FORTHCOMING 62D05-22-01-03
ANÁLISIS CONJUNTO DE DATOS MULTIVARIANTES: DOS APLICACIONES

Mario Miguel Ojeda Ramírez*[1], Cecilia Cruz López*, Roberto Gallardo del Ángel**
* Facultad de Estadística e Informática. Universidad Veracruzana. México.
** Facultad de Economía. Universidad Veracruzana. México.

ABSTRACT
In this work, with the aim of promoting the use of multivariate data set analysis techniques, a review of the methodology is presented as well as a couple of illustrations: with categorical data and with quantitative data. To give context to the illustrations, a general strategy for the application of this type of techniques is presented. Some recommendations are given on the use of the available R software libraries to apply these novel techniques.

KEYWORDs: Tandem analysis, Correspondence analysis, Principal component analysis, Cluster analysis, R software.

MSC: 62H30

RESUMEN
En este trabajo, con el objetivo de promover el uso de las técnicas de análisis conjunto de datos multivariantes, se presenta una revisión de la metodología, así como un par de ilustraciones: con datos categóricos y con datos cuantitativos. Para dar contexto a las ilustraciones se presenta una estrategia general de aplicación de este tipo de técnicas. Se dan algunas recomendaciones sobre el uso de las librerías de software R disponibles para aplicaciones.

PALABRAS CLAVE: Análisis tándem, Análisis de correspondencia, Análisis de componentes principales, Análisis clúster, Software R.

1. INTRODUCCIÓN

En los análisis de datos de encuestas y de otros datos disponibles en muchos sitios de internet (datos abiertos) se presenta con bastante frecuencia la situación de contar con varias variables categóricas o varias variables continuas. La estrategia de análisis convencional considera realizar análisis marginales seguidos de un análisis de asociación o correlación entre variables. Esto puede dar respuestas a preguntas específicas de investigación, pero regularmente la visión marginal y bivariante de los datos no resulta suficiente, ya que la naturaleza multivariante del problema demanda la consideración de todas las variables. Así, el propósito, en primera instancia, es hacer una descripción del conjunto de unidades (población o muestra) en términos de las asociaciones o correlaciones, pero en segunda instancia se requiere identificar los patrones de agrupación inherentes al interior de las unidades de estudio, lo que implicaría aplicar una segmentación; es decir, formar grupos que, por un lado, maximicen la homogeneidad al interior y que, por el otro, que entre sí sean lo más diferentes posible.
Si las variables fueran cuantitativas, una manera de abordar esta situación consiste primero en reducir la dimensión del problema, utilizando un análisis de componentes principales (ACP), y a continuación realizar un análisis clúster sobre las nuevas variables; es decir, realizar un análisis clúster con los dos o tres primeros componentes principales; este método se conoce como un *análisis tándem* (Hubert y Arabie, 1995; Krazanoswky, 1995). La principal ventaja que tiene este enfoque es que resulta sencillo y es muy intuitivo; no obstante, es posible que produzca asignaciones de clúster no óptimas, puesto que en el primer paso se optimiza la varianza explicada por las variables, y puede ser que al reducir la dimensión la estructura de agrupación se vea afectada. Debe recordarse que el análisis de componentes principales persigue el objetivo es encontrar un pequeño conjunto de combinaciones lineales de las variables que maximicen la varianza explicada; y, dado que el análisis clúster, por otro lado, tiene como objetivo encontrar individuos similares de acuerdo a las variables originales, resulta obvio sospechar que si la agrupación de observaciones se produce

[1] mojeda@uv.mx

en las variables que tengan poco peso en los componentes principales, ésta se perderá al usar esta estrategia secuencial.

También existe el *análisis tándem* en el caso de variables categóricas. Se aplicaría inicialmente un análisis de correspondencia múltiple (Greenacre and Blasius, 1994) y a partir de la reducción de dimensiones se realizaría un análisis clúster de los individuos. Vichi y Kiers (2001) señalaron que existían estudios que demostraban que el *análisis tándem* también tiene problemas dando soluciones no optimas de agrupación cuando las variables son categóricas. Para resolver esta situación, Hwang *et al.* (2006) propusieron métodos que evitan problemas potenciales asociados con el enfoque secuencial cuando se aplica a datos categóricos, pero sólo ilustraron tales métodos con algunos ejemplos específicos. Finalmente, motivados por esta ausencia de conocimiento, Van de Velden *et al.* (2017) desarrollaron el método de análisis conjunto para datos categóricos que denominaron Cluster AC, basándose en una revisión exhaustiva de las propuestas equivalentes que existían hasta la fecha; compararon los métodos existentes, para los que previamente Iodice D'Enza *et al.* (2014) habían expuesto algunas relaciones teóricas e ilustrado las diferencias utilizando un ejemplo con datos reales, y asimismo usando un experimento de simulación bastante extenso. A través de un trabajo detallado, demostraron que su análisis conjunto es ventajoso respecto al *análisis tándem* y a las soluciones conjuntas que se habían propuesto hasta entonces.

Finalmente, Markos *et al.* (2019b) presentaron una revisión de los elementos teóricos y metodológicos que dan sustento a una familia de técnicas multivariantes "ensambladas", que son nuevos procedimientos de análisis multivariante conjunto, los cuales realizan simultáneamente dos procesos de optimización: reducción de dimensionalidad (que trabaja sobre las variables) y agrupación (de individuos) a partir de la creación de clústeres. A partir del trabajo de estos autores, contamos con dos opciones desarrolladas, tanto teórica como numéricamente: (1) cuando los datos son categóricos, se realiza un análisis clúster (AC) conjuntamente con una reducción de dimensionalidad vía el análisis de correspondencia múltiple (ACM) —el que aquí se denominará Clúster AC—; y (2) cuando los datos son cuantitativos, se realiza un análisis clúster (AC) conjuntamente con una reducción de dimensionalidad por el método de ACP, —que se denominará Clúster CP—. Para llevar a la práctica estos procedimientos se han desarrollado librerías especializadas usando el software libre R, las cuales producen salidas numéricas y gráficas que informan sobre la asociación entre variables y la agrupación de los individuos (Markos *et al*., 2019a).

Esta metodología es la que se presenta en este trabajo, la cual en la parte teórica es una síntesis de lo presentado en Van de Velden *et al*. (2017) y Kamarras (2020). En el primer apartado se hace una revisión de la notación y los desarrollos, los que se presentan tanto para el Clúster AC como el Clúster CP. El segundo y tercer apartados se dedican a ilustrar el funcionamiento de las técnicas con una aplicación del Clúster AC y otra del Clúster CP. Finalmente se presenta unos comentarios a manera de conclusiones.

2. **METODOLOGÍA**

Consideremos que $\mathbf{X}$ $(n \times k)$ denota el conjunto de datos con n individuos u objetos y k variables; $\mathbf{U}$ $(n \times c)$ es una matriz indicadora que establece la pertenencia del objeto o individuo a cada uno de los c clústeres. Sea $\mathbf{Y}$ $(c \times m)$ la matriz que contiene las coordenadas de dimensión m para los centroides de los c clústeres; así, $\mathbf{A}$ $(k \times m)$ es la matriz de los pesos factoriales, con $\mathbf{A}'\mathbf{A} = \mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad de orden m; y $\mathbf{E}$ $(n \times k)$ es la matriz de los residuos correspondientes.

K-mean reducido

El método de K-mean reducido (RKM, por sus siglas en inglés) crea centroides para los clústeres en un subespacio de menor dimensión. Después se minimiza la distancia entre los objetos del espacio completo y los 'cuasi' centroides en este subespacio, en donde representan el modelo ajustado para el RKM:

$$\mathbf{X} = \mathbf{UYA}' + \mathbf{E} \tag{1}$$

La función de perdida que se minimiza para obtener el resultado deseado se establece con:

$$f_{RKM}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{UYA}'\|^2. \tag{2}$$

Aquí, $\|\cdot\|$ denota la norma Frobenius. Mediante la minimización de la ecuación (2), se obtienen los valores para la matriz de cargas y la asignación de los objetos a los diferentes clústeres. En la práctica, este método es equivalente a maximizar la varianza entre los clústeres en el espacio reducido (Vichi *et al*., 2019).

K-mean factorial

El método k-mean factorial (FKM, por sus siglas en inglés) proyecta todos los objetos en un subespacio de menor dimensión. A partir del cual minimiza la distancia que va de los centroides hasta el subespacio de las proyecciones (Vichi y Kiers, 2001). La diferencia con el método RKM, es que sólo los centroides están en el subespacio. Para una mayor profundidad de esta comparación teórica vea Timmerman et al. (2010), en donde

se discuten las similitudes y diferencias de estos dos métodos. En esencia el modelo ajustado para FKM es (Vichi y Kiers, 2001):

$$\mathbf{XAA'} = \mathbf{UYA'} + \mathbf{E} \tag{3}$$

La función de perdida que de nuevo debe minimizarse para obtener el mejor ajuste de la ecuación (3), se da en (Vichi y Kiers, 2001):

$$f_{FKM}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \|\mathbf{XAA'} - \mathbf{UYA'}\|^2 = \|\mathbf{XA} - \mathbf{UY}\|^2. \tag{4}$$

Minimizando esta ecuación se deriva el ajuste óptimo de la matriz de carga y la asignación de los objetos a los clústeres. Efectivamente, el método FKM minimiza la suma de cuadrados de las distancias entre los puntos de los datos proyectados a sus respectivos centroides (Vichi *et al*., 2019).

Agrupación (clustering) y reducción de dimensión

El método agrupación y reducción de dimensión (CDR, por sus siglas en inglés) es un método que generaliza el análisis tándem, el RKM y el FKM. La función de pérdida se muestra en Vichi *et al*. (2019):

$$f_{CDR}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \alpha\|\mathbf{X} - \mathbf{XAA'}\|^2 + (1-\alpha)\|\mathbf{XA} - \mathbf{UY}\|^2. \tag{5}$$

En esencia, la función de pérdida del CDR es una combinación de la función de pérdida en el análisis tándem y la función de pérdida de FKM, la cual fue establecida en la ecuación (4). La función de pérdida en el análisis tándem es igual a:

$$f_{Tándem}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{FA'}\|^2, \tag{6}$$

en donde el **F** óptimo es igual a **XA** (Vichi *et al*., 2019).

Para $\alpha = 0$, la función de pérdida del análisis tándem desaparece y es igual a la función de pérdida de FKM.

Para $\alpha = 1$, la función de pérdida de FKM desaparece y se queda la función de pérdida del análisis tándem.

Una propiedad notable del modelo CDR es que si $\alpha = 0.5$ la función de pérdida coincide con la función de pérdida de RKM (Ecuación 2). Esto es:

$$\begin{aligned} f_{CDR,\alpha=0.5} &= 0.5(\|\mathbf{X} - \mathbf{XAA'}\|^2 + \|\mathbf{XA} - \mathbf{UY}\|^2) \\ &= 0.5(\|\mathbf{X}\|^2 + 2tr(\mathbf{A'X'XA}) + tr(\mathbf{A'X'XA}) - 2tr(\mathbf{X'UYA}) + \|\mathbf{UYA}\|^2) \\ &= 0.5(\|\mathbf{X} - \mathbf{XAA'}\|^2) = 0.5 f_{RKM} \end{aligned} \tag{7}$$

La multiplicación escalar de la función de pérdida RKM no influye en la solución óptima. El modelo permite variables nominales y/u ordinales, pero como no se usa en este trabajo se omite la explicación. Para más información del modelo CDR consulte Vichi *et al*. (2019).

Análisis de correspondencia y clúster

Supongamos que se tienen datos de n individuos sobre p variables categóricas recopilados en una matriz **Z** de super indicadores de dimensión $(n \times Q)$, donde $\mathrm{Q} = \sum_{j=1}^{p} q_j$.

La pertenencia al clúster se puede considerar como una variable categórica y se puede codificar usando una matriz de indicadores, llamada $\mathbf{Z}_k$. Al considerar la asociación de los clústeres con las variables categóricas, se construye una tabla de tabulación cruzada de pertenencia a los clústeres con las variables categóricas como $\mathbf{F} = \mathbf{Z}_K'\mathbf{Z}$, donde $\mathbf{Z}_K$ es la matriz $(n \times K)$ de indicadores de pertenencia a los clústeres. Aplicando AC a esta matriz produce valores de escala óptimos para filas (clústeres) y columnas (categorías) de tal manera que la varianza entre clústeres sea máxima. Esto es, los clústeres son separados óptimamente con respecto a la distribución de las variables categóricas. Similar y simultáneamente, las categorías con diferente distribución son óptimamente separadas de los clústeres.

Usando las definiciones presentadas se dice que

$$\mathbf{P} = \tfrac{1}{np}\mathbf{F} \tag{8}$$

De modo que para $\mathbf{P} - \mathbf{rc'}$, se obtiene

$$\mathbf{P} - \mathbf{P1}_Q\mathbf{1'}_K\mathbf{P} = \frac{1}{np}\left(\mathbf{F} - \frac{1}{np}\mathbf{F1}_Q\mathbf{1'}_K\mathbf{F}\right) = \frac{1}{np}\left(\mathbf{Z'}_K\mathbf{Z} - \frac{1}{n}\mathbf{Z'}_K\mathbf{1}_n\mathbf{1'}_n\mathbf{Z}\right) = \frac{1}{np}\mathbf{Z'}_K\mathbf{MZ},$$

donde $\mathbf{M} = \mathbf{I}_n - \mathbf{I}_n\mathbf{I'}_n/n$. Además, se define una matriz diagonal $\mathbf{D}_z$ tal que $\mathbf{D}_z\mathbf{1}_Q = \mathbf{Z'1}_n$ y deja $\mathbf{D}_k = \mathbf{Z'}_K\mathbf{Z}_K$, una matriz diagonal con tamaños de clúster. La función objetivo AC

$$\max\phi'_{\text{ca}}(\mathbf{B}) = \text{traza}\mathbf{B}'\mathbf{D}_C^{1/2}\overline{\mathbf{P}}'\overline{\mathbf{P}}\mathbf{D}_C^{1/2}\mathbf{B} \tag{9}$$

sujeto a

$$\mathbf{B}'\mathbf{D}_C\mathbf{B} = \mathbf{I}_K.$$

Y donde $\overline{\mathbf{P}}$ es la matriz de residuos estandarizados y $\mathbf{B}$ es una matriz de coordenadas fila y columna de rango k, donde k es la dimensionalidad de la aproximación. Y $\mathbf{D}_c$ es una matriz diagonal.
Para $\mathbf{P}$ definida en (8) y se convierte en:

$$\max\phi'_{\text{clusca}}(\mathbf{Z}_K, \mathbf{B}) = \frac{1}{np^2}\text{traza}\mathbf{B}'\mathbf{Z}'\mathbf{M}\mathbf{Z}_K\mathbf{D}_K^{-1}\mathbf{Z}'_K\mathbf{M}\mathbf{Z}\mathbf{B} \tag{10}$$

sujeto a:

$$\frac{1}{np}\mathbf{B}'\mathbf{D}_z\mathbf{B} = \mathbf{I}_k.$$

Teniendo en cuenta que con respecto a la notación típica AC, $\mathbf{D}_r = (1/n)\mathbf{D}_k$ y $\mathbf{D}_c = (1/np)\mathbf{D}_z$. Además, k denota la dimensionalidad de las cuantificaciones a escala óptima. Es decir, la dimensionalidad de la solución de AC. Este k debe ser elegido por el usuario y es menor que el rango de la matriz. Esto significa que, k es menor o igual a $\min(K-1, Q-1)$.
Al definir $\mathbf{B}^* = \frac{1}{\sqrt{np}}\mathbf{D}_z^{1/2}\mathbf{B}$, se puede re-escribir (10) como

$$\max\phi_{\text{clusca}}(\mathbf{Z}_K, \mathbf{B}^*) = \frac{1}{p}\text{traza}\mathbf{B}^{*\prime}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{Z}'\mathbf{M}\mathbf{Z}_K\mathbf{D}_K^{-1}\mathbf{Z}'_K\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^*, \tag{11}$$

Sujeto a $\mathbf{B}^{*\prime}\mathbf{B}^* = \mathbf{I}_k$.
Para $\mathbf{Z}_k$ fijo, la solución para $\mathbf{B}^*$ puede ser obtenida directamente de la descomposición del valor propio

$$\frac{1}{p}\mathbf{D}_z^{-1/2}\mathbf{Z}'\mathbf{M}\mathbf{Z}_K\boldsymbol{D}_K^{-1}\mathbf{Z}'_K\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-1/2} = \mathbf{B}^*\boldsymbol{\Lambda}^2\mathbf{B}^{*\prime}. \tag{12}$$

La solución apropiadamente escalada para la categoría se convierte así en:

$$\mathbf{B} = \sqrt{np}\mathbf{D}_z^{-1/2}\mathbf{B}^*. \tag{13}$$

Además de las cuantificaciones de categorías optimas, se necesita determinar el clúster óptimo de asignación $\mathbf{Z}_k$. Esto es, $\mathbf{Z}_k$ debe determinarse de tal manera que (11) sea un máximo. Como $\mathbf{Z}_k$ es una matriz de indicadores, esto no es un problema trivial. Sin embargo, para $\mathbf{B}^*$ fijo, este problema de optimización puede ser re-expresado como un problema de agrupamiento de k-means. Es decir, maximizar $\phi(\mathbf{Z}_K, \mathbf{B}^*)$ con respecto de $\mathbf{Z}_k$ es equivalente a resolver el siguiente objetivo de k-means:

$$\min\ \phi'_{\text{clusca}}(\mathbf{Z}_K, \mathbf{G}) = \left\|\sqrt{\frac{n}{p}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^* - \mathbf{Z}_K\mathbf{G}\right\|^2, \tag{14}$$

donde $\mathbf{G}$ es la matriz con las medias del clúster.
Para ver la equivalencia entre los dos objetivos, se tiene:

$$\mathbf{Y} = \sqrt{\frac{n}{p}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^*, \tag{15}$$

Y re-escribir el objetivo de k-medias (14) como:

$\min \phi'_{\text{clusca}}(\mathbf{Z}_k, \mathbf{G}) = \|\mathbf{Y} - \mathbf{Z}_K\mathbf{G}\|^2$.

Note que $\mathbf{Y}$ contiene las coordenadas del sujeto después de una escala óptima de las categorías. Resolviendo este problema de k-medias con respecto al rendimiento $\mathbf{G}$

$$\mathbf{G} = (\mathbf{Z}'_k\mathbf{Z}_K)^{-1}\mathbf{Z}'_K\mathbf{Y} = \mathbf{D}_k^{-1}\mathbf{Z}'_k\mathbf{Y}. \tag{16}$$

Insertando esto en el objetivo de k-medias se tiene

$\phi'_{\text{clusca}}(\mathbf{Z}_k, \mathbf{G}) = \|\mathbf{Y} - \mathbf{Z}_K\mathbf{G}\|^2 = \text{traza}\ \mathbf{Y}'\mathbf{Y} + \text{traza}\ \mathbf{G}'\mathbf{D}_K\mathbf{G} - 2\ \text{traza}\ \mathbf{G}'\mathbf{Z}'_K\mathbf{Y} = \text{traza}\ \mathbf{Y}'\mathbf{Y} +$
$\text{traza}\ \mathbf{Y}'\mathbf{Z}_K\mathbf{D}_K^{-1}\mathbf{D}_K\mathbf{D}_K^{-1}\mathbf{Z}_K{}'\mathbf{Y} - 2\ \text{traza}\ \mathbf{Y}'\mathbf{Z}_K\mathbf{D}_K^{-1}\mathbf{Z}_K{}'\mathbf{Y} = \text{traza}\ \mathbf{Y}'\mathbf{Y} - \text{traza}\ \mathbf{Y}'\mathbf{Z}_K\mathbf{D}_K^{-1}\mathbf{Z}_K{}'\mathbf{Y}$.

Por lo tanto, minimizar el objetivo de k-means con respecto a $\mathbf{Z}_k$ y $\mathbf{G}$ equivale a maximizar

$$\text{traza } \mathbf{Y}'\mathbf{Z}_K\mathbf{D}_K^{-1}\mathbf{Z}_{K'}\mathbf{Y} = n \text{ traza} \frac{1}{p}\mathbf{B}^{*\prime}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{Z}'\mathbf{M}\mathbf{Z}_K\mathbf{D}_z^{-1}\mathbf{Z}'_K\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\mathbf{B}^*, \quad (17)$$

sobre $\mathbf{Z}_k$ solamente, mientras que se sustituye (16) por $\mathbf{G}$, y por lo tanto es equivalente a (11). Por lo tanto, para $\mathbf{B}^*$ fijo, se puede encontrar una asignación al clúster $\mathbf{Z}_k$ aplicando el algoritmo de k-medias a $\mathbf{Y}$. La asignación resultante al clúster $\mathbf{Z}_k$ produce un valor mejorado para la función objetivo. Usando el nuevo $\mathbf{Z}_k$, se actualizan los valores de escala óptimos $\mathbf{B}^*$ usando (12). La iteración de este proceso produce el siguiente algoritmo para el clúster AC (Van de Velden *et al*., 2017:163):

1. Genera una asignación inicial al clúster $\mathbf{Z}_k$ (Por ejemplo, asignando sujetos al azar a los clústeres).
2. Encuentra cuantificaciones de categoría $\mathbf{B}^*$ usando (12).
3. Use (15) para construir una configuración inicial para los sujetos $\mathbf{Y}$.
4. Encuentre actualizaciones para $\mathbf{Z}_k$ aplicando la agrupación de k-medias a $\mathbf{Y}$ (usando (16) para obtener una matriz inicial de medias de los clústeres).
5. Repetir el procedimiento (es decir, volver al paso 2) usando $\mathbf{Z}_k$ para la matriz de asignación de clústeres hasta la convergencia. Es decir, hasta que $\mathbf{Z}_k$ (y por tanto $\mathbf{Y}$ y $\mathbf{G}$) permanezca constante.

La convergencia está garantizada ya que el valor de la función objetivo (11) nunca disminuye en pasos posteriores. Aunque no hay garantía de que el óptimo obtenido sea global. Se pueden utilizar inicios aleatorios para reducir las posibilidades de encontrar un óptimo local.
Note que la constante n, la cual aparece en (17), no afecta el óptimo clúster de asignación. Sin embargo, es conveniente vincular el problema de k-medias con el problema de AC. En particular, usando la fórmula de transición (9) en la cual se verifica facilmente que las medias de agrupamiento $\mathbf{G}$ obtenidas en el paso de k-medias en el algoritmo son idénticas a la matriz (principal) de coordenadas de fila obtenida del AC aplicado a $\mathbf{P}$ como se define en (8).
El método propuesto clúster AC puede verse como un ACM de tabulaciones cruzadas de pertenencia al clúster para variables categóricas. Las filas de la matriz de datos representan los clústeres y las coordenadas de fila obtenidas maximizan la varianza entre los clústeres. De (10), está claro que la solución para filas y columnas constituye un biplot de medias y atributos de clúster. Por lo tanto, las proyecciones de los puntos del clúster sobre vértices de atributos proporcionan aproximaciones al clúster por asociaciones de entre las categorías próximas. Las normalizaciones típicas de AC no conducen necesariamente a una distribución similar en los puntos fila y columna. En consecuencia, una visualización conjunta de puntos fila y columna no es muy informativa. Esto se puede reparar sin dañar la propiedad del biplot multiplicando las coordenadas de un conjunto por una constante y el otro conjunto por la inversa de esa constante. En el contexto de biplots existen algunas propuestas para hacer frente a tales problemas (Ver Gower *et al*., 2010; Gower *et al*., 2011). Aquí se propone usar una constante γ de tal manera que la desviación cuadrática media del origen es la misma en ambos conjuntos de puntos. Es decir, definir:

$$\mathbf{G}_s = \gamma\mathbf{G} \quad \text{y} \quad \mathbf{B}_s = \frac{1}{\gamma}\mathbf{B},$$

donde:

$$\gamma = \left(\frac{K}{Q} \text{ traza}\mathbf{B}'\mathbf{B}/\text{traza } \mathbf{G}'\mathbf{G}\right)^{1/4},$$

así que resulta:

$$\frac{1}{K}\text{traza } \mathbf{G}_s'\mathbf{G}_s = \frac{1}{Q}\text{traza } \mathbf{B}_s'\mathbf{B}_s.$$

Usar estas matrices traza de coordenadas reescaladas en lugar de las $\mathbf{G}$ y $\mathbf{B}$ originales facilita una visualización, la cual permite interpretar la composición del clúster a partir de las asociaciones con las categorías próximas.

3. RESULTADOS

De la aplicación de Clúster AC: La lectura literaria en la universidad

La lectura es una actividad que los estudiantes, los profesores y los trabajadores universitarios practican profusamente. Más generalmente, es posible aseverar que es una actividad cotidiana y fundamental: no se cuestiona si los estudiantes, académicos, funcionarios y trabajadores de las instituciones de educación

superior (IES) son lectores; es decir, se asume que deberían serlo. La lectura es reconocida como el medio para acceder al conocimiento, para lograr el aprendizaje y asimismo para sustentar el desarrollo de las competencias que garantizan la formación profesional. Pero cuando se habla así, se piensa en los textos académicos, de estudio y de trabajo, no en textos de literatura, no en textos que se lean por el puro placer de leerlos, por divertimento o esparcimiento.

Se ha reconocido que realizar cotidianamente la lectura por placer se considera un buen hábito; de hecho, se habla del hábito de la lectura, refiriéndose a la lectura que no es obligatoria. Se han argumentado muchos beneficios de este tipo de lectura, ya que la mueve una motivación de alto nivel, con lo que se activa la imaginación y se generan procesos fisiológicos e intelectuales que naturalmente generan el desarrollo de competencias asociadas a las habilidades del pensamiento, fundamentales para el desarrollo de las competencias académicas.

Jarvio Fernández y Ojeda Ramírez (2018) presentan un estudio para identificar la práctica de lectura de literatura en la comunidad de la Universidad Veracruzana, de Veracruz, México; analizan un conjunto de variables de una encuesta en la que se preguntó a estudiantes, académicos, funcionarios y trabajadores; clasificados por sexo (M y F), región (Xalapa y otra) y área académica (BA=Biológico-Agropecuaria, CS=Ciencias de la Salud, EA=Económico-Administrativa, HA=Humanidades y Artes NC=No procede y T= Técnica). Los autores presentaron (Jarvio Fernández y Ojeda Ramírez, 2018) una serie de análisis descriptivos e inferenciales y obtienen una serie de conclusiones que aquí no se retomarán. Para ilustrar el uso de la técnica de Clúster AC, se usarán, aparte de las variables independientes ya mencionadas, tres variables dependientes: acerca del gusto por la lectura de literatura (Poco, Regular o Mucho), acerca de qué tanto la practican (Poco, Regular y Mucho) y sobre el concepto que tienen de la lectura literaria. Esta última variable fue una pregunta abierta que se procesó usando análisis de datos textuales, identificando tres tipos de concepto: Deficiente (Def), Suficiente (Suf) y Excelente (Exc). La base de datos tiene la estructura que se muestra en la Tabla 1.

Tabla 1. Estructura de la base de datos de la ilustración del Clúster AC

>**str(L1)**	
'data.frame':	1311 obs. of 7 variables:
$ SEX:	Factor w/2 levels “F”, “M”: 1 1 1 1 1 1 1 2 1 2 …
$ REG:	Factor w/2 levels “Otra”, “Xalapa”: 2 2 1 1 1 1 1 1 1 1 …
$ OCUP:	Factor w/4 levels “Academico”, “Empleado”, …: 2 4 4 2 4 1 1 1 4 4 …
$ AREA:	Factor w/6 levels “BA”, “CS”, “EA”, …: 4 4 4 2 1 4 4 1 4 2 …
$ GUSTO:	Factor w/3 levels “Poco”, “Regular”, …: 3 3 2 3 3 3 3 1 3 2 …
$ DEDIC:	Factor w/3 levels “Poco”, “Regular”, …: 2 1 1 1 2 1 2 1 1 1 …
$ CONCEP:	Factor w/3 levels “Def”, “Suf”, “Exc”: 3 1 1 1 2 1 2 1 2 1 …

Al correr el procedimiento siguiendo las indicaciones dadas en Markos et al. (2019b) se obtiene un resumen que se muestra en la Tabla 2. Lo importante de destacar es la alta resolución que se obtiene en la resolución conjunta de la asociación entre variables y la agrupación de los individuos.

Tabla 2. Resumen de resultados de la salida del Clúster AC.

>**summary(CyAC)**

Solution with 3 clusters of sizes 541 (41.3%), 540 (41.2%), 230 (17.5%) in 2 dimensions.

Cluster centroids:

	Dim.1	Dim.2
Cluster 1	0.0163	-0.0162
Cluster 2	0.0007	0.0223
Cluster 3	-0.0400	-0.0145

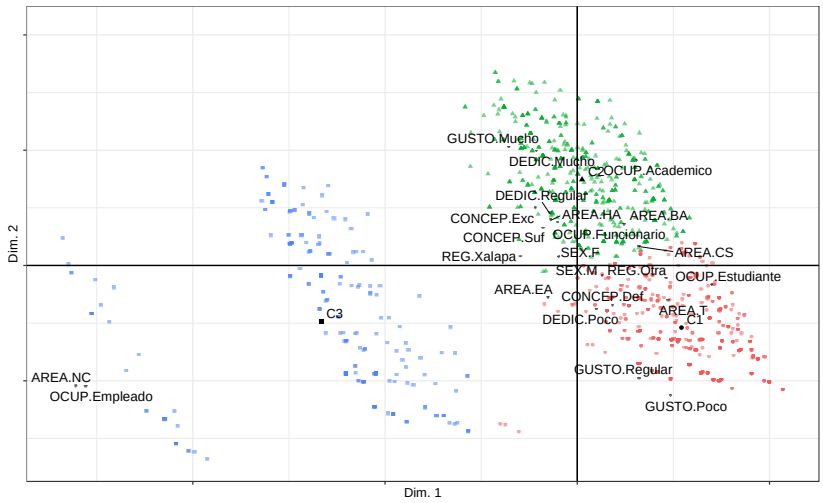
Variable scores:

	Dim.1	Dim 2									
1	-0.0029	0.0021;	2	0.0039	-0.0029	3	0.0139	-0.0035:	4	-0.0088	0.0022
5	0.0092	0.0272;	6	-0.0767	-0.0316	7	0.0210	-0.0052;	8	-0.0031	0.0111
9	0.0073	0.0105;	10	0.0097	0.0049	11	-0.0046	-0.0086;	12	-0.0041	0.0118
13	-0.0782	-0.0315;	14	0.0142	-0.0092	15	0.0146	-0.0339;	16	0.0097	-0.0295
17	-0.0107	0.0307;	18	0.0030	-0.0115	19	-0.0043	0.0135;	20	-0.0064	0.0296
21	0.0055	-0.0106;	22	-0.0053	0.0096	23	-0.0065	0.0148;			

Within cluster sum of squares by cluster:

[1] 0.0855, 0.0987, 0.1378

(between_SS / total_SS = 75.08 %)



En la Figura 1 podemos observar con claridad la formación de los clústeres, la cual se distingue por los colores de los puntos: El (1), en el que se puede apreciar sin problema que es el de los individuos periféricos (C3), que son los empleados, los cuales no tienen área académica de adscripción, son así mismo los que no tienen un comportamiento que pueda describirse en función de las variables dependientes; (2), el de "los lectores literarios" (C2), que son los funcionarios y académicos, de la región de Xalapa, de las áreas académicas de humanidades, artes y biológico-agropecuarias, y también de ciencias de la salud, y de sexo femenino; es aquí donde se **Figura 1. Mapa de Clúster AC de los datos de lectura de literatura en la UV.**

encuentran los que les gusta mucho la lectura, le dedican más tiempo y también tienen los mejores conceptos sobre la lectura; (3) el otro grupo, el de "los menos lectores" (C1) son mayoritariamente estudiantes, de sexo masculino, de otra región diferente a Xalapa, y de las áreas económico-administrativa y técnica; estos son los que se dedican poco a la lectura, que les gusta poco y que tienen conceptos deficientes de la misma. Esta descripción tiene una claridad diáfana, y permite focalizar acciones de promoción de la lectura en la institución.

De la aplicación de Clúster CP: la lectura y la escritura en los tecnológicos de Veracruz

Durante los meses de febrero y marzo de 2021 se realizó una encuesta sobre los hábitos de lectura y escritura entre los estudiantes y los académicos de 22 institutos tecnológicos del sistema nacional TecNM ubicados en el estado de Veracruz. En la Tabla 2 se muestra la estructura de la base de datos, y al pie la definición de cada una de las variables consideradas para este análisis. Salvo las variables de matrícula (MATR) y número de académicos (ACAD), todas las variables se presentan en porcentajes de respuesta.

Tabla 3. Estructura de la base de datos para el análisis de Cúster CP.

>**str(INST)**	
'data.frame':	22 obs. of 23 variables:
$ INST:	Factor w/22 levels "ACAY", "ALAMO", …:4 5 12 20 1 2 3 6 8 …
$ MATR:	int 2262 2773 4476 5968 1261 4415 2190 2909 490 5804 …
$ ACAD:	int 165 95 239 315 70 94 101 106 42 228 …
$ PRAC.P:	num 17.3 16.9 21.8 19.2 19.5 8.1 16.5 11.5 23.7 19.3 …
$ PRAC.M:	num 22.3 18.4 11.5 16.9 20.7 27.3 20.4 16.7 18.7 14.7 …
$ LEES.N:	num 4.6 3.8 7.6 4.9 7.6 4.2 5.2 4.9 4.5 5.6 ...
$ LEES.S:	num 26.1 25 22.9 22.6 25.8 36.1 22.4 26.8 23.2 25 ...
$ P.L:	num 25.8 25.3 23.7 26.2 23.7 15.1 25 22 28.6 25.2 ...
$ M.L:	num 12.2 10.3 9.9 8.8 9 15.1 9 17.1 6.3 8.5 ...
$ NO.LIT:	num 20.6 18 19.1 19.6 17.3 14.3 15.4 17.9 20.5 16.3 ...
$ L.AVAN:	num 6.2 9.6 3.8 5.5 9.2 11.8 5.1 9.8 8.1 5.1 ...
$ NO.PROM:	num 32.1 23.4 22.9 26.8 29.1 23.5 19.9 26 24.1 30 ...
$ PROM.S:	num 11 10.4 17.6 9.8 8.3 16.8 13.5 17.9 14.3 9.8 ...
$ IMPRE:	num 15 14 16.8 20.4 14.5 13.4 17.3 16.3 24.1 12.1 ...
$ DIGIT:	num 33.8 30.3 27.5 31.7 29.6 20.2 33.3 34.1 27.7 32.2 ...
$ ESC.P:	num 17.7 15.9 19.9 18.8 15.7 6.7 12.2 17.9 14.3 21.8 ...
$ ESC.M:	num 29.2 28.7 35.1 30 31 38.7 26.3 38.2 31.3 20.9 ...
$ E.DEF:	num 11.5 10.3 12.2 12.6 11.4 10.9 12.2 10.6 13.4 11.1 ...
$ E.EXC:	num 13.4 13.8 16 15.8 9.5 16 7.7 5.7 8.9 13.2 ...
$ ESC.FIS:	num 19.4 26 14.5 19.9 27.3 28.6 19.2 25.2 23.2 20.7 ...
$ ESC.DIG:	num 27.7 19.3 32.8 32.3 22.3 14.3 30.8 30.1 26.8 26.3 ...

$ LyE.P:	num 3.4 2.3 5.3 3.4 1.9 1.7 1.9 3.3 4.5 4.6 ...
$ LyE.M:	num 63.1 64.6 71 63.2 65.9 80.7 67.3 64.2 72.3 60.1 ...

MATR=Matrícula; ACAD=Número de académicos; PRAC.P=Practican poco la lectura; PRAC.M=Practican mucho la lectura; LEES.N=No leen; LEES.S=Leen semanalmente;
P.L=Leen poco fuera de lo obligatorio; M.L=Leen mucho fuera de lo obligatorio; NO.LIT=No leen literatura; L.AVAN=Lectores avanzados; NO.PROM=No promueven lo que leen; PROM.S=Promueven siempre; IMPRE=Leen fundamentalmente impreso; DIGIT=Leen fundamentalmente en digital; ESC.P=Escriben poco; ESC.M=Escriben mucho; E.DEF=Se consideran con escritura deficiente; E.EXC=Se consideran con escritura excelente; ESC.FIS=Prefieren escribir en físico; ESC.DIG=Prefieren escribir en digital; LyE.P=Consideran poco importante la lectura y escritura; LyE.M=Consideran muy importante la lectura y escritura.

Excepto MATR y ACAD, el resto de las variables están dados en porcentajes. Para ilustrar los datos se presenta en la Figura 2 el despliegue de estrellas, que tiene el propósito de mostrar las similitudes y diferencias entre los institutos comparados de manera visual y rápida.

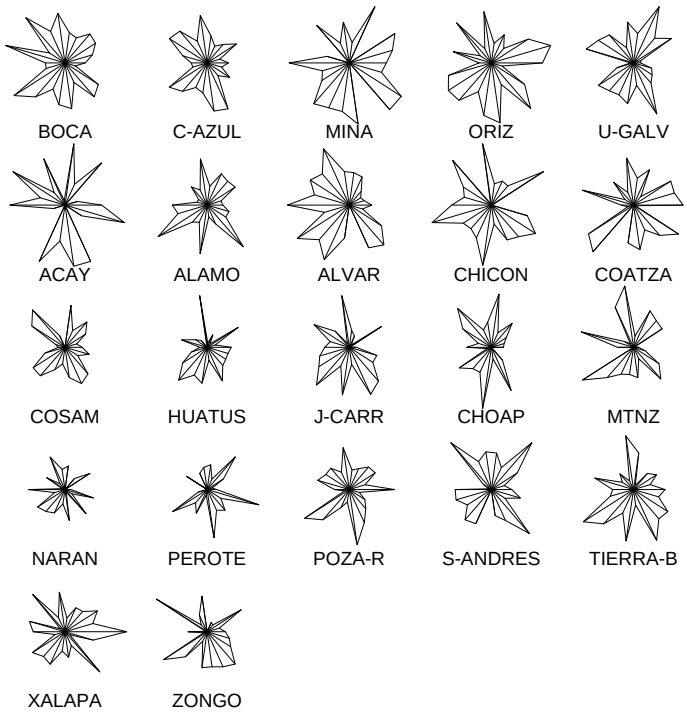


Figura 2. Despliegue de estrellas de los 22 institutos tecnológicos a partir de las 23 variables.

La salida de la corrida de Clúster CP es la que aparece en la Tabla 4.

Tabla 4. Resumen de resultados de la salida del Clúster CP.

>**summary(outINST)**
Solution with 3 clusters of sizes 14 (63.6%), 7 (31.8%), 1 (4.5%) in 2 dimensions. Variables were mean centered and standardized.

Cluster centroids:

	Dim.1	Dim.2
Cluster 1	0.1258	1.1420
Cluster 2	-1.4591	-2.2467
Cluster 3	8.4529	-0.2611

Variable scores:

	Dim.1	Dim 2
MATR:	0.0769	-0-4443
ACAD:	-0.0425	-0.4012
PRAC.P:	-0.2944	-0.0355
PRAC.M:	0.2208	-0.0573
LEES.N:	-0.0257	-0.2747
LEES.S:	0.2909	-0.0126
P.L:	-0.2533	-0.0566

M.L:	0.2140	-0.1814	
NO.LIT:	-0.1002	-0.0762	
L.AVAN:	0.2407	0.1967	
NO.PROM:	-0.0202	-0.3780	
PROM.S:	0.1979	-0.0948	
IMPRE:	0.0556	0.1388	
DIGIT:	-0.3456	0.0108	
ESC.P:	-0.3366	-0.0787	
ESC.M:	0.2642	-0.0990	
E.DEF:	0.0103	-0.0024	
E.EXC:	0.1735	-0.4473	
ESC.FIS:	0.1998	0.2343	
ESC.DIG:	-0.2854	-0.0532	
LyE.P:	-0.0633	-0.1853	
LyE.M:	0.3046	0.0381	

Within cluster sum of squares by cluster:
[1] 22.5376 13.1767 0.0000
(between_SS / total_SS = 79.7 %)
Clustering vector

BOCA	C-AZUL	MINA	ORIZ	U-GALV	ACAY	ALAMO	ALVAR	CHICON	COATZA
2	1	2	2	1	3	1	1	1	2

COSAM	HUATUS	J-CARR	CHOAP	MTNZ	NARAN	PEROTE	POZA-R	S-ANDRES	TIERRA-
1	1	1	1	2	1	1	2	1	1

XALAPA	ZONGO
2	1

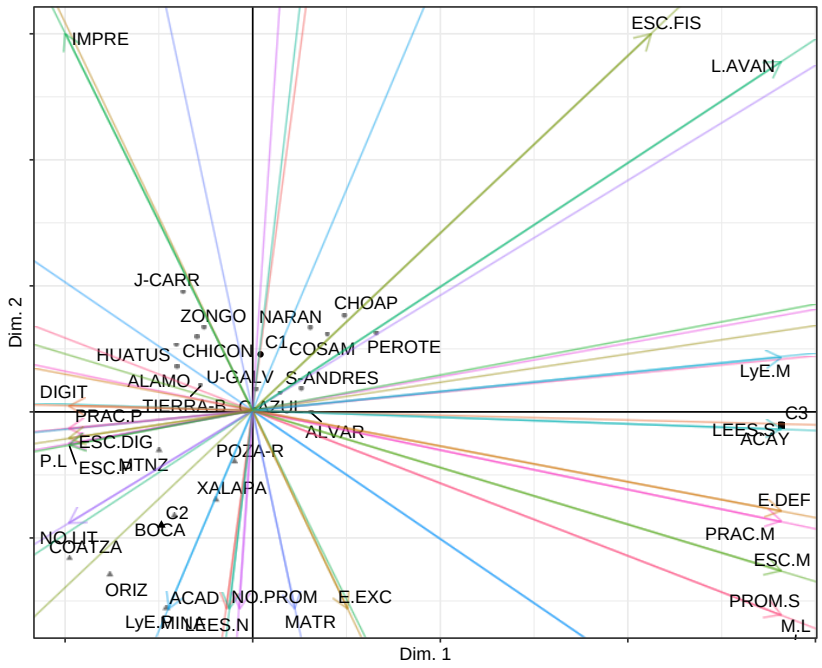


Figura 3. Biplot de las 23 variables que muestran los tres clústeres formados.
Nota: Los colores son elegidos unicamente para distinguir las variables.

El correspondiente biplot que muestra la agrupación y la asociación de las variables aparece en la Figura 3, en la que se puede detectar que el clúster marginal (C3), que se corresponde solamente con el instituto de Acayucan (ACAY), que está caracterizado porque tiene un desempeño muy destacado en la lectura y la escritura, salvo que un porcentaje importante que considera que su escritura es deficiente. Por otro lado, el clúster de Boca del Río (BOCA), Minatitlán (MINA), Orizaba (ORIZ), Coatzacoalcos (COATZA), Martínez (MTNZ), Poza Rica (POZA-R) y Xalapa (XALAPA) que se caracterizan por las dedicaciones y opiniones deficitarias; en general son institutos en los que se tienen bajos porcentajes.

4. CONCLUSIONES

Cuando el objetivo de segmentar un conjunto de unidades de estudio implica la caracterización de las asociaciones de las variables que determinan la agrupación el análisis multivariante conjunto es una alternativa inmediata. Cabe hacer notar que en Markos *et al*. (2019ª) aparecen otros ejemplos de uso de las librerías y comandos, que el lector podrá seguir con facilidad. La claridad de las salidas y la interpretación de

los despliegues gráficos hace que la utilidad de las librerías disponibles sea muy valorada. La aplicación de las técnicas de análisis multivariante conjunto ofrece una opción muy destacada para optimizar las estrategias de análisis de datos. Se recomienda ampliamente como una estrategia en la parte de análisis definitivos. Una ilustración adicional donde se presenta la estrategia completa de análisis se puede encontrar en Domínguez Reyes y Ojeda Ramírez (2021).

RECEIVED: OCTOBER, 2021.
REVISED: JANUARY, 2022.

REFERENCIAS

[1] DOMÍNGUEZ REYES, J. G. y OJEDA RAMÍREZ, M. M. (2021): Un estudio de la relación entre el capital escolar, situación laboral y la opinión de los egresados del posgrado: el caso de la Universidad Veracruzana. **Revista Latinoamericana de Políticas y Administración de la Educación**. 15, 117-129.
[2] GREENACRE, M. and BLASIUS, J. (Ed.) (1994): **Correspondence analysis in the social sciences**. London: Academic Press.
[3] GOWER, J. C., LUBBE, S. G. and LE ROUX, N. J. (2011): **Understanding biplots**. New York: Wiley.
[4] GOWER, J. C., GROENEN, P. J. F. and VAN DE VELDEN, M. (2010): Area biplots. **Journal of Computational and Graphical Statistics**, 19, 46–61.
[5] HWANG, H., DILLON, W. R. and TAKANE, Y. (2006): An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. **Psychometrika**, 71, 161–171.
[6] HUBERT, L., and ARABIE, P. (1985): Comparing partitions. **Journal of classification**, 2(1), 193-218.
[7] IODICE D'ENZA, A., VAN DE VELDEN, M. and PALUMBO, F. (2014): On joint dimension reduction and clustering of categorical data. In Vicari, D., Okada, A., Ragozini, G. and Weihs, C. (Eds.). **Analysis and modeling of complex data in behavioral and social sciences**. Switzerland: Springer International Publishing.
[8] JARVIO FERNÁNDEZ, A. O. y OJEDA RAMÍREZ, M. M. (2018): La lectura no utilitaria en la universidad en la era digital. Un análisis multivariante que ubica el texto impreso en la lectura de literatura. **Palabra Clave (La Plata),** 7(2), e051. https://doi.org/10.24215.
[9] KAMARRAS, M. (2020): **Parameter selection for clustering and dimension reduction.** Bachellor Thesis: Erasmus University.
[10] KRAZANOSWKI, W. J. (Ed.) (1995): **Recent advances in descriptive multivariate analysis**. United Kingdom: Oxford University Press.
[11] MARKOS A., IODICE D'ENZA A, & VAN DE VELDEN, M. (2019a): clustrd: Methods for joint dimension reduction and clustering. R package version 1.3.6-2. Available in https://CRAN.R-project.org/package=clustrd. **Consulted**, 10-7,2021.
[12] MARKOS, A., IODICE D'ENZA, A. and VAN DE VELDEN, M. (2019b): Beyond tandem analysis: Joint dimension reduction and clustering in R. **Journal of Statistical Software (Online)**, 91(10).
[13] TIMMERMAN, M. E., CEULEMANS, E., KIERS, H. A., and VICHI, M. (2010): Factorial and reduced K-means reconsidered. **Computational Statistics & Data Analysis**, 54(7), 18581871.
[14] VAN DE VELDEN M, IODICE D'ENZA A, and MARKOS, A. (2019): Distance-based clustering of mixed data. **Wiley Interdisciplinary Reviews: Computational Statistics**, 11(3), e1456.
[15] VAN DE VELDEN M, IODICE D'ENZA A, and PALUMBO F. (2017): Cluster correspondence analysis. **Psychometrika**, 82, 158–185.
[16] VICHI, M., and KIERS, H. A. L. (2001): Factorial k-means analysis for two-way data. **Computational Statistics and Data Analysis**, 37, 49–64.
[17] VICHI, M., VICARI, D. and KIERS, H. A. L. (2019): Clustering and dimension reduction for mixed variables. **Behaviormetrika**, 64, 243-269.