1}^q v_{j\alpha_k} P(y_k = j | x_k = i) \quad (6)$$

y las coordenadas de proyección de la j -ésima categoría de \mathbf{y}_k :

$$\varphi_{j\alpha_k} = v_{j\alpha_k}, \quad j, \alpha_k = 1, \dots, q, \quad k = 1, \dots, K. \quad (7)$$

2.2. STATIS-CGK

Seguidamente se describen los distintos elementos requeridos para la aplicación de la fase de interestructura de la adaptación del STATIS, basado en el producto escalar de Frobenius.

2.2.1. Etapa 1: estudio dentro del k -ésimo bloque

En el STATIS-CGK se define un estudio como la triplete de la forma

$$((x_k, y_k), M, D) \quad (8)$$

Donde:

1. $\{x_1, \dots, x_K\}$ son K variables cualitativas, con p modalidades cada una.
2. $\{y_1, \dots, y_K\}$ son K variables cualitativas, con q modalidades cada una.
3. (x_k, y_k) , $k = 1, \dots, K$, representan bloques apareados no simétricos, que contiene la medición de dos variables cualitativas sobre n individuos.
4. M es una matriz definida positiva que define la métrica utilizada para construir las distancias entre los individuos.
5. D es una matriz definida positiva, usualmente diagonal, utilizada en el cálculo de las interrelaciones entre esas variables.

Para desarrollar una metodología análoga a STATIS que sirva para abordar esta investigación, se requiere identificar con claridad objetos cuya estructura permita analizar las relaciones entre los bloques no simétricos (x_k, x_K) , $k = 1, \dots, K$. En adelante utilizaremos la siguiente notación

1. $\{X_1, \dots, X_K\}$ son K matrices disyuntivas, asociadas a K variables cualitativas $\{x_1, \dots, x_K\}$ con p modalidades cada una.
2. $\{Y_1, \dots, Y_K\}$ son K matrices disyuntivas, asociadas a K variables cualitativas $\{y_1, \dots, y_K\}$ con q modalidades cada una.

Definición 2 (Definición de objeto no simétrico). Dado un estudio $((x_k y_k), M, D)$, se define el objeto mediante la siguiente expresión

$$W_k = \frac{1}{\sqrt{VT(Y_k)}} (P_{X_k} - P_m) Y_k$$

en este caso particular para cada escenario k se utiliza $M = D_{pk} = (X_k^t X_k)^{-1}$.

El objeto W_k resulta ser una matriz de orden $(n \times n) \times (n \times q) = n \times q$; como puede observarse el orden de este arreglo resulta independiente de p , el número de modalidades de la variable explicativa, más si depende del número de individuos bajo estudio y el número de categorías q de las variables criterio. Estos objetos no son matrices cuadradas como en el STATIS clásico y menos aún simétricas.

Definición 3 (Producto interno de Frobenius). Se define el producto interno de tipo Frobenius entre dos matrices W_h y W_k , de dimensión $n \times p$, de la forma

$$\langle W_h | W_k \rangle_F = \text{traza}(W_h^t W_k) = \frac{1}{\sqrt{VT(Y_K)VT(Y_h)}} \text{traza}(\tilde{Y}_{ch}^t \tilde{Y}_{cK}). \quad (9)$$

2.2.2. Etapa 2: producto escalar entre objetos de distintos bloques

Definición 4 (Matriz de productos internos S). Se denotará mediante S a la matriz de orden $K \times K$ que contiene los productos internos de Frobenius entre los objetos dos a dos.

$$S = \begin{pmatrix} \langle W_1 | W_1 \rangle_F & \cdots & \langle W_1 | W_j \rangle_F & \cdots & \langle W_1 | W_k \rangle_F \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \langle W_i | W_1 \rangle_F & \cdots & \langle W_i | W_j \rangle_F & \cdots & \langle W_i | W_k \rangle_F \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \langle W_k | W_1 \rangle_F & \cdots & \langle W_k | W_j \rangle_F & \cdots & \langle W_k | W_k \rangle_F \end{pmatrix} \quad (10)$$

Sustituyendo las expresiones que definen a W_h y W_k , el término general queda:

$$S_{hk} = \frac{1}{\sqrt{VT(Y_k)VT(Y_h)}} \text{traza}(Y_h^t P_{X_h}^t P_{X_k} Y_k) = \frac{1}{\sqrt{VT(Y_h)VT(Y_k)}} \text{traza}(\tilde{Y}_{ch}^t \tilde{Y}_{ck}) \quad (11)$$

Donde P_{X_k} es la matriz de proyección sobre las columnas de la matriz X_k y así mismo $\tilde{Y}_h = (\tilde{Y}_{(1,h)}, \dots, \tilde{Y}_{(q,h)})$ e $\tilde{Y}_k = (\tilde{Y}_{(1,k)}, \dots, \tilde{Y}_{(q,k)})$ son las proyecciones de las columnas de Y_h y Y_k sobre los respectivos espacios columna X_h y X_k .

Un resultado muy importante es que sobre la diagonal de la matriz S se encuentran los índices de asociación de Goodman-Kruskal correspondientes a cada uno de los K pares de bloques:

$$S_{kk} = \frac{1}{\sqrt{VT(Y_k)VT(Y_k)}} \text{traza}(Y_k^t P_{(X_k)}^t P_{(X_k)} Y_k) = \tau_{y_k \cdot x_k}. \quad (12)$$

La construcción de la matriz S se hace para obtener por un lado un gráfico comparativo del nivel de asociación de los bloques apareados, y por otro lado ser una herramienta que facilita la reconstrucción de las distancias entre los objetos en el espacio euclídeo usual, con una interpretación de interés a los efectos del análisis.

La traza de la matriz S es en sí misma una medida de la asociación contenida entre los K bloques bajo estudio. **El elemento S_{hk} fuera de la diagonal principal de la matriz S , constituye una medida agregada de las covarianzas entre las estimaciones mínimo-cuadrática de las variables criterio en las ocasiones h y k .**

2.2.3. Etapa 3: distancia entre los objetos de distintos bloques

Definición 5 (Distancia entre objetos). Se define una distancia entre objetos basada en la norma dada de forma natural por el producto interno de Frobenius:

$$d_F^2(W_k, W_h) = \|W_k - W_h\|_F^2 \frac{1}{\sqrt{VT(Y_h)VT(Y_k)}} \langle W_k - W_h | W_k - W_h \rangle_F = \tau_{y_k \cdot x_k} + \tau_{Y_h \cdot X_h} - 2\langle W_h | W_k \rangle_F. \quad (13)$$

Donde el último término es una medida de la covarianza entre las estimaciones de y en ambos bloques. De modo que mientras mayor sea esta covariabilidad, menor será la distancia entre las representaciones de los bloques. Es decir, es posible cuantificar en una medida de distancia F , las diferencias existentes entre las estructuras definidas por las estimaciones mínimo-cuadrática del bloque de las q variables criterio en dos ocasiones diferentes que son obtenidas en función de los correspondientes bloques de variables explicativas.

En resumen, la medida de distancia propuesta para comparar los objetos de interés será mayor cuanto más grandes sean los respectivos índices de asociación de Goodman-Kruskal en las dos ocasiones consideradas y tanto más pequeña cuanto mayor sea la medida agregada de las covarianzas entre las estimaciones mínimo-cuadrática de la variable criterio en las ocasiones h y k .

2.2.4. Etapa 4: espacio de representación de la interestructura

La matriz que define las representaciones en el STATIS es la matriz S que contienen los productos escalares de Frobenius entre objetos de los distintos bloques.

La representación deseada se obtendrá al efectuar un análisis de componentes principales sobre la matriz de productos interno de interés, en este caso S , con la finalidad de construir un espacio de representación óptimo para los objetos W_k , determinado por las direcciones principales que definen los índices de asociación de Goodman-Kruskal:

$$\sum_{k=1}^K \frac{1}{VT(Y_h)} \text{traza} (Y_k^t (P_{X_k} - P_m)) Y_k \quad (14)$$

De esta manera el posicionamiento de un bloque apareado en ese espacio, determinado por su distancia al origen de coordenadas, deberá ser medida por una variación del índice de asociación de Goodman-Kruskal del bloque en cuestión

$$\|W_k\|_F^2 = \text{traza} (W_h^t W_k) = \frac{\text{traza} (Y_k^t X_k (X_k^t X_k)^{-1} X_k^t Y_k)}{VT(Y_h)} = \tau_{y_k x_k}. \quad (15)$$

Los resultados deseados se consiguen al efectuar la descomposición espectral de S , descrita mediante la expresión $S = G T G^t$, siendo $G = (G_1, \dots, G_K)$ la matriz cuyas columnas son los autovectores de la matriz S asociados con los correspondientes autovalores ordenados de forma descendente y organizados sobre la diagonal de la matriz

$$\begin{pmatrix} t_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & t_k \end{pmatrix} \quad (16)$$

Un resultado fundamental que se deriva de esta factorización, es la posibilidad de descomponer índice de asociación de Goodman-Kruskal

$$\sum_{k=1}^K \tau_{y_k \cdot x_k} = \sum_{k=1}^K t_k. \quad (17)$$

Es decir, al diagonalizar la matriz $Y_k^t (P_{X_k} - P_m) Y_k$ se obtiene que la traza de la matriz S , es $\sum_{k=1}^K \tau_{y_k \cdot x_k}$, donde cada autovalor está asociado con una dirección principal del análisis de correspondencias no simétrico (ACNS).

La asociación de los objetos capturada por el k -ésimo eje se mide mediante la relación

$$\frac{t_k}{\sum_{k=1}^K \tau_{y_k \cdot x_k}} \quad (18)$$

En este orden de ideas, las coordenadas de proyección γ^{α_k} , de los bloques sobre el eje α^{α_k} quedan de la forma

$$\gamma^{\alpha_k} = \sqrt{\lambda_{\alpha_k}} a^{\alpha_k}, \quad \alpha_k, k = 1, \dots, K, \quad (19)$$

En la representación gráfica se tendrá que en los bloques más alejados del origen, la capacidad de \mathbf{x}_k para explicar \mathbf{y}_k será mayor (mayor será el coeficiente $\tau_{\mathbf{y}_k \cdot \mathbf{x}_k}$).

2.2.5. Estrategia denominada ICI

En el caso de esta investigación en particular se pretende comparar la variabilidad de las variables x_K en función de las matrices y_k , $k = 1, \dots, K$, en términos del coeficiente $\tau_{y_K \cdot x_k}$ de Goodman - Kruskal. Para ello se procederá a construir un marco semejante al de la metodología STATIS, que permita extender el análisis a objetos que contengan toda la información necesaria para reconstruir el coeficiente $\tau_{Y_K \cdot X_k}$ de Goodman - Kruskal en cada bloque no simétrico (x_k, y_k) , $k = 1, \dots, K$. Así las cosas, se propone un símil de la estrategia denominada ICI (Interestructura - Compromiso - Intraestructura) para este caso en particular. Esto quiere decir que se lleva a cabo un procedimiento de tres fases que responde a los objetivos de:

1. Identificar los bloques de información (x_k, y_k) , $k = 1, \dots, K$, que son similares (interestructura).
2. Construir un marco de representación común a todos los bloques de información (x_k, y_K) .
3. Describir las diferencias o similitudes entre los individuos sobre un espacio compromiso (intraestructura). Utilizando el compromiso como pieza fundamental se va a analizar el comportamiento de los individuos y los bloques que de alguna manera expliquen las razones o causas de las semejanzas y/o diferencias detectadas en el análisis de la interestructura.

Cabe destacar, que un objeto definido de acuerdo a los principios de la metodología STATIS, sería una matriz cuadrada simétrica, con características que resultan beneficiosas en algunos aspectos. Sin embargo, los objetos definidos en esta investigación resultan ser matrices rectangulares no simétricas. Sobre la base del producto interno de Frobenius, y con medidas de asociación y correlación matricial establecidas bajo los principios que orientan este concepto, se construyen matrices de similaridad entre objetos que aceptan representaciones gráficas sobre espacios de baja dimensión, las cuales permiten apreciar gráficamente el parecido y/o diferencia en relación con el $\tau_{\mathbf{y}_k \cdot \mathbf{x}_k}$ de Goodman-Kruskal, entre los distintos bloques no simétricos.

3. Aplicación del STATIS-CGK

Según publica la Superintendencia Financiera de Colombia en el capítulo 2 de la circular externa básica contable y financiera 100 de 1995 ² el riesgo de crédito (RC) es la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que un deudor o contraparte incumpla sus obligaciones.

La Superintendencia indica además que las entidades vigiladas deben evaluar permanentemente el riesgo incorporado en sus activos crediticios, tanto en el momento de otorgar créditos como a lo largo de la vida de los mismos, incluidos los casos de reestructuraciones. Para tal efecto, las entidades deben diseñar y adoptar un Sistema de Administración del Riesgo Crediticio (SARC). Los elementos básicos que deben componer el SARC son: políticas de administración del RC, procesos de administración del RC, modelos internos o de referencia para la estimación o cuantificación de pérdidas esperadas, sistema de provisiones para cubrir el RC, procesos de control interno.

Por lo anterior, las entidades financieras deben establecer esquemas eficientes de administración y control del riesgo de crédito al que se exponen en el desarrollo del negocio, en resonancia a su propio perfil de riesgo, segmentación de mercado, según las características de los mercados en los que opera y de los productos que ofrece; por lo tanto es necesario que cada entidad desarrolle su propio esquema de trabajo, que asegure la calidad de sus activos y además permita identificar, medir, controlar (mitigar) y monitorear la materialización de los diferentes riesgos a las que están expuestas como bancos.

²<https://www.superfinanciera.gov.co/publicacion>

La implementación de técnicas estadísticas para la identificación, medición, control (mitigación) y monitoreo del riesgo en las entidades financieras se hace cada vez más necesario para tomar mejores decisiones. Dicha implementación implica el tener que analizar tablas de datos.

Uno de los objetivos del análisis de las tablas de datos tradicionales es caracterizar los clientes, destacando sus similitudes y diferencias. Es posible considerar diferentes generalizaciones de estas matrices, al tomar en cuenta la oportunidad en que son observadas las variables y en el rol que éstas asumen. Las posibilidades pueden ampliarse aún más si además se diferencian dos casos según las variables sean cualitativas o cuantitativas.

Uno de los métodos propuestos para el análisis simultáneo de tablas múltiples de datos en varias ocasiones es el STATIS, cuya finalidad es explorar las similitudes entre unas estructuras denominadas objetos, que resumen la información de los individuos.

En este trabajo se presenta una adaptación de la técnica propuesta, sobre un conjunto de datos reales, conformado por 8 bloques de datos apareados donde cada bloque contiene la medición de dos variables cualitativas sobre 786 clientes de una entidad financiera. El objetivo de la aplicación es presentar un análisis que permita determinar si la información de las centrales de riesgo colombianas (medida a través de un estado de riesgo externo) tiene alguna relación con la clasificación de riesgo crediticio de una determinada entidad financiera.

3.1. Descripción del problema

En esta sección se ilustra la técnica propuesta, aplicada sobre un conjunto de datos reales, conformado por $K=8$ bloques de datos apareados (x_k, y_k) , $k = 1, \dots, 8$, donde cada bloque contiene la medición de dos variables cualitativas sobre 786 individuos. En este análisis se dispone de dos variables cualitativas:

1. Una variable x que corresponde a la calificación de riesgo (a nivel de cliente) ³ de cada cliente con 5 modalidades: A, B, C, D y E.
2. Una variable y que corresponde a un estado externo de riesgo (a nivel de cliente) que establece la entidad con información externa. Esta variable tiene 4 modalidades o categorías: Estado 1, Estado 2, Estado 3 y Estado 4.

se asume que la variable cualitativa x explica a la variable cualitativa y . Estos 8 bloques están constituidos por un conjunto de variables cualitativas de interés para una entidad financieras. Adicionalmente, cada bloque:

1. Contiene la medición de dos variables cualitativas sobre 786 individuos.
2. Representa un trimestre entre los años 2016-2017.

A lo largo de este documento, se detallará los bloques de datos explicando su estructura lógica, explorando en las siguientes secciones los bloques (x_k, y_k) , $k = 1, \dots, 8$, desde el punto de vista de la metodología STATIS-CGK.

3.2. Variables analizadas

1. **Calificación del riesgo crediticio.** Según la norma de la SuperFinanciera de Colombia ⁴ los contratos deben clasificarse en una de las siguientes categorías de riesgo crediticio:
 - Categoría A o “riesgo normal”.
 - Categoría B o “riesgo aceptable, superior al normal”.

³ Variable supeditada por el capítulo 2 de la Superfinanciera.

⁴<https://www.superfinanciera.gov.co/publicacion>

- Categoría C o “riesgo apreciable”.
- Categoría D o “riesgo significativo”.
- Categoría E o “riesgo de incobrabilidad”.

A continuación detallaremos esta segmentación de la normativa Colombiana.

- Categoría “A”: Crédito con riesgo crediticio NORMAL. Los créditos calificados en esta categoría reflejan una estructuración y atención apropiadas.
- Categoría “B”: Crédito con riesgo ACEPTABLE. Los créditos calificados en esta categoría están aceptablemente atendidos y protegidos, pero existen debilidades que potencialmente pueden afectar, transitoria o permanentemente, la capacidad de pago del deudor.

Tabla 1: Condiciones objetivas suficientes para la obtención de la calificación crediticia

		No. de meses en mora (rango)			
MODALIDAD CREDITO	DE	Categoría B	Categoría C	Categoría D	Categoría E
Vivienda		Más de 2 hasta 5	Más de 5 hasta 12	Más de 12 hasta 18	Más de 18
Consumo		Más de 1 hasta 2	Más de 2 hasta 3	Más de 3 hasta 6	Más de 6
Microcrédito		Más de 1 hasta 2	Más de 2 hasta 3	Más de 3 hasta 4	Más de 4
Comercial		Más de 1 hasta 3	Más de 3 hasta 6	Más de 6 hasta 12	Más de 12

- Categoría “C”: Crédito deficiente, con riesgo APRECIABLE. Se califican en esta categoría los créditos o contratos que presentan insuficiencias en la capacidad de pago del deudor.
- Categoría “D”: Crédito de difícil cobro, con riesgo SIGNIFICATIVO. Es aquél que tiene cualquiera de las características del deficiente, pero en mayor grado, de tal suerte que la probabilidad de recaudo es altamente dudosa.
- Categoría “E”: Crédito IRRECUPERABLE. Es aquél que se estima incobrable.

Nota: la tabla 1 resume las condiciones objetivas suficientes para que un crédito sea clasificado en las categorías anteriormente descritas.

2. **Calificación del riesgo a nivel de cliente.** Aun cuando la calificación de riesgo presentada es a nivel de crédito. Para obtener una calificación de riesgo a nivel de cliente se coloca la calificación del crédito con mayor exposición. Es decir, el análisis presentado en esta trabajo se lleva a cabo a nivel de cliente. Para fines prácticos, a lo largo de este documento el término calificación de riesgo a nivel de cliente hace referencia a la calificación crediticia de mayor exposición.

3. **Estado externo de riesgo (cliente).** Segmentación que realiza la entidad financiera para realizar un seguimiento a todos los clientes compartidos ⁵, a partir del comportamiento de pago que estos presenten, tanto dentro de la entidad como en el sector financiero (esta variable incorpora datos de la central de riesgo ⁶). ⁷

Este estado es construido por la entidad de manera trimestral para todos los clientes que tienen productos de crédito vigentes con la entidad y que presentan deudas con el sector. Los estados según el comportamiento presentado son:

⁵Los clientes compartidos son aquellos que mantienen una operación vigente con la entidad y que, adicionalmente, presentan al menos una obligación con otra entidad del sector financiero.

⁶La central de riesgo es una empresa privada, independiente de las instituciones financieras, de las comerciales y de las gubernamentales, que tiene como fin concentrar y proporcionar a sus empresas afiliadas, la información referente al comportamiento que han tenido las personas en sus créditos

⁷Las entidades financieras utilizan los servicios de las centrales de riesgo porque esta información proporciona una innovadora herramienta de apoyo a la toma de decisiones en la evaluación, prevención del riesgo crediticio y gestión de clientes. Estos servicios, aunque no son gratuitos, le permiten el acceso de manera sencilla a la más actualizada y completa base de datos de información de incumplimientos.

- a) Estado 1. Constituido por aquellos clientes que al ser consultado en las centrales de riesgo en el trimestre cumple sus obligaciones financieras con el sector y cumple con la entidad.
- b) Estado 2. Constituido por aquellos clientes que al ser consultado en las centrales de riesgo en el trimestre incumple sus obligaciones financieras con el sector y cumple con la entidad.
- c) Estado 3. Constituido por aquellos clientes que al ser consultado en las centrales de riesgo en el trimestre cumple sus obligaciones financieras con el sector e incumple con la entidad.
- d) Estado 4. Constituido por aquellos clientes que al ser consultado en las centrales de riesgo en el trimestre incumple sus obligaciones financieras con el sector e incumple con la entidad.

La tabla 2 resume la información anteriormente descrita.

Tabla 2: Resumen Estado externo de riesgo (cliente)

Comportamiento		
Estado	En el Sector financiero sin incluir la entidad	Con la entidad
Estado 1	Cumple	Cumple
Estado 2	Incumple	Cumple
Estado 3	Cumple	Incumple
Estado 4	Incumple	Incumple

3.3. Descripción de la muestra analizada

La muestra analizada en este capítulo forma parte de un estudio trimestral que realiza una entidad financiera. Para fines académicos se trabajará con $n=786$ clientes que fueron observados trimestralmente durante 24 meses ($K=8$). Para fines prácticos utilizaremos la notación $2016Ti$ $i = 1, \dots, 4$, indica el i -ésimo trimestre del año 2016⁸, de igual manera se utiliza la notación $2017Ti$ $i = 1, \dots, 4$. Los $n=786$ clientes fueron evaluados en cada trimestre ($K=8$) en cada una de las variables mencionadas (estado externo de riesgo y calificación de riesgo). La muestra analizada en este capítulo forma parte de un análisis trimestral que realiza una entidad financiera, donde se evalúa información del sector financiero y permite obtener una medición global de los clientes compartidos. La aplicación de esta investigación se centrará en 786 individuos que serán analizados en los 8 trimestres (2016-2017).

3.4. Análisis univariado de los datos

Todos los bloques contienen información a nivel de cliente de dos variables: estado externo de riesgo y la calificación de riesgo. En la tabla 3 presentamos los resultados del estado externo de riesgo de 786 clientes a lo largo de estos 8 trimestres.

Tabla 3: Comportamiento de la variable estado de riesgo cliente a lo largo del período de tiempo bajo estudio-Centro de gravedad de la nube de perfiles columna (en $R^q, q = 4$).

	2016T1		2016T2		2016T3		2016T4		2017T1		2017T2		2017T3		2017T4	
	Clientes		Clientes		Clientes		Clientes		Clientes		Clientes		Clientes		Clientes	
Estado 1	751	96 %	748	95 %	722	92 %	643	82 %	451	57 %	318	40 %	215	27 %	69	9 %
Estado 2	3	0 %	3	0 %	17	2 %	3	0 %	81	10 %	4	1 %	5	1 %	9	1 %
Estado 3	2	0 %	4	1 %	15	2 %	56	7 %	70	9 %	76	10 %	84	11 %	116	15 %
Estado 4	30	4 %	31	4 %	32	4 %	84	11 %	184	23 %	388	49 %	482	61 %	592	75 %
Total	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %

⁸Por ejemplo

1. 2016T1: Se refiere al primer trimestre de 2016, correspondiente a los meses enero, febrero y marzo de 2016.
2. 2016T4: Se refiere al cuarto trimestre de 2016, correspondiente a los meses octubre, noviembre y diciembre de 2016.

Observe que a medida que se avanza en el tiempo los clientes pasan de estados de riesgo bajos a estados de riesgo altos (ver figura (1)), esto es medido a través del vector de centro de gravedad de esta variable. De manera análoga, la tabla 4 presenta la calificación de riesgo de estos 786 clientes. A medida que se

Tabla 4: Comportamiento de la variable calificación de riesgo cliente a lo largo del período de tiempo bajo estudio -Centro de gravedad de la nube de perfiles fila (en R^p , $p = 5$).

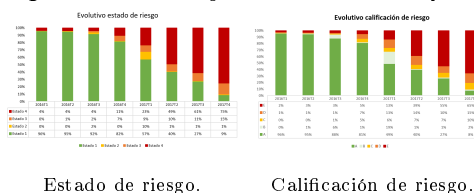
	2016T1		2016T2		2016T3		2016T4		2017T1		2017T2		2017T3		2017T4	
	Clientes		Clientes		Clientes		Clientes		Clientes		Clientes		Clientes		Clientes	
A	751	96 %	744	95 %	690	88 %	636	81 %	385	49 %	312	40 %	209	27 %	61	8 %
B	3	0 %	7	1 %	49	6 %	10	1 %	147	19 %	9	1 %	10	1 %	14	2 %
C	2	0 %	2	0 %	10	1 %	42	5 %	46	6 %	53	7 %	58	7 %	78	10 %
D	11	1 %	10	1 %	11	1 %	58	7 %	103	13 %	107	14 %	79	10 %	119	15 %
E	19	2 %	23	3 %	26	3 %	40	5 %	105	13 %	305	39 %	430	55 %	514	65 %
Total	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %	786	100 %

avanza en el tiempo los clientes pasan de calificación de riesgo baja a calificación de riesgo alta (ver figura 1), esto es medido a través del vector de centro de gravedad de esta variable. La figura 1 representa la

Tabla 5: Matriz de productos internos S

	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
W_1	1.00000013	0.86631278	0.64426035	0.41195924	0.23599688	0.15641101	0.11862325	0.06184779
W_2	0.86631278	0.90908378	0.65534145	0.41423125	0.23307192	0.15391475	0.11463736	0.05820963
W_3	0.64426035	0.65534145	0.77109895	0.40219283	0.21790601	0.13848237	0.09718957	0.02865672
W_4	0.41195924	0.41423125	0.40219283	0.897001	0.41604316	0.26061323	0.17725116	0.05812477
W_5	0.23599688	0.23307192	0.21790601	0.41604316	0.76669806	0.44321497	0.31315822	0.1171675
W_6	0.15641101	0.15391475	0.13848237	0.26061323	0.44321497	0.90577747	0.62099695	0.32452613
W_7	0.11862325	0.11463736	0.09718957	0.17725116	0.31315822	0.62099695	0.89988276	0.44351571
W_8	0.06184779	0.05820963	0.02865672	0.05812477	0.1171675	0.32452613	0.44351571	0.7905103

Figura 1: Centro de gravedad de la nube de puntos.



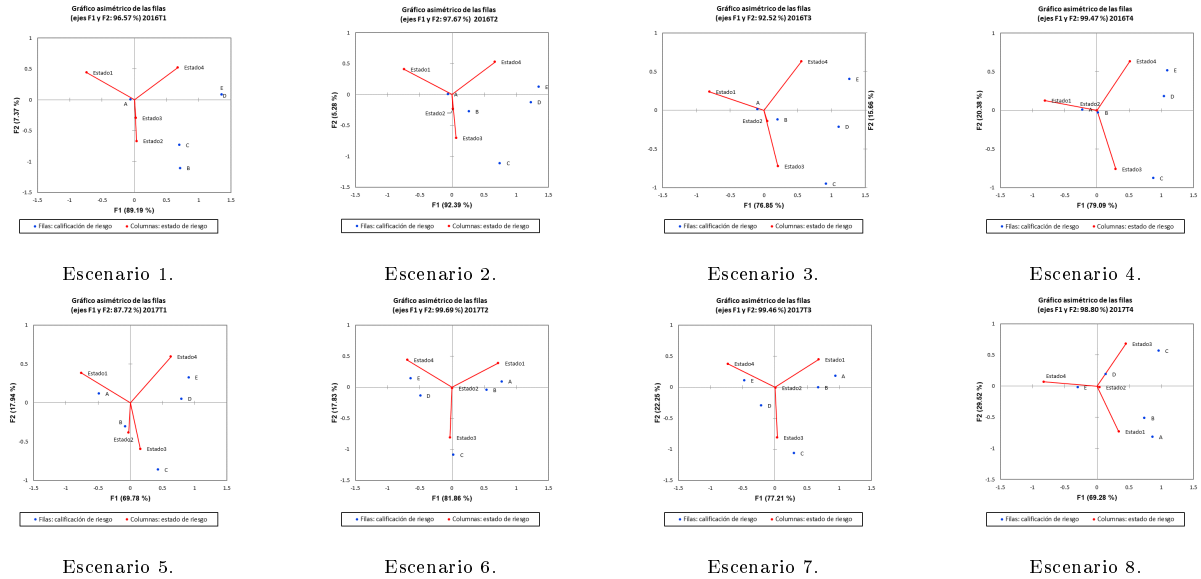
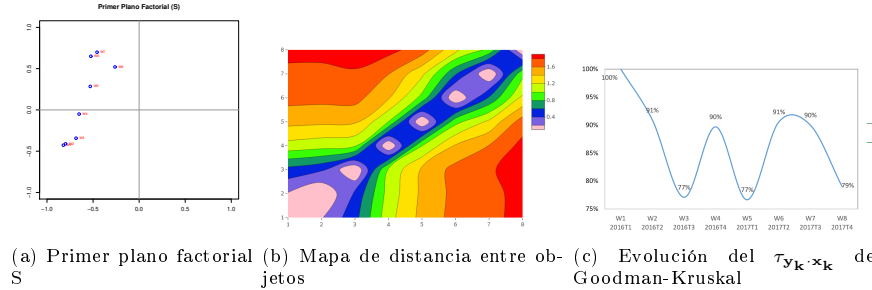
Estado de riesgo.

Calificación de riesgo.

evolución de los centros de gravedad de las variables bajo estudio y nos permite indicar que en el periodo de tiempo analizado (2016-2017) se presenta un deterioro en el comportamiento de pago de los clientes de la entidad analizada.

4. Resultados obtenidos

En el gráfico (3), se puede observar que el primer eje factorial, divide a los objetos en dos grupos diferenciales: un primer grupo constituido por los objetos W_1 , W_2 , W_3 y W_4 y un segundo grupo constituido por el resto de los elementos. Observe en el gráfico 3((a) y (b)) que W_8 está más alejando de los objetos W_1 , W_2 y W_3 . Al realizar un ACNS, obtenemos que el primer grupo se caracteriza por objetos correspondientes a los 4 trimestres del 2016, presentando participación de categoría A - estado 1 superior al 80 % y categoría E - estado 4 inferior al 12 % (ver gráfico (1) y tablas (4) y (3)) donde se presenta la información de los centros de gravedad de la nube de puntos en los espacios R^p y R^q). El segundo grupo se caracteriza por estar constituido por los trimestres del 2017 que presentan participación de la tupla categoría A - estado 1 inferior al 60 % y en general categoría E - estado 4 con participación superior al 20 %.

Figura 2: Análisis de correspondencias no simétrico (ACNS) de los bloques apareados de información $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, \dots, K = 8$.**Figura 3:** Gráficos para el análisis

Es natural que surja la pregunta: ¿Qué pudo haber sucedido para que se produzcan migraciones de calificaciones y estados de riesgo tan pronunciadas para este conjunto de clientes? para dar una respuesta en primer lugar analizaremos las variaciones porcentuales anuales del PIB (ver tabla (6)), según cifras del DANE⁹ indican que en el primer trimestre del año 2016 se produjo un crecimiento de la economía, que se ubicó en 2,5 %, inferior a las expectativas de los analistas, que señalaban 2,8 % (Bloomberg) y 3 % (Bancolombia). Sin embargo, la variación porcentual anual del PIB en el 2015 fue de 3.05, en el 2016 de 2.04 y en el 2017 de 1.77. En este orden desde el punto de vista macro económico se evidencia una contracción de la economía. La segunda parte de la respuesta, esta relacionado con el comportamiento migratorio de los centros de gravedad (ver gráfico (1), y las tablas (3) y (4)).

Así las cosas, lo anteriormente comentado se evidencia en el comportamiento del τ_{y_k, x_k} de Goodman-Kruskal (ver tabla (6) y gráfico (3)(c).) Como puede observarse los escenarios W_1, W_2, W_4, W_6 y W_7 presentan alta capacidad de explicar la y (estado de riesgo) en términos de la x (calificación de riesgo), mientras que el resto de los escenarios presentan una mediana capacidad explicativa de la y en función de las x . Constituyendo el 2016, el año que presenta más ocasiones con alta capacidad explicativa de las y en función de las x . Finalmente pero no menos importante, a partir de tabla (6) y el gráfico (3)(c) queda en evidencia que el mínimo valor τ_{y_k, x_k} se alcanzó en el primer trimestre de 2017; denotando una mediana capacidad de explicar la y (estado de riesgo) en términos de la x (calificación de riesgo) medido a través de un τ_{y_k, x_k} de Goodman de 76.7 %. Esto se debe principalmente a la propuesta de la reforma tributaria de la Ley 1819 del 2016 que propone cárcel para los evasores de impuestos que oculten

⁹Departamento Administrativo Nacional de Estadística.

Tabla 6: Evolución $\tau_{\mathbf{y}_k \cdot \mathbf{x}_k}$, $k = 1, \dots, K = 8$.

Escenario	Variabilidad total	Variabilidad explicada	Tau	Variaciones porcentuales anuales del PIB (DANE)		
W_1	67.2798982	67.28107	100.00000 %	2016	I	2.49
W_2	72.9083969	66.2820272	90.91138 %		II	2.36
W_3	120.832061	93.1756387	77.11169 %		III	1.50
W_4	247.005089	221.5649	89.70054 %		IV	1.84
W_5	469.564885	360.01428	76.66976 %	2017	I	1.55
W_6	458.442748	415.247106	90.57775 %		II	1.69
W_7	422.603053	380.293028	89.98823 %		III	2.26
W_8	316.83715	250.464085	79.05136 %		IV	1.58

o inventen información a partir de \$100 millones por entre 48 y 108 meses. Además, quienes evadan más de \$5.000 millones pagarán entre 4 y 9 años de cárcel. Los evasores de menos de \$100 millones recibirán multa del 20 %. Esto implica que existieron muchos pequeños comerciantes informales que tienen ingresos entre \$40 millones y \$100 millones, que empezaron pagar renta y por ende se produjo un impacto en la capacidad de pago de los clientes.

5. Conclusiones

1. Se realizó una adaptación del método STATIS a un escenario bloques apareados no simétricos $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, 2, \dots, K$. La adaptación desarrollada se aplica a un conjunto de datos, mediante la descomposición del estadístico τ de Goodman-Kruskal, permitiendo obtener estimaciones mínimo cuadrática de \mathbf{y} , obtenidas al proyectarla ortogonalmente sobre el espacio generado por las categorías de \mathbf{x} .
2. La conclusión de la aplicación es que efectivamente la información de las centrales de riesgo (medido através de la variable estado de riesgo) presenta una relación con la clasificación de riesgo de la entidad. Dado que el proceso trimestral de consulta a centrales de riesgo es un proceso que representa costo para la entidad, se recomienda utilizar la segmentación presentada en la tabla (7).

Tabla 7: Estado de riesgo recomendado.

Calificación de riesgo	Estado recomendado
A	Estado 1
B	Estado 2
C	Estado 3
D	Estado 4
E	Estado 4

Referencias

- [1] Lauro, N. (1984), "L'analyse non symetrique des correspondances". : Data Analysis and Informatics III. Diday, E. Coll. (Eds.) Elsevier, North-Holland. 433-446
- [2] Lavit, C. (1985), Application de la méthode statis. Statistique et Analyse des Données, Vol 10, n.1.