

STATIS-CNS2: an exploratory analysis technique for qualitative non-symmetric data

Jennyfer Combariza * , Guillermo Ramírez , Maura Vasquez 

Área de Postgrado en Estadística y Actuariado de la Facultad de Ciencias Económicas y Sociales de la Universidad Central de Venezuela. Caracas, Venezuela.



<https://doi.org/10.54139/revinguc.v29i2.174>

Abstract.- The problem raised in this research focuses on the search for a methodology, based on STATIS, that allows comparing and simultaneously explaining the influence of an explanatory qualitative variable x as determinant of a criterion categorical variable y on H occasions. To this end, a Frobenius scalar product is defined between the objects, which allows us to conceptualize a statistical distance between objects, based on the Goodman-Kruskal statistic τ .

An application of the proposed technique on a set of real data from the financial field is also presented made up of 4 blocks of paired data, where each block contains the measurement of two qualitative variables on a sample of 20316 individuals, with the purpose of measuring the impact it has on the information emanating from the Colombian credit bureaus the segmentation used for the calculation of provisions both in local accounting and accounting under IFRS9 regulations.

Keywords: STATIS; τ Goodman-Kruskal; non-symmetric correspondence analysis.

STATIS-CNS2: una técnica de análisis exploratorio para datos no simétricos cualitativos

Resumen.- El problema planteado en esta investigación se centra en la búsqueda de una metodología, basada en el STATIS, que permita comparar y explicar simultáneamente la influencia que tiene una variable cualitativa explicativa x como determinante de una variable categórica criterio y en H ocasiones. Con este fin se define un producto escalar tipo Frobenius entre los objetos, que permite conceptualizar una distancia estadística entre objetos, en función del estadístico τ de Goodman-Kruskal.

Se presenta además una aplicación de la técnica propuesta sobre un conjunto de datos reales del ámbito financiero conformado por 4 bloques de datos apareados, donde cada bloque contiene la medición de dos variables cualitativas sobre una muestra de 20316 individuos, con el propósito de medir el impacto que tiene sobre la información emanada de las centrales de riesgo colombianas la segmentación utilizada para el cálculo de provisiones tanto en la contabilidad local como contabilidad bajo normativa NIIF9.

Palabras clave: STATIS; τ de Goodman-Kruskal; análisis de correspondencias no simétrico.

Received: 01 de julio, 2022.

Accepted: 03 de agosto, 2022.

1. Introducción

En el mundo de la ciencia de datos, saber qué datos tenemos y cómo los podemos utilizar es un tema determinante para la solución de problemas no sólo en el ámbito académico. Por esta razón el disponer de diversas herramientas de análisis exploratorio es clave. La técnica que vamos a

presentar en este artículo es una herramienta de análisis exploratorio que permite representar un conjunto particular de datos cualitativos no simétricos longitudinales en un único espacio de representación.

D'Ambra y Lauro [1] han propuesto el análisis de correspondencias no simétrico (ACNS) para obtener un espacio de representación de datos cualitativos no simétricos. Esta técnica, basada en la descomposición del estadístico τ de Goodman-Kruskal, plantea obtener estimaciones mínimo cuadráticas de una variable cualitativa y , obtenidas al proyectarla ortogonalmente sobre el espacio generado por las modalidades de una variable x . En

* Autor para correspondencia:

Correo-e: jennyfer.combariza@gmail.com (J. Combariza)

el caso que se disponga la medición de estas dos variables en diferentes instantes de tiempo surge la necesidad de analizar varios conjuntos de datos y realizar un estudio comparativo de sus estructuras, lo que nos lleva al problema de cómo trasladar los resultados de los distintos ACNS a un único espacio de representación.

La conocida metodología STATIS (Structuration des Tableaux A Trois Indices de la Statistique También conocido como ACT. ACT: Analyse Conjointe de Tableaux) fue inicialmente diseñada para el tratamiento de información generada al caracterizar n individuos según p variables continuas en H ocasiones diferentes y simétricas.

La propuesta que se ofrece STATIS-CNS2 (Las siglas CNS2 significa que se está estudiando un escenario cualitativo, no simétrico con asociación medida a través del τ de Goodman-Kruskal y el dos significa que son analizadas sólo dos variables: una variable criterio y como función de una variable independiente x) tiene como propósito fundamental exhibir un espacio de representación único que permita analizar los cambios que se producen en la influencia que tiene una variable explicativa categórica x sobre una variable categórica a explicar y , a lo largo de H ocasiones.

La organización de este artículo es la siguiente:

- El objetivo de la sección 1.1 de preliminares, es presentar la visión estratégica del STATIS-CNS2. Para esto se debe partir de la visión estratégica del STATIS clásico, seguidamente se enseña la configuración de los datos que se requieren para el análisis de correspondencias no simétrico (ACNS), se muestra un resumen de las etapas del ACNS de los profesores D'Ambra y Lauro [1], se describe el τ de Goodman-Kruskal y finalmente se presenta la visión estratégica de la investigación.
- En la sección 2 se presenta todo el fundamento teórico para la comprensión de la Metodología STATIS-CNS2.
- En la sección 3 se expone una aplicación del STATIS-CNS2 sobre una muestra de datos reales.

1.1. Preliminares

En esta sección presentamos los conceptos básicos requeridos para poder entender la adaptación desarrollada: uno de ellos es la configuración de datos, es importante que el lector tenga claro qué características deben tener los datos para poder ser analizados con STATIS-CNS2. Otro concepto importante es entender las etapas a seguir para realizar el análisis de correspondencias no simétrico (ACNS) así como el uso del τ de Goodman-Kruskal para evaluar la intensidad de la fuerza con que una variable cualitativa explica a otra. Finalmente pero no menos importante presentamos la propuesta de la estrategia ICI (Interestructura - Compromiso - Intraestructura) para este caso particular.

1.1.1. STATIS: estrategia ICI

La metodología del análisis multivariante denominada STATIS inicialmente introducida por Escoufier [2] se sustenta en las propiedades de operadores asociados a H diferentes matrices de datos cuantitativos en un escenario simétrico. El STATIS es un procedimiento de análisis en etapas, mediante una estrategia denominada ICI (interestructura - compromiso - intraestructura). 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 que son similares (**interestructura**).
2. Proveer un bloque resumen (**compromiso**).
3. Describir las diferencias o similitudes entre los individuos sobre un espacio compromiso (**intraestructura**).

En la Figura 2 se presenta un resumen de la estrategia ICI para el STATIS.

1.1.2. Configuración de los datos en el ACNS

En el ACNS (análisis de correspondencias no simétrico) se consideran dos variables cualitativas x e y :

1. Una variable explicativa x con p modalidades $A_1, \dots, A_i, \dots, A_p$.
2. Una variable explicada y con q modalidades $B_1, \dots, B_j, \dots, B_q$.

Así mismo, se considera $(\mathbf{x}_h, \mathbf{y}_h)$, $h = 1, \dots, H$, bloques apareados no simétricos, que contienen la medición de dos variables cualitativas sobre n individuos.

Las variables cualitativas son observadas sobre los mismos n individuos en H ocasiones. En el caso de diseños longitudinales se utilizará el término ocasión para identificar cada uno de los h posibles instantes de medición, en tanto que los términos situación y condición se reservarán para los casos en que h no esté vinculado al paso del tiempo sino a clasificaciones conceptuales propias a los significados que tengan las variables que se estén analizando. Para fines de desarrollo teórico utilizamos el término ocasión.

Para cada ocasión h , los valores observados de las variables x e y sobre los n individuos pueden organizarse sobre matrices disyuntivas completas de la forma:

$$X_h = (x_{sih})_{1 \leq s \leq n, 1 \leq i \leq p} \quad Y_h = (y_{sjh})_{1 \leq s \leq n, 1 \leq j \leq q}$$

siendo, según las ecuaciones (1), (2):

$$x_{sih} = \begin{cases} 1 & \text{si el individuo } s \text{ presenta la} \\ & \text{i-ésima modalidad de la} \\ & \text{variable } x \text{ en la} \\ & \text{h-ésima ocasión} \\ 0 & \text{en otro caso.} \end{cases} \quad (1)$$

$$y_{sjh} = \begin{cases} 1 & \text{si el individuo } s \text{ presenta la} \\ & \text{j-ésima modalidad de la} \\ & \text{variable } y \text{ en la} \\ & \text{h-ésima ocasión} \\ 0 & \text{en otro caso.} \end{cases} \quad (2)$$

A lo largo de esta investigación indicar que el individuo s presenta la i -ésima modalidad de la variable x en la h -ésima ocasión es equivalente a indicar que el individuo s presenta la modalidad A_i de la variable x en la h -ésima ocasión. Lo anterior vale también para las modalidades de la variable y en la h -ésima ocasión, por simplicidad utilizaremos j -ésima modalidad para indicar la modalidad B_j en la h -ésima ocasión. En la Figura 1 $k_{.j}$ es el número de individuos que presentan la modalidad B_j de la variable y y satisfacen la condición $\sum_{j=1}^q k_{.j} = n$ y $k_{.j}$ es el número de individuos que presentan la modalidad B_j de la variable y y satisfacen la condición $\sum_{j=1}^q k_{.j} = n$.

1.1.3. Índice de Goodman-Kruskal

Un aspecto central en esta investigación tiene que ver con analizar la influencia que la variable categórica x (con p modalidades) ejerce sobre la variable categórica y (q modalidades), y estudiar los cambios que pudieran producirse a lo largo de varias ocasiones.

En esta sección presentaremos un índice estadístico τ de Goodman-Kruskal [3] que nos permite obtener una medida de la intensidad de la fuerza con que la variable x explica a la y . El mismo, tiene la siguiente representación, de acuerdo con la ecuación (3) :

$$\tau_{y \cdot x} = \frac{\sum_{j=1}^q \sum_{i=1}^p \frac{f_{ij}^2}{f_{i.}} - \sum_{j=1}^q f_{.j}^2}{1 - \sum_{j=1}^q f_{.j}^2}, \quad (3)$$

$$\text{donde } f_{.j} = \frac{k_{.j}}{n}, \quad f_{i.} = \frac{k_{i.}}{n}, \quad f_{ij} = \frac{k_{ij}}{n}.$$

1.1.4. Análisis interno de cada bloque con un enfoque ACNS

El tratamiento analítico de la información dentro del bloque h , plantea efectuar un ACNS [1], lo que implica analizar la variabilidad de y_h explicada por x_h , mediante la descomposición del estadístico $\tau_{y_h \cdot x_h}$ de Goodman-Kruskal, de acuerdo con las siguientes etapas:

Etapas 1. Para cada ocasión h , se obtienen las estimaciones mínimo cuadráticas de y_h , sobre el espacio generado por las modalidades de la variable x_h , expresada en la ecuación (4):

$$\tilde{Y}_h = X_h(X_h^t X_h)^{-1} X_h^t Y_h. \quad (4)$$

En este arreglo, la estimación mínimo cuadrática que corresponde a la variable y_h en un individuo en el que está presente la modalidad i de la variable x_h queda definida por el vector de probabilidades condicionales estimadas:

$$(\tilde{P}(y_h = 1 | x_h = i) \cdots \tilde{P}(y_h = j | x_h = i) \cdots \tilde{P}(y_h = q | x_h = i))$$

El centro de gravedad de las filas de $\tilde{Y}_h$ es el vector:

$$(\tilde{P}(y_h = 1) \cdots \tilde{P}(y_h = j) \cdots \tilde{P}(y_h = q))$$

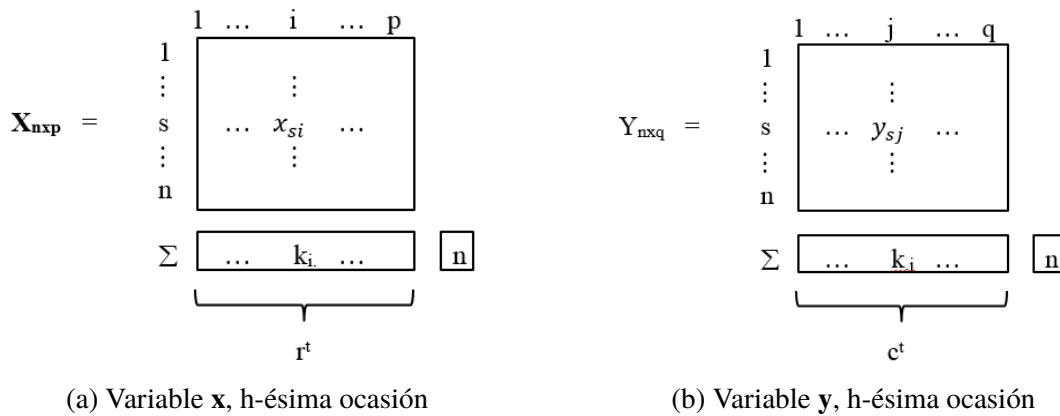


Figura 1: Matrices disyuntivas h -ésima ocasión

La matriz de estimaciones centrada $\tilde{\mathbf{Y}}_{ch}$, expresada en la ecuación (5):

$$\tilde{\mathbf{Y}}_{ch} = (\mathbf{P}_{X_h} - \mathbf{P}_m)\mathbf{Y}_h \quad (5)$$

tiene como fila genérica:

$$\begin{aligned} & (\tilde{P}(y_h = 1 | x_h = i) - P(y_h = 1)) \\ & \dots \tilde{P}(y_h = j | x_h = i) - P(y_h = j) \\ & \dots \tilde{P}(y_h = q | x_h = i) - P(y_h = q) \end{aligned}$$

Etapas 2. Se construye el cociente entre la variabilidad de las estimaciones mínimo cuadráticas (variabilidad explicada) y la variabilidad total de la variable y_h , encontrándose que este cociente coincide con el estadístico de Goodman-Kruskal:

1. Variabilidad total $VT(h)$: según la ecuación (6)

$$VT(h) = \text{traza}(\mathbf{Y}_{ch}^t \mathbf{Y}_{ch}) \quad (6)$$

2. Variabilidad explicada $VE(h)$: según la ecuación (7)

$$VE(h) = \text{traza}(\tilde{\mathbf{Y}}_{ch}^t \tilde{\mathbf{Y}}_{ch}) \quad (7)$$

3. τ de Goodman-Kruskal (8):

$$\begin{aligned} \tau_{y_h \cdot x_h} &= \frac{\frac{VE(h)}{n}}{\frac{VT(h)}{n}} \\ &= \frac{\sum_{j=1}^q \sum_{i=1}^p \frac{f_{ijh}^2}{f_{i \cdot h}} - \sum_{j=1}^q f_{\cdot jh}^2}{1 - \sum_{j=1}^q f_{\cdot jh}^2} \end{aligned} \quad (8)$$

Etapas 3. Se determinan las direcciones principales de un ACNS obtenidas mediante la descomposición espectral de la matriz $\tilde{\mathbf{Y}}_h^t \tilde{\mathbf{Y}}_h$ que a su vez genera una descomposición de $\tau_{y_h \cdot x_h}$ en sumandos, cada uno de los cuales mide la porción de la variabilidad de y_h explicada por x_h , captada por la correspondiente dirección principal.

Con este fin se obtienen los autovalores y autovectores de la matriz $\tilde{\mathbf{Y}}_h^t \tilde{\mathbf{Y}}_h$:

$$\tilde{\mathbf{Y}}_h^t \tilde{\mathbf{Y}}_h v^{\alpha_h} = \lambda_{\alpha_h} v^{\alpha_h}, \alpha_h = 1, \dots, q, h = 1, \dots, H$$

Calculando además para cada dirección el cociente $\frac{\lambda_{\alpha_h}}{\tau_{y_h \cdot x_h}}$. Este cociente mide la proporción de la variabilidad de y_h explicada por x_h , captada por la α_h -ésima dirección principal.

Etapas 4. Para cada ocasión h , se construye una representación biplot de las filas (probabilidades condicionales de $y|x = i$ para las distintas modalidades de la variable y) y las columnas (probabilidades condicionales de $y = j$ para las distintas modalidades de la variable x) de la matriz $\tilde{\mathbf{Y}}_h$. El vector de proyección de las filas sobre el eje α_h :

$$\begin{aligned} \psi^{\alpha_h} &= \tilde{\mathbf{Y}}_h v^{\alpha_h}, \\ \alpha_h &= 1, \dots, q, \\ h &= 1, \dots, H \end{aligned}$$

así que la proyección de un individuo s que posea la modalidad i de x_h (proyección de la i -ésima modalidad de x_h) viene dada por:

$$\psi_{s\alpha_h} = \sum_{j=1}^q v_{j\alpha_h} (P(y_h = j | x_h = i)),$$

$$\alpha_h = 1, \dots, q,$$

$$h = 1, \dots, H$$

y las coordenadas de proyección de las columnas de la j -ésima modalidad de $\mathbf{y}_h$:

$$\varphi_{j\alpha_h} = v_{j\alpha_h}, \quad \mathbf{j}, \alpha_h = 1, \dots, q, \quad \mathbf{h} = 1, \dots, H$$

1.1.5. STATIS-CNS2: Estrategia ICI

En el caso de esta investigación en particular se pretende comparar la variabilidad de la variable $\mathbf{x}_h$ en función de la variable $\mathbf{y}_h$, entre H ocasiones, en términos del coeficiente $\tau_{y_h \cdot x_h}$ de Goodman - Kruskal.

Para ello se procede 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 $\tau_{y_h \cdot x_h}$ de Goodman - Kruskal en cada bloque no simétrico $(\mathbf{x}_h, \mathbf{y}_h)$, $h = 1, \dots, H$.

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 $(\mathbf{x}_h, \mathbf{y}_h)$, $h = 1, \dots, H$, que son similares entre sí. (**interestructura**).
2. Construir un marco de representación común a todos los bloques de información $(\mathbf{x}_h, \mathbf{y}_h)$, $h = 1, \dots, H$, (**compromiso**).
3. Describir las diferencias o similitudes entre los individuos (**intraestructura**), utilizando como pieza fundamental el ACNS para analizar el comportamiento de los individuos en cada ocasión y los bloques que de alguna manera expliquen las razones o causas de las semejanzas y/o diferencias entre los individuos.

En la Figura 2 se presenta un resumen de la estrategia ICI para el STATIS.

En la siguiente sección se presentan los elementos teóricos necesarios para efectuar la

adaptación del STATIS a la problemática planteada en esta investigación.

2. Metodología STATIS-CNS2

En esta sección se presenta un resumen de los distintos elementos requeridos para el desarrollo de la metodología, para más detalle de los cálculos contemplados en esta sección se recomienda consultar el capítulo 3 en [4].

2.1. Elementos básicos del STATIS-CNS2

Definición 1 El estudio para el h -ésimo bloque se define como la tripleta $E_h = ((\mathbf{X}_h, \mathbf{Y}_h), \mathbf{M}_h, \mathbf{D}_h)$, donde:

1. $\mathbf{X}_h$ y $\mathbf{Y}_h$, son las matrices disyuntivas correspondientes a las variables cualitativas $\mathbf{x}$ e $\mathbf{y}$, en la h -ésima ocasión sobre n individuos.
2. $\mathbf{M}_h = (\mathbf{X}_h^t \mathbf{X}_h)^{-1}$ es una matriz definida positiva que define la métrica utilizada para construir las distancias entre los individuos.
3. $\mathbf{D}_h = \frac{1}{\sqrt{VT(h)}} \mathbf{I}_n$ siendo $VT(h)$ la variabilidad total de $\mathbf{y}$ en el bloque h .

Definición 2 (Definición de objeto no simétrico)

Dado el estudio para el h -ésimo bloque E_h , se define el objeto no simétrico de acuerdo con la ecuación (9):

$$\mathbf{W}_h = \mathbf{X}_h \mathbf{M}_h \mathbf{X}_h^t \mathbf{Y}_h = (\mathbf{P}_{\mathbf{X}_h} - \mathbf{P}_m) \mathbf{Y}_h \quad (9)$$

donde la c indica en este caso que el arreglo correspondiente está centrado, $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$, donde $\mathbf{P}_{\mathbf{X}}$ es la matriz de proyección sobre el espacio generado por las modalidades de la x en la h -ésima ocasión. y $\mathbf{P}_m = \frac{(jj^t)}{n}$, siendo $j = (1, \dots, 1)^t$.

El objeto $\mathbf{W}_h$ 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 modalidades q de las

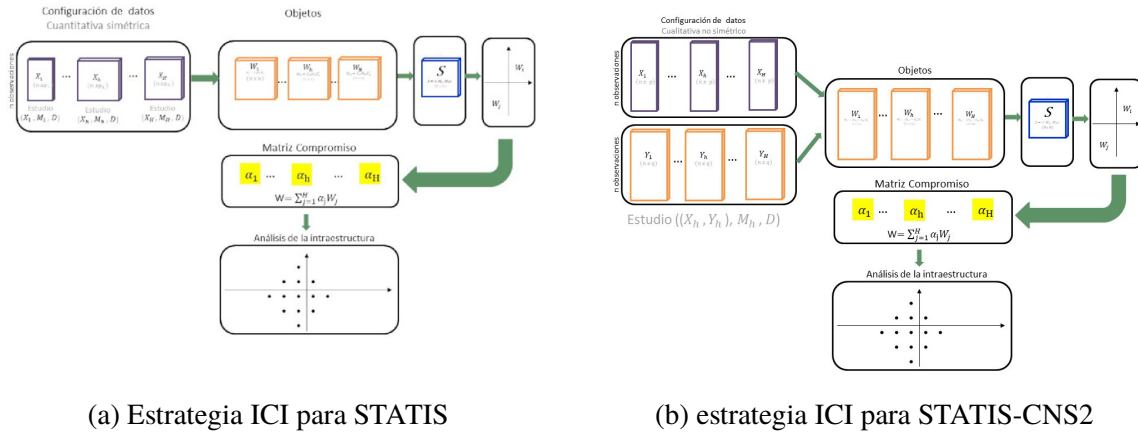


Figura 2: Estrategia ICI de cada una de las metodologías

variables criterio. Estos objetos no son matrices cuadradas como en el STATIS clásico y menos aún simétricas.

El número de columnas del bloque del objeto depende del número de modalidades de la variable categórica y , las columnas de $\mathbf{W}_h$, se definen mediante la notación $\mathbf{W}_{lh}$ para $l = 1, \dots, q$. Cada vector $\mathbf{W}_{lh}$, $l = 1, \dots, q$ contiene las estimaciones mínimo-cuadráticas de la variable criterio y , medida en la ocasión h , en función de las variables explicativas $\mathbf{x}$.

Definición 3 (producto escalar de Frobenius)

Se define el producto escalar de tipo Frobenius entre dos matrices $\mathbf{W}_h$ y $\mathbf{W}_l$, de dimensión $n \times q$, según la ecuación (10):

$$\langle \mathbf{W}_h | \mathbf{W}_l \rangle_{FCNS2} = \text{traza}((\mathbf{D}_h \mathbf{W}_h)^t \mathbf{D}_l \mathbf{W}_l) = \text{traza}(\mathbf{W}_h^t \mathbf{D}_h \mathbf{D}_l \mathbf{W}_l) \quad (10)$$

con rango de valores sobre la recta real, tanto positivos como negativos.

2.1.1. Producto escalar Frobenius en el momento h

El producto escalar Frobenius en el momento h se expresa según la ecuación (11):

$$\begin{aligned} \langle \mathbf{W}_h | \mathbf{W}_h \rangle_{FCNS2} &= \text{traza}((\mathbf{D}_h \mathbf{W}_h)^t \mathbf{D}_h \mathbf{W}_h) \\ &= \frac{1}{\sqrt{VT(h)VT(h)}} \text{traza}(\tilde{\mathbf{Y}}_{ch}^t \tilde{\mathbf{Y}}_{ch}) \\ &= \tau_{y_h \cdot x_h} \end{aligned} \quad (11)$$

así que el producto escalar de un objeto consigo mismo, en la h -ésima ocasión, indica una relación de variabilidad o asociación entre variables medido a través del $\tau_{y_h \cdot x_h}$.

2.1.2. Producto escalar Frobenius en los momentos h y l

El producto escalar Frobenius en los momentos h y l se expresa según la ecuación (12).

$$\begin{aligned} \langle \mathbf{W}_h | \mathbf{W}_l \rangle_{FCNS2} &= \text{traza}((\mathbf{D}_h \mathbf{W}_h)^t \mathbf{D}_l \mathbf{W}_l) \\ &= \text{traza}(\mathbf{W}_h^t \mathbf{D}_h \mathbf{D}_l \mathbf{W}_l) \\ &= \frac{1}{\sqrt{VT(h)VT(l)}} \text{traza}(\tilde{\mathbf{Y}}_{ch}^t \tilde{\mathbf{Y}}_{cl}) \end{aligned} \quad (12)$$

Este producto queda definido como función de las covarianzas entre las estimaciones de y en el bloque h y en el bloque l .

Definición 4 (Distancia entre objetos) Se define una distancia entre objetos basada en la norma dada de forma natural por el producto escalar de Frobenius, expresado en la ecuación (13):

$$\begin{aligned}
 d_{FCNS2}^2(\mathbf{W}_h, \mathbf{W}_l) &= \|\mathbf{W}_h - \mathbf{W}_l\|_{FCNS2}^2 \\
 &= \tau_{y_h \cdot x_h} + \tau_{y_l \cdot x_l} \\
 &\quad - 2\langle \mathbf{W}_h | \mathbf{W}_l \rangle_{FCNS2}
 \end{aligned} \quad (13)$$

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 D_{FCNS2} , las diferencias existentes entre las estructuras definidas por las estimaciones mínimo-cuadráticas del bloque de las q variables criterio en dos ocasiones diferentes que son

$$S = \begin{pmatrix} \langle \mathbf{W}_1 | \mathbf{W}_1 \rangle_{FCNS2} & \cdots & \langle \mathbf{W}_1 | \mathbf{W}_i \rangle_{FCNS2} & \cdots & \langle \mathbf{W}_1 | \mathbf{W}_H \rangle_{FCNS2} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \langle \mathbf{W}_i | \mathbf{W}_1 \rangle_{FCNS2} & \cdots & \langle \mathbf{W}_i | \mathbf{W}_i \rangle_{FCNS2} & \cdots & \langle \mathbf{W}_i | \mathbf{W}_H \rangle_{FCNS2} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \langle \mathbf{W}_H | \mathbf{W}_1 \rangle_{FCNS2} & \cdots & \langle \mathbf{W}_H | \mathbf{W}_i \rangle_{FCNS2} & \cdots & \langle \mathbf{W}_H | \mathbf{W}_H \rangle_{FCNS2} \end{pmatrix}$$

Sustituyendo las expresiones que definen a $\mathbf{W}_h$ y $\mathbf{W}_l$, el término general queda expresado según la ecuación (14):

$$\begin{aligned}
 S_{hl} &= \frac{1}{\sqrt{VT(h)VT(l)}} \text{traza}(\tilde{\mathbf{Y}}_{ch}^t \tilde{\mathbf{Y}}_{cl}) \\
 &= \frac{1}{\sqrt{VT(h)VT(l)}} \text{traza}(\mathbf{Y}_h^t (\mathbf{P}_{X_h}^t \mathbf{P}_{X_l} - \mathbf{P}_m) \mathbf{Y}_l)
 \end{aligned} \quad (14)$$

donde $\mathbf{P}_{X_h}$ es la matriz de proyección sobre el espacio generado por las modalidades de la x en la h -ésima ocasión y así mismo $\tilde{\mathbf{Y}}_h = (\tilde{\mathbf{Y}}_{(1,h)}, \dots, \tilde{\mathbf{Y}}_{(q,h)})$ y $\tilde{\mathbf{Y}}_l = (\tilde{\mathbf{Y}}_{(1,l)}, \dots, \tilde{\mathbf{Y}}_{(q,l)})$ son las proyecciones de las y sobre el espacio generado por las modalidades de las x en las ocasiones h y l .

Un resultado muy importante es que sobre la diagonal principal de la matriz $\mathbf{S}$ se encuentran los índices de asociación de Goodman-Kruskal correspondientes a cada uno de los H pares de

obtenidas en función de los correspondientes bloques de variables explicativas.

A continuación, se presenta el análisis que permite obtener la representación euclídea de los objetos $\mathbf{W}_h$, $h = 1, \dots, H$.

2.1.3. Espacio de representación de la interestructura

Se define a continuación un arreglo matricial que contiene los productos escalares de Frobenius entre los objetos de los distintos bloques.

Definición 5 (Matriz de productos escalares $\mathbf{S}$)

Se denotará mediante $\mathbf{S}$ a la matriz de orden $H \times H$ que contiene los productos escalares de Frobenius entre los objetos dos a dos:

bloques, expresados en la ecuación (15):

$$S_{hh} = \frac{1}{VT(h)} \text{traza}(\mathbf{Y}_h^t (\mathbf{P}_{X_h} - \mathbf{P}_m) \mathbf{Y}_h) \quad (15)$$

La construcción de la matriz $\mathbf{S}$ 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 $\mathbf{S}$ es en sí misma una medida de la asociación contenida entre los H bloques bajo estudio.

El elemento S_{hl} fuera de la diagonal principal de la matriz $\mathbf{S}$, constituye una medida agregada de las covarianzas entre las estimaciones mínimo-cuadráticas de la variable criterio y en las ocasiones h y l . En efecto, quedando expresado en la ecuación (16).

$$\begin{aligned} \text{traza}(\mathbf{W}_h^t \mathbf{W}_h) &= \frac{1}{VT(h)} \text{traza}(\tilde{\mathbf{Y}}_h^t \tilde{\mathbf{Y}}_h) \\ &= \sum_{j=1}^q \tilde{\mathbf{Y}}_{jh}^t \tilde{\mathbf{Y}}_{jh} \end{aligned} \quad (16)$$

Como indicamos previamente la matriz que define las representaciones en el STATIS-CNS2 es la matriz $\mathbf{S}$ que contienen los productos escalares de Frobenius entre objetos de los distintos bloques. Con la finalidad de construir un espacio de representación óptimo para los objetos $\mathbf{W}_h$ se realiza un análisis de componentes principales sobre $\mathbf{S}$, determinado por las direcciones principales que definen los índices de asociación de Goodman-Kruskal $\tau_{y_h \cdot x_h}$. 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, según la ecuación (17):

$$\begin{aligned} \|\mathbf{W}_h\|_{FCNS2}^2 &= \text{traza}(\mathbf{W}_h^t \mathbf{W}_h) \\ &= \frac{VE(h)}{VT(h)} = \tau_{y_h \cdot x_h} \end{aligned} \quad (17)$$

Se efectúa la descomposición espectral de esta matriz $\mathbf{S}$ con el propósito de hallar un espacio de representación para los objetos de los distintos bloques. Se obtienen entonces los autovalores y autovectores: $SG^\alpha = t_\alpha G^\alpha$, $\alpha = 1, 2, \dots, H$.

Es decir la descomposición espectral de $\mathbf{S}$, queda descrita mediante la expresión $\mathbf{S} = \mathbf{G}\mathbf{T}\mathbf{G}^t$, siendo $\mathbf{G} = (\mathbf{G}^1, \dots, \mathbf{G}^h, \dots, \mathbf{G}^H)$ la matriz cuyas columnas son los autovectores de la matriz $\mathbf{S}$ asociados con los correspondientes autovalores ordenados de forma descendente y organizados sobre la diagonal de la matriz:

$$\begin{pmatrix} t_1 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \ddots & 0 \\ 0 & \cdots & t_h & \cdots & 0 \\ 0 & \ddots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \cdots & t_H \end{pmatrix}$$

encontrándose con esta factorización la posibilidad de descomponer el índice de asociación de Goodman-Kruskal (18):

$$\text{traza}(\mathbf{S}) = \sum_{\alpha=1}^H \tau_{y_\alpha \cdot x_\alpha} = \sum_{\alpha=1}^H t_\alpha \quad (18)$$

Es decir, al diagonalizar la matriz $\mathbf{Y}_h^t (\mathbf{P}_{\mathbf{X}_h} - \mathbf{P}_m) \mathbf{Y}_h$ se obtiene que la traza de la matriz $\mathbf{S}$ es $\sum_{\alpha=1}^H \tau_{y_\alpha \cdot x_\alpha}$, donde cada autovalor está asociado con una dirección principal del análisis de correspondencias no simétrico (ACNS). Por lo tanto, la h -ésima dirección de este espacio, G^α , capta una porción igual a: $\frac{t_\alpha}{\sum_{\alpha=1}^H \tau_{y \cdot x_h}}$ de la medida global en que la x explica a la y a lo largo de los bloques.

2.2. Construcción de un objeto compromiso

El objeto compromiso resulta ser una pieza fundamental de análisis en la metodología STATIS. El propósito del objeto compromiso es el de realizar una comparación global de las H ocasiones. Este objeto es requerido en un intento de establecer un espacio de referencia sobre el cual sea posible representar a todos los elementos de interés en el análisis, que se denomina espacio compromiso. El objeto en referencia se construye como un promedio ponderado de los objetos que identifican a los diferentes bloques de datos $(\mathbf{x}_h, \mathbf{y}_h)$, $h = 1, \dots, H$, en el que los coeficientes usualmente se definen en función de las coordenadas del autovector de la matriz $\mathbf{S}$, asociado al primer autovalor.

Se plantea construir el objeto compromiso como una combinación lineal de los objetos de la forma en la ecuación (19)

$$\begin{aligned} \mathbf{W}_{\text{comp}} &= \sum_{h=1}^H \alpha_h \mathbf{W}_h \\ &= \sum_{h=1}^H \alpha_h \tilde{\mathbf{Y}}_{ch} \\ &= \sum_{h=1}^H \alpha_h (\mathbf{P}_{\mathbf{X}_h} - \mathbf{P}_m) \mathbf{Y}_h \end{aligned} \quad (19)$$

de manera que se maximice globalmente una medida de covarianza con los objetos analizados

en el sentido del producto interno definido sobre los objetos según la ecuación (20):

$$\begin{aligned} \langle \mathbf{W}_h | \mathbf{W}_1 \rangle_{FCNS2} &= \text{traza}(\mathbf{W}_h^t \mathbf{D}_h \mathbf{D}_1 \mathbf{W}_1) \\ &= \frac{1}{\sqrt{VT(h)VT(l)}} \text{traza}(\tilde{\mathbf{Y}}_{ch}^t \tilde{\mathbf{Y}}_{cl}^t) \end{aligned} \quad (20)$$

La función objetivo entonces, es de la forma: $\sum_{h=1}^H \langle \mathbf{W}_{comp} | \mathbf{W}_h \rangle_{FCNS2}^2 = \alpha^t \mathbf{S} \mathbf{S}^t \alpha$, con la restricción: $\sum_{h=1}^H \alpha_h^2 = 1$.

La expresión definida $\mathbf{W}_{comp}$ destaca la naturaleza del objeto compromiso, en la que cada sumando es, salvo un escalar, la proyección de $\mathbf{Y}_h$, sobre las columnas de $\mathbf{X}_h$ correspondiente, siendo por lo tanto el compromiso de dimensión $n \times q$, perteneciente al espacio generado por las matrices $\mathbf{W}_1, \dots, \mathbf{W}_h, \dots, \mathbf{W}_H$.

El problema reside en la determinación del vector $\alpha^t = (\alpha_1, \dots, \alpha_h, \dots, \alpha_H)$ de coeficientes del objeto compromiso.

Se considera en primer lugar que: $\langle \mathbf{W}_{comp} | \mathbf{W}_1 \rangle_{FCNS2} = \langle \sum_{h=1}^H \alpha_h \mathbf{W}_h | \mathbf{W}_1 \rangle_{FCNS2} = \alpha^t \mathbf{S}^l$, siendo $\mathbf{S}^l$ la l -ésima columna de la matriz de productos interno $\mathbf{S}$. Por lo tanto la función a maximizar es $\sum_{l=1}^H \langle \mathbf{W}_{comp} | \mathbf{W}_l \rangle_{FCNS2}^2 = \alpha^t \mathbf{S} \mathbf{S}^t \alpha = \alpha^t \mathbf{S}^2 \alpha$ sujeto a que α sea un vector normalizado.

Para hallar el vector que maximice la expresión, con la condición dada se define el lagrangiano $L = \alpha^t \mathbf{S} \mathbf{S}^t \alpha - \lambda(\alpha^t \alpha - 1)$

La solución se obtiene de la forma usual: $\frac{\partial L}{\partial \alpha} = 2\mathbf{S} \mathbf{S}^t \alpha - 2\lambda \alpha = 0 \Rightarrow \mathbf{S} \mathbf{S}^t \alpha = \lambda \alpha$

Por lo tanto el vector de α de coeficientes buscado es el autovector normalizado de la matriz simétrica $\mathbf{S} \mathbf{S}^t$ asociado con su mayor autovalor $(t_1)^2$, en la forma:

$$\alpha = \frac{\mathbf{G}^1}{\|\mathbf{G}^1\|_{FCNS2}} = \frac{\mathbf{G}^1}{(\mathbf{G}^{1t} \mathbf{G}^1)^{1/2}}.$$

3. Aplicación del STATIS-CNS2

Según publica la Superintendencia Financiera de Colombia en el capítulo 2 de la circular externa básica contable y financiera 100 de 1995

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

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 consonancia con 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 los que están expuestas como bancos.

3.1. Configuración de los datos

En esta sección se ilustra la técnica propuesta, aplicándola sobre un conjunto de datos reales referidos a dos variables cualitativas $\mathbf{x}_1, \mathbf{x}_2$ e $\mathbf{y}$, medidas sobre 20316 individuos a lo largo de los 4 trimestres correspondientes al año 2020. Las variables consideradas son:

1. Una variable $\mathbf{x}_1$ que corresponde a la calificación de riesgo (a nivel de cliente) de cada cliente con 5 modalidades: A, B, C, D y E.
2. Una variable $\mathbf{x}_2$ que corresponde al staging de riesgo (a nivel de cliente) de cada cliente con 3 modalidades: stage1, stage2 y stage3.
3. Una variable $\mathbf{y}$ que corresponde a un estado 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 parte del supuesto que: (1) la calificación interna del riesgo de cada cliente (variable $\mathbf{x}_1$) explica de alguna manera el estado obtenido con información externa (variable $\mathbf{y}$) y (2) el staging-NIIF9 de cada cliente (variable $\mathbf{x}_2$) explica de alguna manera el estado obtenido con información externa (variable $\mathbf{y}$).

A lo largo de esta sección, se detallan los bloques de datos explicando su estructura lógica, explorando en las siguientes secciones por un lado el bloque (x_{1h}, y_h) y por el otro el bloque (x_{2h}, y_h) , $h = 1, \dots, 4$, desde el punto de vista de la metodología STATIS-CNS2.

3.1.1. Definición de variables

A continuación presentaremos la definición detallada de cada variable.

1. **Calificación del riesgo crediticio. (variable x_1)** Según la norma de la SuperFinanciera de Colombia <https://www.superfinanciera.gov.co/publicacion> 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”, categoría C o “riesgo apreciable”, categoría D o “riesgo significativo” y categoría E o “riesgo de incobrabilidad”.

Para más detalle de la definición de esta variable consulte el capítulo 4 de [4].

2. **Staging de riesgo (variable x_2)**

La NIIF 9 (Norma Internacional de Información Financiera 9, <https://www.bis.org/press/p161011.htm>) establece una serie de reglas y definiciones para los cálculos de provisiones a través de la pérdida esperada de los instrumentos financieros. Su correcta implementación requiere un conjunto de reglas y parámetros, entre ellos la definición de staging de los clientes al momento de cada fecha de reporte.

- Stage 3: Se considerará que el cliente se encuentra en stage 3 si satisface alguna de las siguientes condiciones: default > 60 días y calificación de riesgo D y E.
- Stage 2: el cliente se encuentra en este stage si satisface alguna de las siguientes condiciones: default > 30 días, o cuando la calificación es mayor a A, es decir, su valor puede ser B, C, D o E o cuando la calificación es A y la altura de mora es mayor a 16 días.

- Stage 1: Lo que no cumpla con las condiciones anteriores.

3. **Estado de riesgo (variable y)** Segmentación que realiza la entidad financiera para realizar un seguimiento a todos los clientes compartidos (Los 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), 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. (La 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)) . Las 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 consultados en las centrales de riesgo en el trimestre cumplen sus obligaciones financieras con el sector y cumplen con la entidad.
- b) Estado 2: constituido por aquellos clientes que al ser consultados en las centrales de riesgo en el trimestre incumplen sus obligaciones financieras con el sector y cumplen con la entidad.
- c) Estado 3: constituido por aquellos clientes que al ser consultados en las centrales de riesgo en el trimestre cumplen sus obligaciones financieras con el sector e incumplen con la entidad.

- d) Estado 4: constituido por aquellos clientes que al ser consultados en las centrales de riesgo en el trimestre incumplen sus obligaciones financieras con el sector e incumplen con la entidad.

Es importante observar que la variable x_1 es una variable con normativa colombiana utilizada por las entidades financieras para el cálculo de provisiones en contabilidad local mientras que la variable x_2 es utilizada para el cálculo de provisiones en contabilidad de normativa internacional.

3.1.2. Descripción de la muestra analizada

La muestra analizada forma parte de un estudio trimestral de endeudamiento que realiza una entidad financiera. Se trabaja con $n=20316$ clientes que fueron observados trimestralmente durante 12 meses ($H=4$ trimestres). Para fines prácticos utilizaremos la notación $2020T_i$, $i = 1, \dots, 4$, para indicar el i -ésimo trimestre del año 2020. Los $n=20316$ clientes fueron evaluados en cada trimestre en cada una de las variables mencionadas (estado de riesgo, calificación de riesgo y staging).

3.1.3. Análisis univariado de los datos

El año 2020 el coronavirus Covid-19, produjo un gran impacto financiero a nivel mundial. Esto se puede apreciar fácilmente en la distribución que vemos en las Tablas 1, 2 y 3. Los clientes deterioraron tanto en calificación como en staging de riesgo. No obstante, se observa un aumento más acentuado para los stages en los trimestres 3 y 4 (9 % y 17 %). Dado que este aumento representa un gran impacto en el cálculo de provisiones bajo contabilidad NIIF9 para la entidad, se solicita hacer un análisis exploratorio más detallado de los datos para detectar si existe algún error en el proceso de obtención de los datos o si estos valores son los correctos y efectivamente la junta directiva debe hacer frente a estos valores de provisiones.

3.2. Índice de Goodman-Kruskal

El analista de datos al recibir la solicitud de hacer un análisis exploratorio detallado, decide estudiar la influencia (por separado) que las variables

categorías calificación (x_1) y staging (x_2) ejercen sobre la variable estado de riesgo (y), y estudiar los cambios que pudieran producirse a lo largo de 2020. Para esto decide utilizar el índice estadístico τ de Goodman-Kruskal para medir la intensidad de la fuerza con la cual, por un lado, la variable calificación explica el estado de riesgo y por el otro, la intensidad con la cual la variable staging explica el estado de riesgo.

Los resultados están expuestos en la figura 3, como podemos apreciar la intensidad de la fuerza con la cual, la calificación explica el estado de riesgo es superior en los 4 trimestres a la intensidad de la fuerza con la que el staging explica el estado de riesgo. Además los valores del índice de Goodman-Kruskal en el caso de NIIF9 es inferior al 30 % en todo el año. El analista de datos sugiere que se debe realizar una revisión de las variables utilizadas para el cálculo de la calificación de riesgo para los trimestres 2 y 3 del año 2020 porque los valores del índice de Goodman-Kruskal bajan drásticamente a 20,1 % y 9,68 %, respectivamente. Presentando valores por encima del 70 % para los trimestres 1 y 4.

Con este primer análisis exploratorio la información debe ser evaluada por un perfil de ingeniería de datos, quien al consultar el origen de los datos confirma que la información es correcta y recomienda hablar con el dueño del proceso para ver si algún proceso macroeconómico puede estar causando este comportamiento en los datos.

3.3. Análisis interno de cada bloque con un enfoque ACNS

El analista de datos, decide continuar con el análisis exploratorio. Tomando en consideración la configuración de los datos, decide realizar un análisis de correspondencias no simétrico (ACNS). Al examinar los gráficos de biplot del ACNS (Figura 4) se ve principalmente que para los trimestres 1 y 4: los clientes con calificación A están asociados al estado 1, los clientes con calificación B están asociados a los estado 3 y 2, y las calificaciones C,D y E al estado 4. En cambio para los trimestres 2 y 3 hay muchos clientes con calificación B,C, D E y estado de riesgo 1. Dado que este resultado no es el esperado y ya se ha

Tabla 1: 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 $R^p, p = 5$).

	2020T1		2020T2		2020T3		2020T4	
A	19785	97 %	19666	97 %	18962	93 %	18159	89 %
B	326	2 %	389	2 %	963	5 %	667	3 %
C	60	0 %	27	0 %	145	1 %	616	3 %
D	67	0 %	49	0 %	63	0 %	546	3 %
E	78	0 %	185	1 %	183	1 %	328	2 %
Total	20316	100 %	20316	100 %	20316	100 %	20316	100 %

Observe que a medida que se avanza en el tiempo los clientes migran de calificación de riesgo bajas a calificaciones de riesgo altas, esto es medido a través del vector de centro de gravedad de esta variable.

Tabla 2: Comportamiento de la variable staging de riesgo cliente a lo largo del período de tiempo bajo estudio -Centro de gravedad de la nube de perfiles fila (en $R^p, p = 3$).

	2020T1		2020T2		2020T3		2020T4	
Stage1	18797	93 %	18574	91 %	15548	77 %	15504	76 %
Stage2	762	4 %	684	3 %	2907	14 %	1403	7 %
Stage3	757	4 %	1058	5 %	1861	9 %	3409	17 %
Total	20316	100 %	20316	100 %	20316	100 %	20316	100 %

Observe que a medida que se avanza en el tiempo los clientes migran de staging de riesgo bajos a staging de riesgo altos, esto es medido a través del vector de centro de gravedad de esta variable.

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 $R^q, q = 4$).

	2020T1		2020T2		2020T3		2020T4	
Estado 1	19774	97 %	18805	93 %	18485	91 %	18160	89 %
Estado 2	18	0 %	866	4 %	1362	7 %	280	1 %
Estado 3	228	1 %	361	2 %	285	1 %	779	4 %
Estado 4	296	1 %	284	1 %	184	1 %	1097	5 %
Total	20316	100 %	20316	100 %	20316	100 %	20316	100 %

Observe que a medida que se avanza en el tiempo los clientes migran de estados de riesgo bajos a estados de riesgo altos, esto es medido a través del vector de centro de gravedad de esta variable.

validado que los datos son correctos, se procede a hacer una análisis STATIS-CNS2 para determinar si un plano de representación único pueda aportar más información.

3.4. STATIS-CNS2: Estrategia ICI

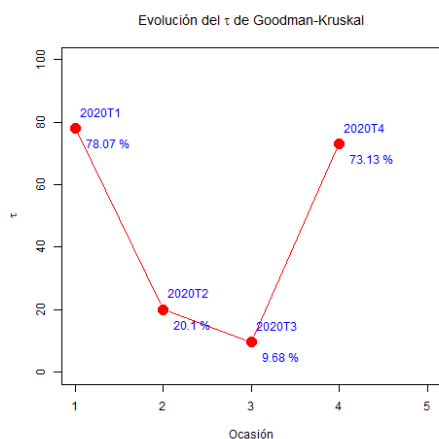
3.4.1. Matriz S

En la Figura 5 (5a, 5b), se puede apreciar que en el plano factorial obtenido en la descomposición de la matriz S , se forman dos grupos, un primer grupo constituido por la información de los trimestres 1 y 4, y otro grupo constituido por los trimestres 2 y 3. Adicional en la figura relacionada con el plano del Staging los puntos están más cerca del origen de coordenadas, este resultado va de la mano con el hecho que τ de Goodman-Kruskal toma

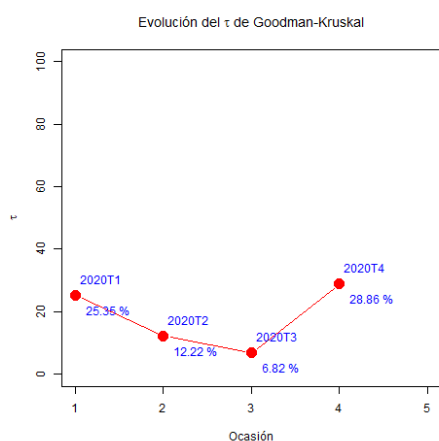
valores muy bajos para la relación no simétrica entre Staging-Estado. A partir de la matriz S se hace la reconstrucción de las distancias entre los objetos (5c, 5d), pudiendo apreciar fácilmente que el objeto W_4 está más alejado de los objetos W_1, W_2 y W_3 (con ambas variables).

3.4.2. Compromiso

El propósito fundamental por el cual el analista de datos realiza la construcción de la matriz compromiso, es disponer de una herramienta que permita realizar una comparación global en los 4 trimestres, es un intento de poder establecer un espacio de referencia sobre el cual sea posible representar a todos los clientes de interés en el análisis.



(a) Calificación-Estado



(b) Staging-Estado

Figura 3: Evolución de τ de Goodman-Kruskal

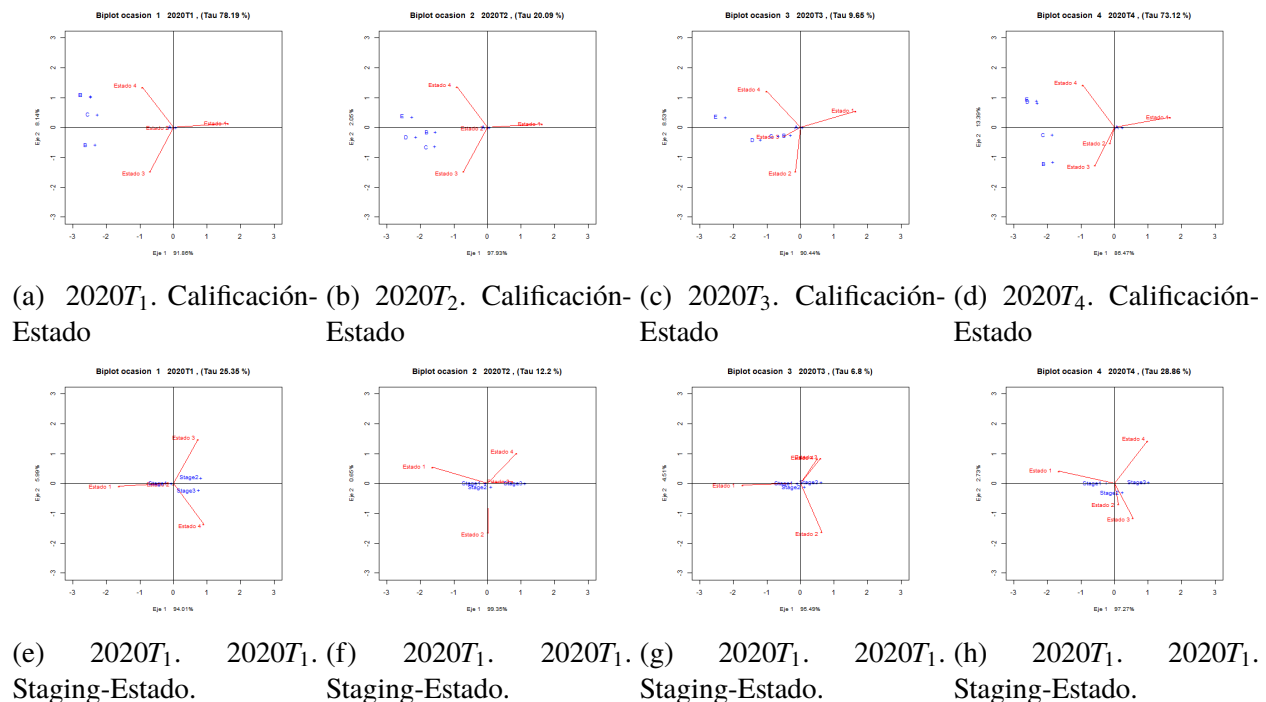
De la Figura 6a podemos deducir que los clientes son segmentados en 4 cuadrantes. De la relación calificación-estado de riesgo se deduce:

- Para el primer cuadrante: tiene 822 clientes (4 % del total). 618 clientes (75 % del total cuadrante) comienzan la calificación de riesgo A (2020T₁) y terminan en calificaciones D y E (2020T₄). Para estos clientes los estados de riesgo comienzan en los perfiles 1 y 2 (2020T₁) y terminan en perfiles 3 y 4 (2020T₄).
- En el segundo cuadrante hay 17684 clientes (87 % del total). De los cuales 16337 clientes (92 % del total cuadrante) se caracterizan por tener la combinación calificación A-Estado 1 en los 4 trimestres (resultado esperado).

- En el tercer Cuadrante hay 349 clientes (2 % del total). 289 clientes (83 % del total cuadrante) tienen la combinación calificación A-Estado de riesgo tipo 1 (a excepción del trimestre 2020T₃ que presentan calificación A con Estados diferentes al perfil 1).
- Finalmente en el cuarto cuadrante hay 1461 clientes (7 % del total). 950 clientes (65 % del total cuadrante) comienzan el primer trimestre (2020T₁) con calificaciones A,B y estado de riesgo 1 y finalizan el 4to trimestre (2020T₄) con calificaciones distintas de A y estados de riesgo diferentes al 1 (se deterioran).

Por otro lado, de la relación calificación-estado de riesgo utilizando 6b podemos deducir que los clientes son segmentados en 4 cuadrantes:

- Para el primer cuadrante: tiene 13740 clientes (68 % del total). 12951 clientes (94 % del total cuadrante) presentan la combinación stage 1 y estado de riesgo 1, en los 4 trimestres bajo estudio.
- En el segundo cuadrante hay 2852 clientes (14 %). De los cuales 1037 clientes (36 % del total cuadrante) se caracterizan por iniciar el primer trimestre (2020T₁) con la combinación stage1-estado 1 y terminar en el último trimestre (2020T₄) en stage 3 y estado de riesgo distinto de 1. 907 clientes (32 % del total cuadrante) que comienzan en stage 1 (2020T₁) y terminan (2020T₄) en stage 3 pero el estado de riesgo siempre es 1 (en los 4 trimestres)(resultado no esperado). 402 clientes (14 % del total cuadrante) que en los 4 trimestres están en stage3 y tienen estado de riesgo 1 (resultado no esperado).
- En el tercer Cuadrante hay 2244 clientes (11 % del total). 920 clientes (41 % del total cuadrante) tienen la combinación stage 1 en 2020T₁, terminan en stage 2 en 2020T₄ y siempre tienen estado de riesgo 1.
- Finalmente en el IV cuadrante hay 1480 clientes (7 % del total). 1038 clientes (70 % del total cuadrante) en los trimestres 2020T₁,



Figuras superiores: cuando la calificación interna del riesgo de cada cliente (variable x_1) explica de alguna manera el estado de riesgo. Figuras inferiores: cuando el staging de cada cliente (variable x_2) explica de alguna manera el estado obtenido con información externa (variable y)

Figura 4: Análisis de correspondencias no simétrico (ACNS) de los bloques apareados de información (x_h, y_h), $h = 1, \dots, H = 4$.

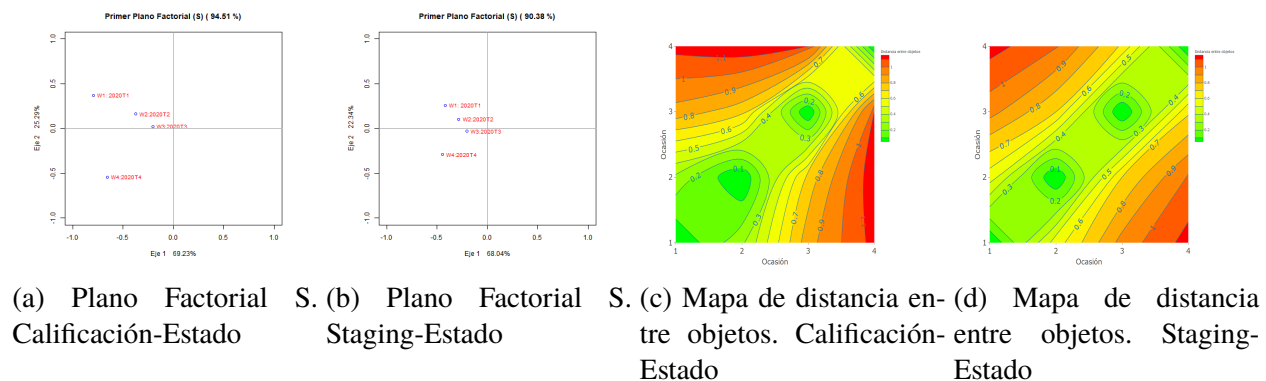
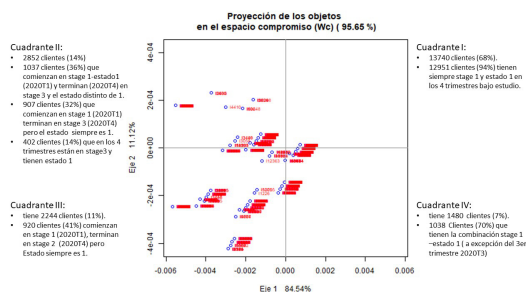


Figura 5: Herramientas para el análisis: Primer plano factorial S y Mapa de distancia entre objetos. Caso a: cuando la calificación interna del riesgo de cada cliente (variable x_1) explica de alguna manera el estado de riesgo. Caso b: cuando el staging de cada cliente (variable x_2) explica de alguna manera el estado de riesgo obtenido con información externa (variable y)

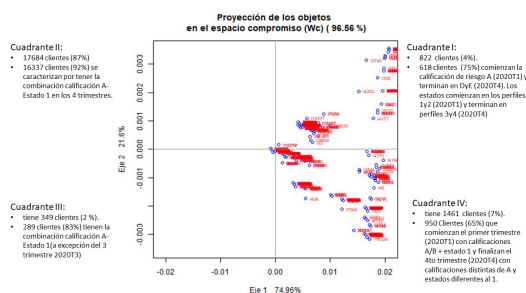
2020T₂ y 2020T₄ tienen la combinación stage 1-estado de riesgo 1. A excepción del 3er trimestre (2020T₃) donde presentan valores de alto riesgo.

3.4.3. Conclusiones de la aplicaciones

Como queda en evidencia el realizar directamente un análisis exploratorio utilizando la metodología STATIS-CNS2, permite al analista de datos, por una lado, identificar visualmente qué



(a) Plano Compromiso. Calificación-Estado



(b) Plano Compromiso. Staging-Estado

Figura 6: Gráfico superior: cuando la calificación interna del riesgo de cada cliente (variable x_1) explica de alguna manera el estado obtenido con información externa (variable y). Gráfico inferior: cuando el staging de cada cliente (variable x_2) explica de alguna manera el estado obtenido con información externa (variable y)

conjunto de datos son parecidos y/o diferentes entre sí, y por el otro, permite apreciar la intensidad de las fuerza con que las variables explican al estado de riesgo (Figura 5). Una vez identificados las diferencias y/o similitudes entre los bloques de datos, el beneficio de esta técnica de análisis exploratorio es la obtención de un espacio de referencia sobre el cual es posible representar a todos los clientes de interés en el análisis (Figuras 6a y 6b).

Para finalizar el ejemplo debemos indicar que durante el año 2020 debido a la pandemia en Colombia desde marzo de ese año, cuando se declaró el aislamiento obligatorio, el Gobierno colombiano a través de la superintendencia financiera lanzó diversas alternativas a fin de aliviar los impactos financieros causados por el COVID-19. Una de tales medidas fue realizar una normalización artificial de la cartera, en el sentido

que personas que entraron en mora durante ese período de tiempo, quedaban marcadas con 0 días a nivel de sistema. Este proceso de alivio, afectó directamente variables que dependían en su cálculo de los días de mora del cliente. Por esta razón el analista de datos detectó un comportamiento atípico en sus resultados, especialmente para los trimestres 2020T₂ y 2020T₃.

El disponer de herramientas de análisis exploratorio como el STATIS-CNS2 permite detectar no sólo que existe un problema a nivel de variables sino permite ubicar exactamente cuáles son los clientes donde se nota el comportamiento atípico. Una vez que se valida la calidad de los datos y el experto de negocio ayuda a entender que el resultado es correcto, el analista de datos debe dejar nota de esta información ya que el siguiente paso es analizar si esos clientes se van o no se van a eliminar para la construcción de futuros modelos estadísticos que vayan a incluir en su desarrollo información del año 2020.

Finalmente, como conclusión de este ejercicio con datos reales los resultados de provisiones tanto en contabilidad local como internacional son correctos y la junta directiva debe hacer frente al desembolso de las mismas.

4. Conclusiones

Con esta investigación se aporta una técnica de análisis exploratorio que permite adaptar la metodología STATIS para efectuar la comparación simultánea de los objetos que identifican las estructuras de interdistancias correspondientes a H ocasiones de bloques cualitativos, bajo una perspectiva no simétrica.

La tarea de mayor envergadura con la que nos enfrentamos consistió en definir un producto escalar adecuado entre objetos que permitiese comparar las estructuras que describen las interdistancias entre individuos caracterizados por los diferentes conjuntos de variables cualitativas. Adicionalmente el producto escalar permitió por un lado definir y conceptualizar una distancia estadística entre objetos y por el otro construir un espacio de baja dimensión, donde es posible efectuar comparaciones entre los diferentes conjuntos

de datos con relaciones no simétricas evaluados en términos del estadístico $\tau_{y_h \cdot x_h}$ de Goodman-Kruskal.

La propuesta planteada en esta investigación da origen a nuevas preguntas y formulación de otros problemas, pudiendo convertirse en un línea para futuras investigaciones. El hecho de desarrollar la adaptación para dos variables cualitativas explicativas y/o diferentes individuos pero mismas variables, abren fácilmente problemas para tesis doctorales y/o postdoctorales.

5. Referencias

- [1] N. Lauro, “L’analyse non symetrique des correspondances,” in *Data Analysis and Informatics III*, ed., E. . C. Diday, Ed. North-Holland: Elsevier, 1984.
- [2] Y. Escoufier, “Objectifs et procédures de l’analyse conjointe de plusieurs tableaux de données,” in *Statistique et Analyse des Données*, ed. France: Association pour la statistique et ses utilisations, 1985, vol. 10, no. 1, pp. 1–10.
- [3] L. Goodman and W. Kruskal, “Measures of association for cross classifications, IV: simplification of asymptotic variances,” *Journal of the american statistical association*, vol. 67, no. 338, pp. 415–421, 1972. <https://doi.org/10.2307/2284396>
- [4] J. Combariza, “Una adaptación del método statis para el análisis no simétrico de bloques de variables cualitativas,” Tesis doctoral, Universidad Central de Venezuela, Caracas, Venezuela, 2019.