

---

# Estudio de las propiedades del método de clustering k-modes. Aplicación a la toma de decisiones empresariales en el marco del covid en URUGUAY

---

Diego Araújo

Mauro de la Vega

Ramón Álvarez-Vaz

Preprint

Nº5/22

2do semestre , 2022

Universidad de la República.

Facultad de Ciencias Económicas y de Administración,

Instituto de Estadística (IESTA)

Montevideo, Uruguay.



Esta obra está bajo una Licencia Creative Commons Atribución - NoComercial - CompartirIgual 4.0 Internacional.

Forma de citación sugerida para este documento:

Araújo, Diego. Dela Vega, Mauro Álvarez-Vaz, Ramón (2022). *Estudio del método de clustering k-modes aplicado a la toma de decisiones empresariales en el marco del covid en URUGUAY* (Serie Documentos de Trabajo; Nº5/22). Montevideo: Universidad de la República. Facultad de Ciencias Económicas y de Administración, Instituto de Estadística.

<https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/10518>

Diego Araújo <sup>1</sup>; Mauro de la Vega <sup>2</sup>; Ramón Álvarez-Vaz <sup>3</sup>

*Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República*

## Resumen

En el marco de un estudio en el campo de las finanzas corporativas, desarrollado durante la pandemia en Uruguay, acerca del IMPACTO COVID-19 EN LAS EMPRESAS URUGUAYAS en la toma de decisiones sobre reducción de personal, teletrabajo, capacitación y ambiente laboral entre otras se elabora una tipología de la toma de decisiones empresariales usando el método de clustering kmodos de perfil modal sobre variables de tipo binarias. Se analizan varios escenarios que dan cuenta de la cantidad de perfiles de toma de decisiones en base a la cantidad de grupos. En particular se estudia la variabilidad de algunas métricas que permiten evaluar la homogeneidad de los grupos y decidir una posible solución a la cantidad de grupos. Esa variabilidad se estudia en función del arranque el método , que es aleatorio en la elección de centros, viendo de ese modo la dependencia con la solución encontrada y también en la estructura de los datos por lo cual se trabaja con muestras de aprendizaje, variando el tamaño de la misma mediante remuestreo.

**Palabras clave:** Clustering, K-Modes, Remuestreo, Variables Categóricas.

**CÓDIGOS JEL:** .

**Clasificación MSC2010:** .

---

<sup>1</sup> *email:* [diego.araujo@fcea.edu.uy](mailto:diego.araujo@fcea.edu.uy), ORCID:<https://orcid.org/0000-0002-7303-2738>

<sup>2</sup> *email:* [maurodelavega@hotmail.com](mailto:maurodelavega@hotmail.com) , ORCID:

<sup>3</sup> *email:* [ramon.alvarez@fcea.edu.uy](mailto:ramon.alvarez@fcea.edu.uy), ORCID: <https://orcid.org/0000-0002-2505-4238>

## ABSTRACT

Within the framework of a study in the field of corporate finance, developed during the pandemic in Uruguay, about the COVID-19 IMPACT ON COMPANIES URUGUAYAS in decision-making on downsizing, telecommuting, training and work environment, among others, a typology of business decision making is elaborated using the modal profile k-modes clustering method on binary type variables. Several scenarios are analyzed that account for the number of decision-making profiles based on the number of groups. In particular, it studies variability of some metrics that allow evaluating the homogeneity of the groups and decide a possible solution to the number of groups. This variability is studied based on the bootstrap method, which is random in the choice of centers, thus seeing the dependency with the solution found and also in the structure of the data for which we work with learning samples, varying its size using resampling methods.

**Key words:** Categorical variables, Clustering, K-modes, resampling .

**JEL CODES:**.

**Mathematics Subject Classification MSC2010:**.

# 1. Introducción

Existen muchas situaciones en muy variadas disciplinas como la economía, el marketing, la epidemiología, donde la matriz de datos de la que se dispone está formada por datos binarios (unos y ceros) que surgen de trabajar con varias variables aleatorias resultantes de un experimento con 2 resultados posibles en cada caso. Muchas veces interesa analizar la relación entre variables y formar a su vez grupos que den cuenta de esas relaciones.

El trabajo aquí presentado deriva de una investigación que se realizó sobre una muestra de empresas uruguayas, de forma de conocer el impacto que tuvo la pandemia del Covid-19 sobre las finanzas de las mismas, las medidas tuvieron que implementar para poder dar continuidad a sus negocios, las opciones brindadas por el gobierno que fueron mayormente utilizadas, entre otros aspectos.

Para poder recolectar estos datos se utilizó un cuestionario de formato electrónico, en el cuál las respuestas a las preguntas enviadas se pueden codificar como unos y ceros ("Si / No"). Por lo que fue conveniente utilizar algún método de clustering para variables categóricas, de forma de crear el perfil modal de estas empresas y ver como quedaban segmentadas a través de distintos bloques de información.

A partir del tema inicial de la investigación, es de interés estudiar las propiedades del algoritmo en cuanto a su estabilidad bajo distintos escenarios, evaluando su rendimiento a través de distintas métricas y logrando aproximarnos a una cantidad  $k$  óptima de grupos, un problema común en este tipo de análisis. El algoritmo *kmodes* así como su similar para datos numéricos (*k-means*), se encuentra afectado por los arranques aleatorios y en este caso también por el tamaño de la muestra.

Es así que se proponen 2 escenarios:

- **Escenario 1:** Trabajar con una muestra de tamaño fijo y arranques aleatorios.
- **Escenario 2:** Haciendo remuestreo a distintos tamaños de muestra de aprendizaje (70 %, 80 % y 90 %) de la muestra original, para un  $k$  fijo.

Con respecto al ciclo de vida de este trabajo se deja constancia que un primer avance

---

de esta investigación fue presentada en junio de 2022 con un resumen extendido para la XV Semana Internacional de la Estadística y la Probabilidad, de la Facultad de Ciencias Físico Matemáticas de la Benemérita Universidad Autónoma de Puebla con número de preprint <https://doi.org/10.5281/zenodo.7145291>; se complementa posteriormente con el trabajo casi terminado previo a la elaboración de este DT, el que se presentó en las VI Jornadas Argentinas de Econometría de la FCE de la UBA, Buenos Aires, Argentina, en setiembre 2022 que aparece en . <https://doi.org/10.31219/osf.io/rwybh>.

Para asegurar la reproducibilidad de los resultados del análisis realizado, se dispone el código y datos utilizados en un repositorio público en la plataforma **Gitlab** al que se puede acceder a través de este <https://gitlab.com/GimeseIesta2/finanzas-de-empresas-iesta>.

La organización del documento es la siguiente: en la primera sección se introduce el problema bajo estudio,<sup>1</sup>, para luego presentar una breve reseña de los antecedentes ampliando la metodología empleada en ,<sup>2</sup>, en la tercera sección 3 las aplicaciones y resultados en 4 de los diversos análisis para finalmente presentarse las conclusiones y futuros pasos en la quinta sección, 5.

## 2. Metodología

Se utiliza parte de la metodología propuesta por Tsekouras *et al.* para clasificar atributos categóricos, a través del algoritmo mixto *Fuzzy C-modes* (10), y utilizada por Alvarez-Vaz y Massa para encontrar perfiles de infección parasitaria en escolares de Montevideo (1). En ambos trabajos cada individuo es previamente clasificado con algún método de clustering y luego pasa por una etapa de difusión, donde pasa a pertenecer a más de un cluster con diferentes grados de participación o membresía.

En el método original antes mencionado, se utilizaba el algoritmo *k-modes*, que era de tipo *modal*, y siendo un caso particular de un *k-prototipo* descrito por Huang (5). En este caso, el algoritmo tenía una lógica de funcionamiento similar a la del algoritmo *k-means*, y dada la naturaleza de las variables (binarias), era necesario el uso de otras medidas de disimilaridad, usando un método basado en frecuencias para actualizar los modos (11). Por lo tanto, del método mixto original planteado por Tsekouras *et al.* (10) en este trabajo solamente se usa el algoritmo *k-modes* que aplica la siguiente disimilaridad, siendo  $x_i, y_i$

2 individuos de los que se mide los atributos:

$$d(x_i, y_i) = \sum_1^m \delta(x_j, y_j) \quad (1)$$

con:

$$\delta(x_j, y_j) = \begin{cases} 0 & \text{si } x_j = y_j \\ 1 & \text{si } x_j \neq y_j \end{cases} \quad (2)$$

Lo anterior establece la suma de diferencias para cada par de elementos (o individuos)  $x_i, y_i$ , en cada  $j$  atributo, para tener una idea de la disimilaridad de los mismos. Cuanto más cercano a 0 el valor, más parecidos los elementos. En caso contrario, cuanto más cercano a  $m$  (la cantidad de atributos), más diferentes son.

El algoritmo trabaja en 4 pasos, de la siguiente manera:

1. Selecciona  $k$  modos iniciales, que actuarán como centroide de cada cluster, con  $x, y$  variables categóricas binarias en este caso;
2. Calcula la cantidad de disimilitudes de cada individuo con dichos modos y lo agrupará con el centroide que presente menor cantidad de diferencias (mayor similitud);
3. Luego de que todos los individuos han sido asignados, reestima la disimilaridad de los objetos contra el actual modo y si encuentra que un individuo tiene un modo de otro grupo que está más próximo lo reasigna, actualizando los modos de ambos grupos que se modificaron;
4. Se repite el paso 2 y 3 hasta que ningún individuo haya cambiado de cluster y haber visitado todo el conjunto de datos.

El resultado de este algoritmo es, entonces, una partición de los individuos en grupos cuyo representante es el perfil modal, es decir la combinación de respuestas que es más frecuente en cada cluster.

## 2.1. Métricas de validación

Se proponen 3 métricas para realizar la validación interna de los cluster y evaluar tanto la compacidad de los grupos (homogeneidad), como la separación entre ellos.

- **Diferencias Intracluster Absolutas (DIA):** Cantidad total de diferencias que hay dentro de cada cluster, de cada elemento con respecto al perfil modal del grupo, a través de la medida definida en (1). Para un cluster  $k$  dado, se podría definir como:

$$DIA(k) = \sum_{i=1}^{n_k} d(x_i, y_k) \quad (3)$$

dónde  $n_k$  es la cantidad de elementos en el cluster  $k$  y  $y_k$  el centroide del grupo (en este caso el modo o perfil modal).

- **Diferencias Intracluster Relativas (DIR):** Cantidad total de diferencias dentro de cada cluster, pero condicionadas a la cantidad de elementos del grupo. Para un cluster  $k$  dado, se define como:

$$DIR(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} d(x_i, y_k) \quad (4)$$

- **Coficiente de silueta (CS):** Métrica utilizada para calcular que tan similares son las observaciones o elementos en un mismo grupo en comparación con las observaciones de otros grupos.(9). Su valor oscila entre -1 (los grupos se asignan de forma incorrecta) y 1 (los grupos están bien separados entre sí y claramente diferenciados). Para una observación  $i$  se define como:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

donde  $a(i)$  es el promedio de disimilaridades de  $i$  con el resto de observaciones que pertenecen al mismo cluster que  $i$ , que se define como:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C, i \neq j} d(i, j) \quad (6)$$

dónde  $|C_I|$  es la cantidad de elementos en el cluster al cual pertenece  $i$  y  $b(i)$  es el promedio de disimilaridades de  $i$  con los elementos del cluster más cercano al cual  $i$  no pertenece (cluster “vecino”), definido como:

$$b(i) = \min_{i \neq k} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \quad (7)$$

dónde  $|C_k|$  es la cantidad de elementos del cluster vecino.

De (5) se puede observar que el coeficiente de silueta se puede expresar también como:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{si } a(i) < b(i) \\ 0 & \text{si } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{si } a(i) > b(i) \end{cases}$$

De lo anterior, se ve que el coeficiente queda definido en valores entre -1 y 1. Cuanto más cercano a 1, los grupos quedarían bien separados entre sí y claramente diferenciados. Esto se puede dar tanto por un valor de  $a(i)$  pequeño (lo que indicaría que el elemento es similar a otros en su grupo) o por un  $b(i)$  grande (lo que implica gran disimilaridad con el cluster vecino más cercano). Por el contrario, valores negativos cercanos a -1 indicarían que los elementos en los grupos se podrían estar asignando de forma incorrecta.

Esta última métrica en comparación con las primeras dos, da una información más íntegra del agrupamiento global, ya que no tiene en cuenta únicamente lo que sucede dentro del grupo, sino la separación a otros.

### 3. Aplicación

Los datos utilizados surgen de un estudio sobre las medidas implementadas en el contexto de la pandemia de Covid-19 a una muestra de empresas uruguayas y observar como fueron afectadas sus finanzas,(2),(3),(8). Para eso se elabora una aplicación Shiny [AppKmodes](#) que tiene un menú desplegable con 4 pestañas que permite interactuar a las usuarias y usuarios y poder

- **Pestaña 1** Desplegar todas las preguntas que forman parte del cuestionario
- **Pestaña 2** Tener un resumen descriptivo de las empresas de acuerdo a atributos de las mismas que dan cuenta del sector, la forma jurídica, si es de tipo familiar o no, la cantidad de empleados de cada empresa, el volumen de facturación
- los atributos de las personas que responden en cuanto a la propiedad de la empresa y el nivel educativo de las mismas
- **Pestaña 2** Donde poder elegir bloques de variables que se presentan en Tabla 1, y donde se puede controlar las variables que se usan para formar los clusters, la cantidad de grupos y decidir si se quieren visualizar los mismos ordenándolos de acuerdo a la frecuencia de respuestas y el arranque aleatorio del algoritmo
- **Pestaña 4** Donde se puede interactuar y cambiar el número de iteraciones para ver el comportamiento de las métricas .

## Impacto Covid-19 en Empresas Uruguayas

Análisis de cluster mediante k-modes

Desarrollado por Diego Araujo, Ramón Álvarez-Vaz (IESTA)



FACULTAD DE  
CIENCIAS ECONÓMICAS  
Y DE ADMINISTRACIÓN



INSTITUTO  
DE ESTADÍSTICA



UNIVERSIDAD  
DE LA REPÚBLICA  
URUGUAY

Menú Cuestionario Descriptiva Clustering Empresas Algoritmo kmodes

Variable	Pregunta
educacion_ceo	E - Indique su nivel de educación
actividad_emp	F - ¿En que sector se desempeña la empresa?
cant_personal	G - ¿Indique la cantidad de personal que emplea su empresa?
juridica_emp	H - ¿Cuál es la forma jurídica de la empresa?
factura_emp	I - Indique el rango de facturación de la empresa
exporta	J - La empresa exporta al exterior
familiar	K - Es empresa familiar
antiguedad_5	K - ¿La empresa tiene más de 5 años de antigüedad?
afectada_covid	1 - ¿La empresa fue afectada en su actividad diaria por la pandemia del COVID - 19
efecto_positivo	2 - ¿Considera que el efecto de la pandemia fue positivo para la empresa?

Figura 1: Aplicación Shiny AppKmodes



Figura 2: Aplicación Shiny AppKmodes con medidas descriptivas de las empresas

Los datos quedaron resumidos en los siguientes bloques:

- Efectos de la pandemia sobre las finanzas de las empresas (**EF**).
- Medidas implementadas por las empresas para sostener el negocio (**MI**).
- Medidas ofrecidas por el gobierno y utilizadas en las empresas (**MG**).
- Cambios tecnológicos que tuvieron que realizar las empresas (**CT**).

Lo que se presenta en este documento surge de trabajar sobre el bloque EF, que se detalla a continuación:

Variable	Descripción	Bloque	Tipo	Proporción
V1	Uso de indicadores económico-financieros	1	Estado financiero	55.4 %
V2	Flujo de fondos para evaluar liquidez	1	Estado financiero	70.5 %
V3	Afecta Flujo de Efectivo	2	Efecto en la empresa	64.3 %
V4	Afecta Volumen de Ventas	2	Efecto en la empresa	63.4 %
V5	Afecta Capital de Trabajo	2	Efecto en la empresa	50.9 %
V6	Afecta Ciclo de cuentas a cobrar	2	Efecto en la empresa	64,3 %
V7	Afecta Tiempo de abastecimiento Materias Primas	2	Efecto en la empresa	40.2 %

Tabla 1: Bloque de variables de efectos de la pandemia sobre las finanzas de la empresa.

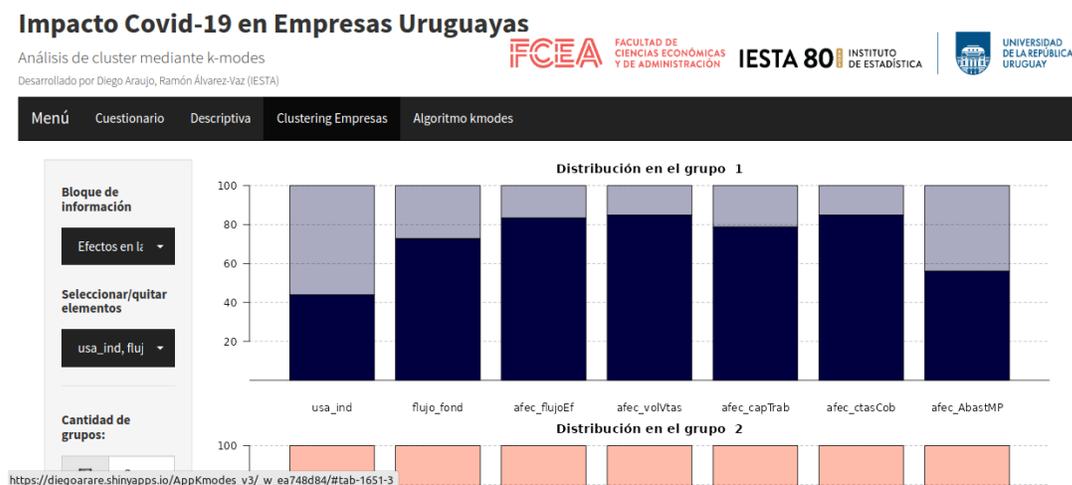


Figura 3: Aplicación Shiny AppKmodes con resultados para los diferentes tamaños de grupo  $n = 2$

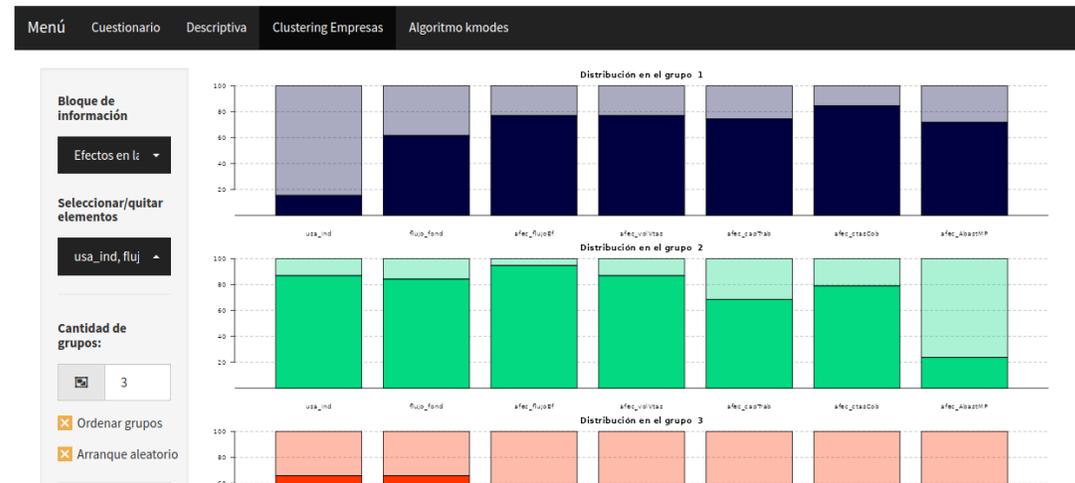


Figura 4: Aplicación Shiny AppKmodes con resultados para los diferentes tamaños de grupo ( $n = 3$ )

En cada uno de los escenarios descriptos, se evalúa el comportamiento de las distintas métricas propuestas, para observar su adecuación al problema y lograr determinar una cantidad de grupos “óptima”.

Un aspecto a tener en cuenta para el análisis es que se trabaja con una muestra de 700 empresas, de las cuales 112 de ellas (un 16%) respondieron a una encuesta electrónica donde se trató de cubrir todos los aspectos necesarios para la investigación, conociendo a su vez atributos de las empresas. Se analizó un eventual sesgo de respuesta (ver más detalles en el siguiente <https://gitlab.com/GimeseIesta2/finanzas-de-empresas-iesta> )

## 4. Resultados

Para el análisis global se trabaja con el software libre R (7), para la determinación de los clusters con el algoritmo presentado en la metodología se usa la librería `k1aR` (11) y el paquete `cluster` (6) para los coeficientes de silueta y paquete `ggplot2` (12) para la visualización.

Las visualizaciones siguientes son una forma de detectar rápidamente los patrones de respuesta dentro de cada grupo. El color más intenso representa la presencia de dicha

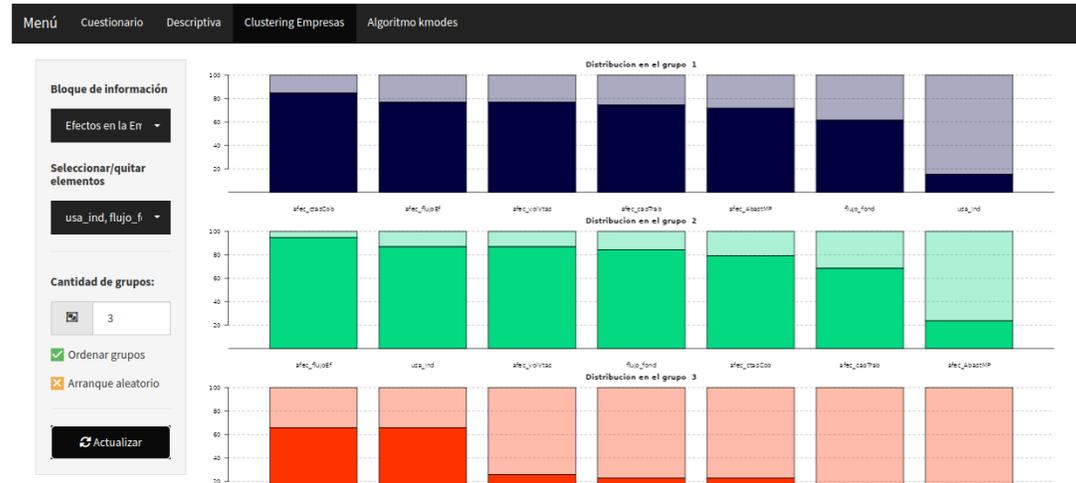


Figura 5: Aplicación Shiny AppKmodes con resultados con diferentes tamaños de iteraciones para  $(n = 2)$  cambiando orden de visualización de acuerdo a frecuencia de respuestas

variable dentro del cluster, es decir el porcentaje de respuestas afirmativas a la pregunta realizada (que se observa en cada columna), en contraste con el color más claro que hace referencia a la respuesta negativa o ausencia de la variable.

En la Tabla 2 se pueden apreciar las métricas de forma resumida, para cada uno de los  $k$  utilizados en el algoritmo. Se puede ver inicialmente que a medida que los elementos se desagregan en más grupos, la métrica de diferencias intracluster absolutas (DIA) disminuye, lo que indica inicialmente que es sensible al valor de  $k$ . Por lo que se podría tomar como referencia mayormente las métricas de diferencias relativas (DIA) y el coeficiente de silueta para determinar una cantidad  $k$  “óptima”.

El coeficiente de silueta, si bien es utilizado generalmente en clustering de variables numéricas, se puede adaptar para variables categóricas construyendo la matriz de disimilaridad en R a través de la función `daisy` (6) y especificando la distancia de *Gower*(4) como medida de disimilaridad. Este es un tipo de distancia para datos mixtos, donde utiliza la medida de disimilaridad mencionada en este trabajo para datos categóricos.

A través del análisis de esta métrica para los distintos valores de  $k$ , señala que el mayor valor (además de no presentar coeficientes negativos), se da para  $k = 2$ . Por lo que sería conveniente utilizar esta cantidad de grupos en las pruebas del escenario 2.

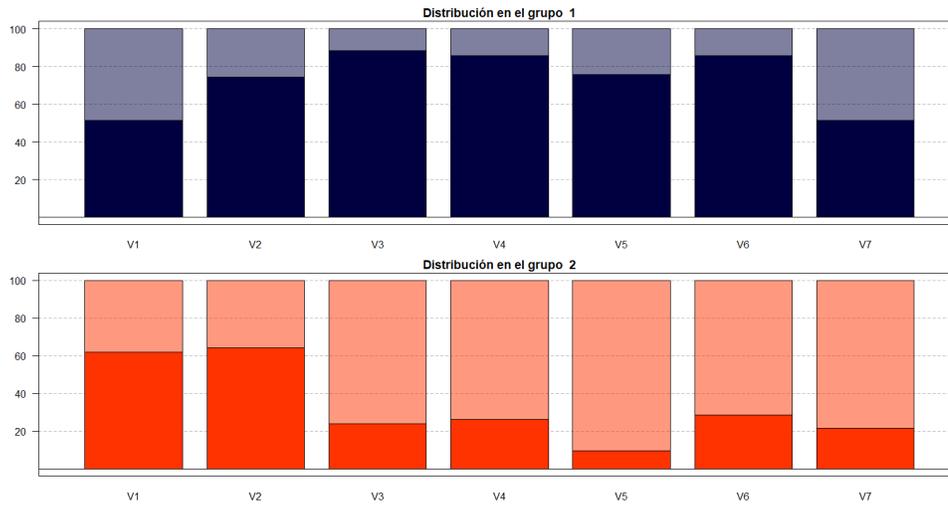


Figura 6: Proporción de las variables en cada grupo (2 grupos).

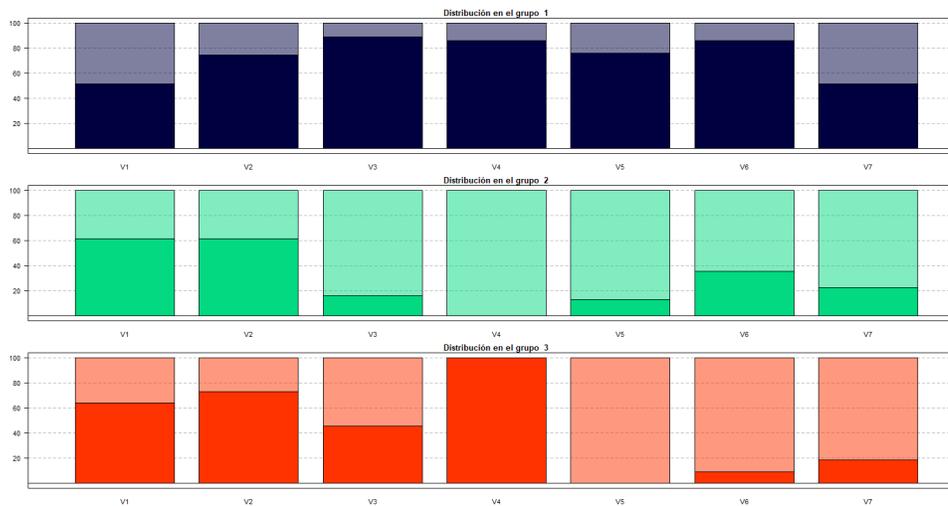


Figura 7: Proporción de las variables en cada grupo (3 grupos).

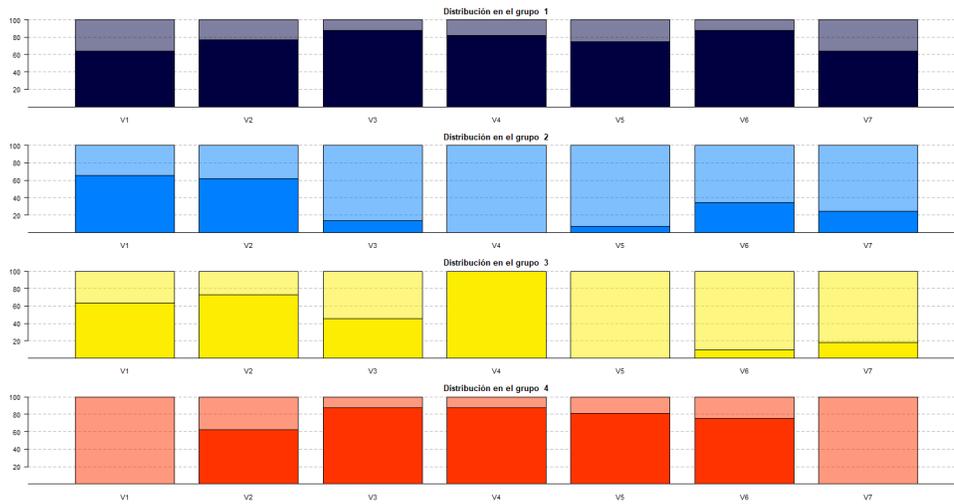
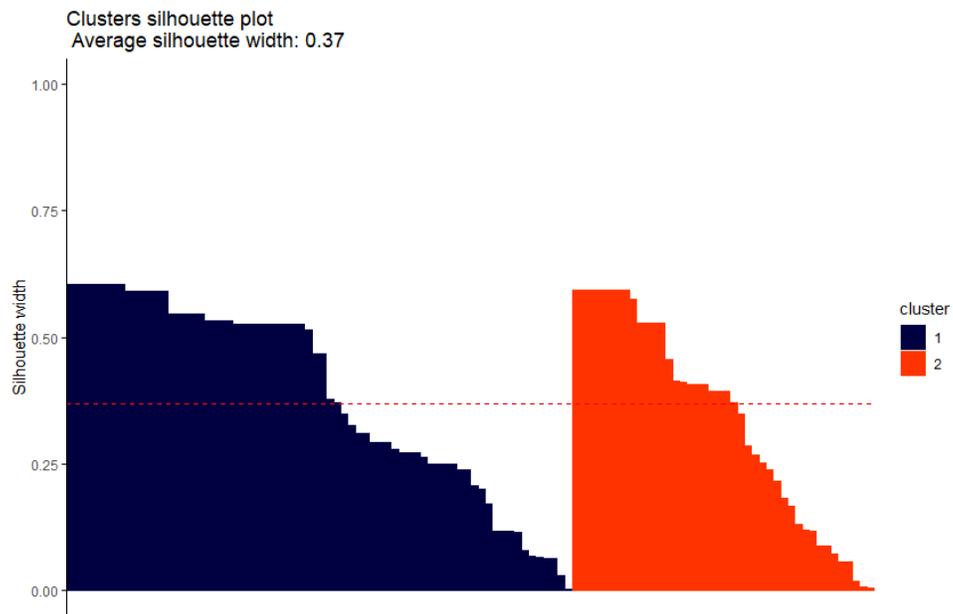
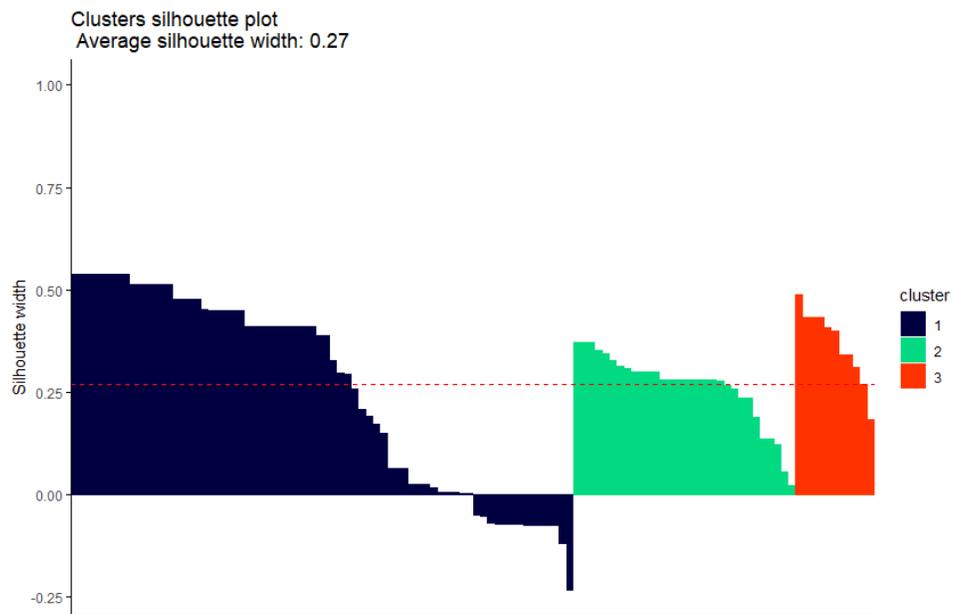


Figura 8: Proporción de las variables en cada grupo (4 grupos).

k-modes con 2 clusters	1	2		
tamaño	70	42		
withindiff (DIA)	131	77		
densidad (DIR)	1.87	1.83		
coef. silh. (CS)	0.39	0.34		
k-modes con 3 clusters	1	2	3	
tamaño	70	31	11	
withindiff (DIA)	131	51	15	
densidad (DIR)	1.87	1.65	1.36	
coef. silh. (CS)	0.26	0.26	0.37	
k-modes con 4 clusters	1	2	3	4
tamaño	56	29	11	16
withindiff (DIA)	91	44	15	17
densidad (DIR)	1.62	1.52	1.36	1.06
coef. silh. (CS)	0.08	0.29	0.27	0.38

Tabla 2: Métricas para validación de los grupos.

Figura 9: Coeficiente de silueta para  $k=2$ .Figura 10: Coeficiente de silueta para  $k=3$ .

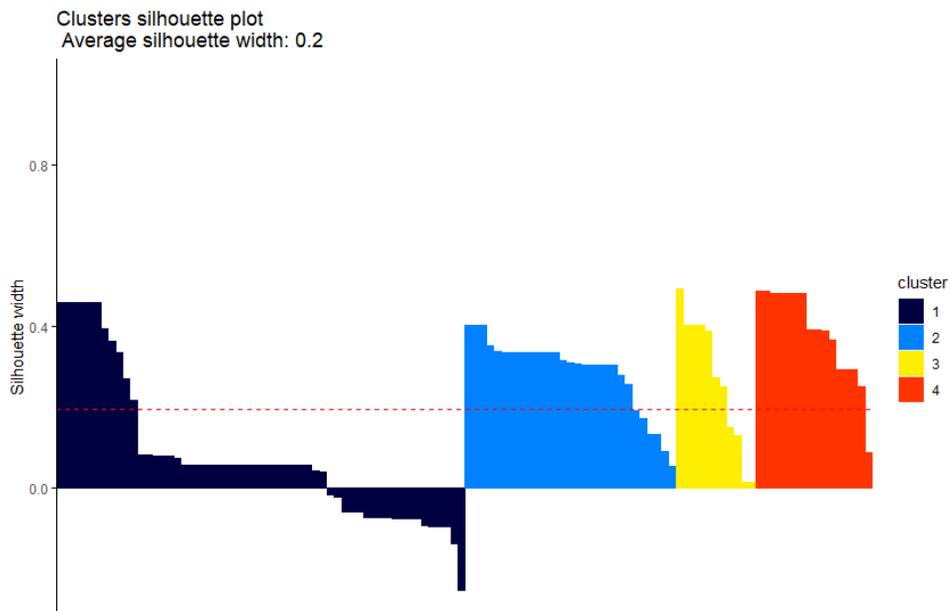


Figura 11: Coeficiente de silueta para  $k=4$ .

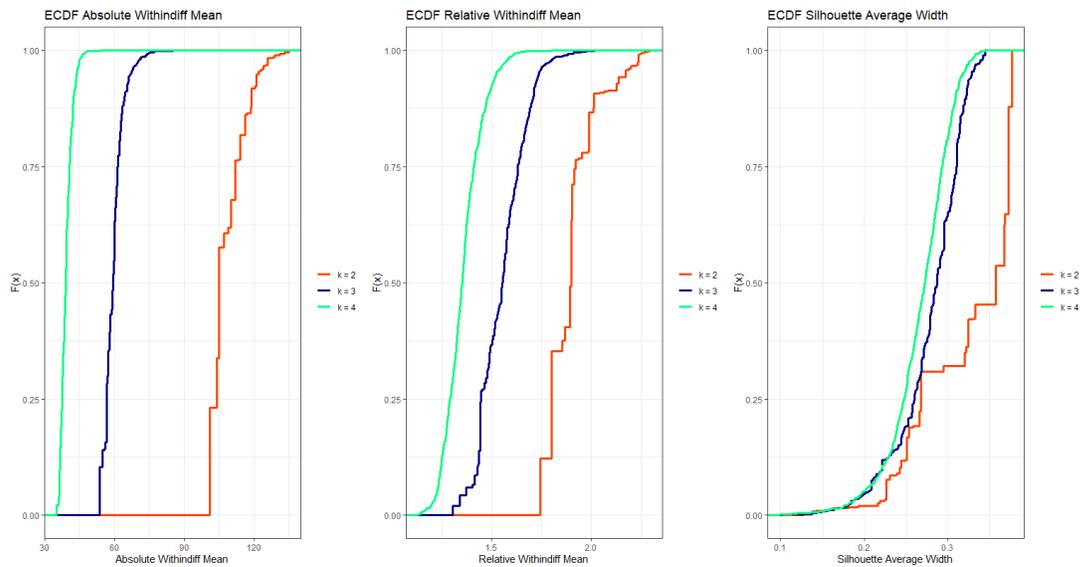


Figura 12: Escenario 1. Métricas para distintos  $k$  y arranques aleatorios. De izquierda a derecha: Dif. intracluster absolutas, relativas y coeficiente de silueta.

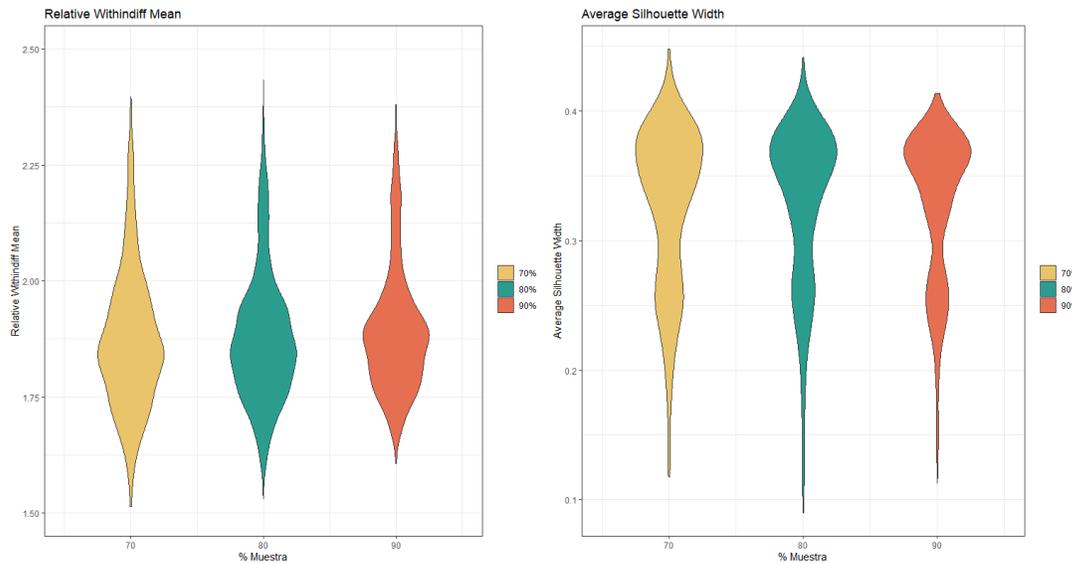


Figura 13: Escenario 2. Métricas para  $k=2$  y remuestreo a distintos niveles. Diferencias relativas (izquierda) y Coeficiente de silueta (derecha).

Los resultados encontrados con lo hecho hasta ahora permiten ver 2 aspectos del algoritmo de clusterización utilizado.

#### 4.1. Análisis Global

Se puede ver que a medida que se va desagregando en más grupos, la heterogeneidad va aumentando, medida a través de 3 métricas que muestra que se llega a un  $k$  'óptimo' en 2, decidiendo en base a la silueta (CS), ya que no hay empresas con valores negativos.

#### 4.2. Escenario 1

Para este escenario se prueban diferentes arranques (ya que es aleatorio), iterando 1000 veces y se ve que la densidad media de pares discordantes (medida local) es mayor para  $k = 2$ , mostrando mayor dispersión relativa y mayor discontinuidad, mientras que la silueta es mayor y presenta un comportamiento con un cambio marcado al llegar a valores de 0.28.

### 4.3. Escenario 2

Teniendo en cuenta que los datos utilizados son una muestra que tiene una tasa de respuesta baja (por lo tanto el tamaño es reducido), toma especial relevancia ver que tanta dependencia hay en los datos y de ahí el probar cambiando el tamaño de muestra. Para  $k = 2$  para la métrica (DIA), se observa que para un tamaño menor de muestra de aprendizaje la distribución está corrida a valores mas bajos pero con mas dispersión, mientras que a mayor tamaño de muestra de aprendizaje la variabilidad de (CS) es un poco menor y con menor dispersión en las iteraciones, casi bimodal.

## 5. Conclusiones y consideraciones finales

Se ha logrado utilizar varias métricas para el análisis de cluster y se observó el comportamiento del algoritmo bajo distintos escenarios. Lá métricas de diferencias intracluster absolutas, si bien nos puede brindar un primer acercamiento a la compacidad de los grupos, vemos que es susceptible al valor de  $k$ , siendo menor en la medida que  $k$  aumenta. Quizás una métrica mas fiable sea tener en cuenta las diferencias intracluster relativas (DIR) que están condicionados a la cantidad de elementos de cada grupo, o el Coeficiente de Silueta (CS), que para cada elemento brinda señales de cohesión a su grupo y de separación con el resto.

Como futuros pasos se propone:

- Estudiar el comportamiento sobre un conjunto con una cantidad mayor de observaciones y ver si estos resultados se confirman.
- Analizar el efecto de incluir o excluir variables con fuerte presencia en una gran cantidad de grupos (que presentan proporciones altas en cada cluster).
- Trabajar con variables categóricas multinominales (politómicas) y ver como funciona el algoritmo modal bajo diferentes esquemas de arranques aleatorios, evaluando su estabilidad como forma de ver la robustez del mismo y también tener una ida del impacto en el tamaño de la muestra de aprendizaje.

- Considerar diferentes algoritmos competitivos basados por ejemplos en medidas de divergencia apropiadas para datos categóricos.

## Referencias

- [1] Álvarez-Vaz, R. y Massa, F. (2012). Determinación de tipologías de infecciones parasitarias intestinales, en escolares mediante, técnicas de clustering sobre datos binarios. Documento de Trabajo Serie DT (12 / 05) - ISSN : 1688-6453, IESTA.
- [2] Brealey, R., Myers, S., y Allen, F. (2011). *Principles of Corporate Finance*. McGraw-Hill / Irwin., New York.
- [3] Graham, J. R. y Harvey, C. R. (2001). The theory and practice of corporate finance: Evidence from the field. *Journal of Financial Economics*, 60(2):187–243.
- [4] Gower, J. C. (1971) A general coefficient of similarity and some of its properties, *Biometrics* 27, 857–874.
- [5] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. in kdd: Techniques and applications. Technical report.
- [6] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2021). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2.
- [7] R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [8] Ross, S., Westerfield, R., y Jaffe, J. (2012). *Finanzas Corporativas*. McGraw-Hill/Interamericana Editores, México.
- [9] Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis *Journal of Computational and Applied Mathematics*, 20:53-65
- [10] Tsekouras, G., Papageorgiou, D., Kotsiantis, S., Kalloniatis, C., y Pintelas, P. (2005). Fuzzy clustering of categorical attributes and its use in analyzing cultural data. *World Academy of Science, Engineering and Technology*, 1:87–91.

- [11] Weihs, C., Ligges, U., Luebke, K., y Raabe, N. (2005). Klar analyzing german business cycles. En Baier, D., Decker, R., y Schmidt-Thieme, L., editores, *Data Analysis and Decision Support*, pp. 335–343, Berlin. Springer-Verlag.
- [12] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

# Instituto de Estadística

---

## Serie Documentos de Trabajo



FACULTAD DE  
CIENCIAS ECONÓMICAS  
Y DE ADMINISTRACIÓN

**IESTA 80** 80 años INSTITUTO  
DE ESTADÍSTICA



UNIVERSIDAD  
DE LA REPÚBLICA  
URUGUAY

Gonzalo Ramirez 1926, Piso 1, Oficina 23 - C.P. 11200 -  
Montevideo, Uruguay  
Teléfono: (598) 2410 2564  
<https://iesta.fcea.udelar.edu.uy/>  
Área Publicaciones

2do semestre, 2022

Nº 5/22