Dom. Cien., ISSN: 2477-8818
Vol. 9, núm. 3. Julio-Septiembre, 2023, pp. 1725-1741

DOI: 10.23857/dc.v9i3.3522

Ciencias Técnicas y Aplicadas
Artículo de Investigación

Comparación de técnicas estadísticas multivariadas usadas en datos cualitativos

Comparison of multivariate statistical techniques used in qualitative data

Comparação de técnicas estatísticas multivariadas usadas em dados qualitativos

Natalia Alexandra Pérez-Londo [I]
nperez@espoch.edu.ec
https://orcid.org/0000-0001-9068-8790

Darwin Saul Lema-Londo [II]
darwinlema@hotmail.es
https://orcid.org/0009-0002-7404-5052

Nataly Alejandra Batallas-Carrillo [III]
natalybatallas17@gmail.com
https://orcid.org/0009-0003-8969-7974

Rubén Antonio Pazmiño-Maji [IV]
ruben.pazmino@espoch.edu.ec
https://orcid.org/0000-0002-6811-7876

Correspondencia: nperez@espoch.edu.ec

*Recibido: 01 de junio de 2023 *Aceptado: 21 de julio de 2023 * Publicado: 21 de agosto de 2023

I. Máster Universitario en Estadística Aplicada, Docente ocasional en Escuela Superior Politécnica de Chimborazo; Riobamba, Ecuador.
II. Ingeniero Mecánico, Docente en Unidad Educativa Vigotsky; Riobamba, Ecuador.
III. Ingeniera en Estadística Informática, Investigadora Independiente; Santo Domingo, Ecuador.
IV. Doctor dentro del Programa de Doctorado en Formación en la Sociedad del Conocimiento, Docente en Escuela Superior Politécnica de Chimborazo; Riobamba, Ecuador.

Resumen

El avance de la estadística ha permitido el perfeccionamiento hacia técnicas estadísticas multivariadas con una gran cantidad de variables. Dichas técnicas permiten la interpretación, análisis y descripción simultánea de varias variables relacionadas a un conjunto de individuos, entonces resulta importante la comparación de las distintas técnicas estadísticas disponibles para evaluar los métodos más óptimos en cuanto al uso del tiempo y memoria del dispositivo. Por ello, en esta investigación se consideraron los métodos MHCPC (agrupamiento jerárquico en componentes principales), MSIM (agrupación jerárquica), Hrarchy (árbol de cohesión) y Hsimy (árbol de similaridad) y se plantearon los siguientes objetivos: comparar bajo el criterio de tiempo de ejecución, los métodos MHCPC, MSIM, Hrarchy y Hsimy para datos cualitativos, y comparar bajo el criterio de cantidad de memoria, los métodos MHCPC, MSIM, Hrarchy y Hsimy para datos cualitativos. Se utilizó un enfoque cuantitativo con un tipo de inferencia inductivo con un diseño experimental – pre experimental por la manipulación de variables de forma aleatoria. Los resultados muestran que se obtuvieron cinco grupos homogéneos, de los cuales en relación al tiempo de ejecución, se debe considerar la menor varianza por cuanto no existen diferencias significativas. Por su parte, respecto al uso de memoria en el grupo 1 se puede utilizar cualquiera de los cuatro métodos, pero para los grupos 2, 3, 4 y 5 los mejores métodos serían MHCPC o SIM porque ocupan menos memoria y son similares entre ellos.

Palabras clave: Técnicas multivariadas; datos cualitativos; agrupamiento jerárquico en componentes principales; agrupación jerárquica; árbol de cohesión; árbol de similaridad.

Abstract

The progress of statistics has allowed the improvement towards multivariate statistical techniques with a large number of variables. These techniques allow the simultaneous interpretation, analysis and description of several variables related to a group of individuals, so it is important to compare the different statistical techniques available to evaluate the most optimal methods in terms of the use of time and memory of the device. For this reason, in this investigation the methods MHCPC (hierarchical grouping in principal components), MSIM (hierarchical grouping), Hrarchy (cohesion tree) and Hsimy (similarity tree) were considered, and the following objectives were set: to compare under the criterion of execution time, the MHCPC, MSIM, Hrarchy and Hsimy methods for qualitative data, and compare under the criterion of amount of memory, the MHCPC, MSIM, Hrarchy and Hsimy methods for qualitative data. A quantitative approach was used with a type of inductive

inference with an experimental - pre-experimental design by manipulating variables randomly. The results show that five homogeneous groups were obtained, of which in relation to execution time, the lowest variance should be considered since there are no significant differences. For its part, regarding the use of memory in group 1, any of the four methods can be used, but for groups 2, 3, 4 and 5 the best methods would be MHCPC or SIM because they take up less memory and are similar to each other.

Keywords: Multivariate techniques; qualitative data; hierarchical grouping into principal components; hierarchical grouping; cohesion tree; similarity tree.

Resumo

O progresso da estatística permitiu o avanço para técnicas estatísticas multivariadas com um grande número de variáveis. Estas técnicas permitem a interpretação, análise e descrição simultâneas de diversas variáveis relacionadas com um grupo de indivíduos, pelo que é importante comparar as diferentes técnicas estatísticas disponíveis para avaliar os métodos mais óptimos em termos de utilização de tempo e memória do dispositivo. Por este motivo, nesta investigação foram considerados os métodos MHCPC (agrupamento hierárquico em componentes principais), MSIM (agrupamento hierárquico), Hrarchy (árvore de coesão) e Hsimy (árvore de similaridade), e foram traçados os seguintes objetivos: comparar sob o critério de tempo de execução, os métodos MHCPC, MSIM, Hrarchy e Hsimy para dados qualitativos, e comparar sob o critério de quantidade de memória, os métodos MHCPC, MSIM, Hrarchy e Hsimy para dados qualitativos. Utilizou-se uma abordagem quantitativa com tipo de inferência indutiva com delineamento experimental - pré-experimental por meio da manipulação de variáveis aleatoriamente. Os resultados mostram que foram obtidos cinco grupos homogêneos, dos quais em relação ao tempo de execução, deve-se considerar o de menor variância, pois não há diferenças significativas. Por sua vez, em relação ao uso de memória no grupo 1, qualquer um dos quatro métodos pode ser usado, mas para os grupos 2, 3, 4 e 5 os melhores métodos seriam MHCPC ou SIM porque ocupam menos memória e são semelhantes aos uns aos outros.

Palavras-chave: Técnicas multivariadas; dados qualitativos; agrupamento hierárquico em componentes principais; agrupamento hierárquico; árvore de coesão; árvore de semelhança.

Introducción

Con el tiempo, el establecimiento de nuevos modelos multicausales, junto con la búsqueda de métodos de estimación más eficaces, el desarrollo de nuevas pruebas no paramétricas y el uso de técnicas multivariadas han ido perfeccionando la metodología estadística. Estos avances han contribuido a mejorar la capacidad de análisis estadístico. En particular, las técnicas multivariadas han evolucionado para adaptarse al análisis de grandes volúmenes de datos y procesos iterativos. Dichas técnicas han sido fundamentales en la evolución de la metodología estadística, especialmente en el análisis de grandes masas de datos. Estas técnicas han proporcionado herramientas más avanzadas para el análisis de datos complejos y han mejorado la capacidad de los investigadores para obtener resultados más precisos y significativos (Sagaró del Campo y Zamora, 2019a).

Además, el uso de técnicas estadísticas multivariadas se ha vuelto cada vez más relevante debido a que la mayoría de los análisis de datos involucran preguntas complejas que abarcan más de dos variables. Como señala Pulido (2020), estas técnicas permiten abordar de manera más efectiva estos cuestionamientos al examinar las interacciones entre múltiples variables y proporcionar una comprensión más completa de las relaciones subyacentes.

Los avances en el campo de la estadística han tenido un impacto considerable en diversas áreas de la investigación, influyendo en el enfoque y las prácticas utilizadas en la recopilación y análisis de datos. La inclusión de técnicas estadísticas y estrategias de ciencia de datos ha proporcionado a los investigadores una comprensión más profunda de los fenómenos y ha mejorado la forma en que analizan los datos (Chan y Galli, 2020). La introducción de técnicas estadísticas más sofisticadas ha permitido a los investigadores abordar preguntas más complejas y explorar relaciones más sutiles entre las variables. Además, el avance de la ciencia de datos ha facilitado la gestión y el análisis de grandes volúmenes de datos, lo que ha llevado a descubrimientos más profundos y conclusiones más sólidas.

Como lo expresan Chan y Galli (2020), en vista de la creciente complejidad del panorama del conocimiento en la actualidad, resulta imperativo ampliar el horizonte de las perspectivas mediante el uso de técnicas multivariadas. Al considerar múltiples variables y fuentes de datos, se pueden obtener perspectivas más holísticas y enriquecedoras sobre el objeto de estudio. En un estudio real, es común que el investigador tenga a su disposición una gran cantidad de variables medidas u observadas en un grupo de individuos y su objetivo es estudiarlas en conjunto para ello cuenta con múltiples técnicas estadísticas multivariadas (Hidalgo, 2019). Por lo antes mencionado, Pulido (2020)

indica que para los datos categóricos existen técnicas como el análisis de conglomerados, análisis implicativo y análisis de correspondencias que permiten diagnosticar la elección del método apropiado para resolver el problema.

El objetivo de comparar técnicas estadísticas multivariadas además de representar de manera simplificada las relaciones causales o predecir, en base a lo observado, dentro de ciertos márgenes de error, facilita encontrar la técnica más eficiente en términos de espacio y tiempo que respondan a los criterios de un margen de error predeterminado. A su vez, las técnicas multivariadas no solo posibilitan el análisis causal del problema en análisis, sino que la comparación entre las distintas técnicas existentes favorece a eliminar posibles sesgos y llegar a resultados válidos, muy apegados a la realidad (Sagaró del Campo y Zamora, 2019a).

El uso de estas técnicas estadísticas multivariadas brinda la oportunidad de ampliar el horizonte y los alcances de los estudios para comprender de manera más profunda el fenómeno de interés a través del tipo de relaciones que pueden establecerse entre las variables. La comparación de ciertas técnicas estadísticas a través de los métodos utilizados ayuda a entender la lógica estadística aplicada para asegurar que las conclusiones sean válidas y las interpretaciones correctas (Chan y Galli, 2020). Al respecto, Pulido (2020) añade que se puede reducir el sesgo al momento de tomar decisiones basadas en evidencias sólidas. Por otra parte, Naranjo et al. (2018), mencionan que al utilizar la técnica óptima se minimiza el espacio de memoria y se reduce el tiempo de procesamiento.

Aunque actualmente, existen muchos programas estadísticos, tanto pagados como gratuitos, que realizan este tipo de modelos, sin embargo, se requiere de un análisis estadístico exhaustivo para interpretar y dar sentido a los resultados (Pulido, 2020). Indiferentemente, la aplicación de técnicas estadísticas por sí solas no resulta eficiente para comprender correctamente los fenómenos estudiados, requiere comprender los fundamentos teóricos y metodológicos de las técnicas estadísticas utilizadas, así como los supuestos y limitaciones asociados con ellas. Es necesario contar con un conocimiento más amplio sobre cómo seleccionar y aplicar las técnicas apropiadas para abordar las preguntas de investigación y los objetivos del estudio (Chan y Galli, 2020).

Las técnicas estadísticas multivariadas cobran especial importancia principalmente en las ciencias sociales y económicas debido a que resume y sintetiza grandes conjuntos de variables en una estructura más manejable y comprensible, todo ello en función a los objetivos de estudio, es decir el análisis multivariado examina cómo las variables se relacionan entre sí y contribuyen conjuntamente para entender y explicar el fenómeno (Álvarez-Perdomo et al., 2022). Esto proporciona una visión

más completa del fenómeno en estudio. El conjunto de técnicas estadísticas multivariadas proporciona una visión multidimensional de la realidad estudiada (Chan y Galli, 2020). Estas técnicas amplían el análisis más allá de una sola variable o una relación de dos variables, posibilitando el examen de las interacciones entre múltiples variables en un conjunto de datos.

Las técnicas estadísticas multivariadas comprenden un conjunto de métodos estadísticos cuyo fin es analizar simultáneamente conjuntos de datos multivariantes, es decir, aquellos en los que se registran múltiples variables para cada individuo u objeto estudiado. Estas técnicas permiten obtener un mejor entendimiento del fenómeno en estudio al proporcionar información que los métodos estadísticos univariantes y bivariantes no pueden obtener (Vargas et al., 2020).

La técnica estadística multivariada "permite la interpretación, análisis y descripción simultánea de varias características o atributos (variables) sobre un conjunto de individuos (objetos o unidades de análisis) correspondientes a una misma variedad designada como población de interés" (Chan y Galli, 2020, pág. 126). Las técnicas multivariantes encuentran la asociación entre variables, mediante el uso de tabulación cruzada correlación parcial y regresiones múltiples. Una de las ventajas destacadas de las técnicas multivariadas es su capacidad para manejar, visualizar e interpretar grandes bases de datos, lo que contribuye a una mejor comprensión de la complejidad de los fenómenos estudiados. Esto amplía el alcance de la inferencia estadística, permitiendo de esta manera un análisis más completo y profundo (Chan y Galli, 2020).

Según la literatura se ha encontrado que los métodos más empleados en las técnicas estadísticas multivariadas son el agrupamiento jerárquico en componentes principales, la agrupación jerárquica, el árbol de cohesión y el árbol de similaridad.

MHCPC (agrupamiento jerárquico en componentes principales)

Según Maugeri et al. (2021), el enfoque de agrupamiento jerárquico en componentes principales combina tres métodos estándar: PCA (Análisis de Componentes Principales), agrupamiento jerárquico y el algoritmo de k-medias. Donde, el agrupamiento jerárquico es una técnica que organiza los datos en grupos o clústeres basados en su similitud, mientras que el algoritmo de k-medias es un método de agrupamiento que asigna los datos a "k" grupos predefinidos. Con respecto a ello, se utiliza PCA para reducir la dimensionalidad de los datos, luego se aplica el agrupamiento jerárquico y el algoritmo de k-medias para obtener grupos más precisos y significativos. La combinación de estos tres métodos tiene como objetivo obtener una mejor solución de agrupamiento.

Este método es útil para desarrollar sistemas de clasificación incorporados y su importancia radica en que no se necesita especifica previamente el número óptimo de clústers “k”. El agrupamiento jerárquico permite explorar las relaciones existentes entre las agrupaciones basado en dos principios: a) pequeña variabilidad dentro de la clase: si los individuos en la misma clase están cerca unos de otros, b) gran variabilidad entre clases: si los individuos en diferentes clases están lejos unos de otros (Koh et al., 2022).

Este método combina la reducción de la dimensionalidad con el agrupamiento para obtener grupos más significativos y comprensibles. Al aplicar el PCA, se pueden identificar las principales dimensiones de variabilidad en los datos, lo que ayuda a simplificar la estructura subyacente. Luego, el agrupamiento jerárquico y el algoritmo de k-medias se utilizan para asignar los datos a grupos en función de su similitud. Esta combinación permite obtener una solución de agrupamiento más precisa y efectiva, facilitando la exploración y el análisis de grandes conjuntos de datos.

MSIM (agrupación jerárquica)

Las técnicas de agrupación jerárquica son métodos utilizados en el análisis de datos para encontrar grupos o clústeres en un conjunto de datos. Estos métodos funcionan de manera recursiva, es decir, se aplican de forma repetida para crear grupos anidados dentro de otros grupos. Existen dos enfoques principales en la agrupación jerárquica: el aglomerativo y el divisivo (Govender y Sivakumar, 2020).

En la agrupación aglomerativa, cada punto de datos se considera inicialmente como un clúster individual y luego se van fusionando sucesivamente los pares de clústeres más similares, lo que resulta en una jerarquía de clústeres cada vez más grandes. Al final del proceso, todos los puntos de datos estarán en un solo clúster (Govender y Sivakumar, 2020). La agrupación jerárquica aglomerativa es un método utilizado en análisis de datos y aprendizaje automático para agrupar objetos o muestras en clústeres o grupos. La característica principal de este enfoque es que genera una secuencia anidada de particiones o divisiones de los datos. En este sentido, se van fusionando los grupos en base a la similitud o proximidad entre ellos, formando así agrupaciones más grandes. Este proceso continúa iterativamente, fusionando grupos similares hasta que todos los objetos se encuentren en un solo grupo o clúster (Martín del Campo, 2019).

Por otro lado, en la agrupación divisiva todos los puntos de datos se agrupan inicialmente juntos en un solo grupo y luego se dividen repetidamente en grupos más pequeños. En cada paso, se selecciona un grupo existente y se divide en subgrupos basados en ciertos criterios. Este proceso

continúa hasta que cada punto de datos se encuentra en su propio clúster individual (Govender y Sivakumar, 2020). Un agrupamiento jerárquico divisivo se representa un conjunto de datos mediante un árbol enraizado. En este árbol, cada hoja del árbol representa un punto de datos individual, mientras que cada nodo interno representa un grupo que contiene sus hojas descendientes. El agrupamiento jerárquico es un proceso de particionar un conjunto de datos en grupos o clústeres de forma recursiva, donde cada paso de la partición se realiza en un nivel de detalle más fino (Cohen-Addad et al., 2019).

Hrarchy (árbol de cohesión)

El árbol cohesivo, también conocido como árbol de cohesión, es una representación gráfica que se utiliza en el análisis estadístico implicativo para visualizar las reglas generadas. La cohesión juega un papel fundamental al estructurar el conocimiento en forma de reglas y meta-reglas. A diferencia de una estructura lineal y simétrica, el árbol cohesivo es no lineal, asimétrico, jerárquico y dinámico, lo que permite ir más allá de la simple articulación de las partes de una tipología clásica. Su objetivo es alcanzar un todo significativo, proporcionando una representación visual que muestra cómo las reglas se relacionan entre sí y forman un sistema coherente (Sagaró del Campo y Zamora-Matamoros, 2019b).

La jerarquía cohesiva se refiere a la estructura jerárquica que se obtiene utilizando el software CHIC (Cohesive Hierarchical Clustering) mediante un enfoque basado en la implicación para agrupar éxitos o elementos similares en conjuntos coherentes. Cuando se crea una jerarquía cohesiva con CHIC, se espera que respete el orden taxonómico presunto dentro de los grupos formados. Esto significa que los elementos que comparten características más similares deberían agruparse más cerca entre sí en la jerarquía, reflejando así una relación taxonómica más cercana. y para el análisis de los datos se utiliza el análisis estadístico implicativo (Fotiadis y Anastasiadou, 2019).

El árbol cohesivo se construye de manera ascendente y está organizado en clases ajustadas y orientadas. Cada nivel del árbol representa una etapa en la que se calcula el índice de cohesión entre cada par de clases ordenadas del nivel anterior. Dicho índice evalúa la fuerza de la consistencia de las variables que están inmersas en la creación de una nueva clase. A medida que se avanza en los niveles del árbol, las clases anteriores se reemplazan y se forma una nueva clase que combina y reúne a las dos clases anteriores. Este proceso continúa sucesivamente hasta que se alcanza un nivel final (Sagaró del Campo y Zamora-Matamoros, 2019b).

Hsimy (árbol de similaridad)

"El dendograma es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud." (Minitab Statistical Software, 2023) El árbol de similitud calcula para cada para de variables la similitud que existe entre ellas. Luego va agregando las mismas clases compuestas por otras clases (Pazmiño-Maji et al., 2017). Este método es muy útil para analizar cómo se formar los conglomerados en cada paso y evaluar los niveles de similitud entre ellos.

En un árbol de similaridad, los factores o elementos se agrupan en función de sus características similares, utilizando una medida de similaridad definida. El árbol de similaridad es una representación visual que muestra las relaciones de similitud entre los factores o elementos. Al aplicar la medida de similaridad a todos los pares de factores, se obtiene una matriz de similaridad que refleja las similitudes entre ellos. Esta matriz se utiliza para construir el árbol de similaridad, donde los factores con mayor similitud se agrupan más cerca entre sí, mientras que aquellos con menor similitud se encuentran más alejados (Sagaró del Campo y Zamora, 2020).

Criterios de comparación

Los criterios de comparación de la complejidad computacional experimental se tomaron con base en el tiempo de ejecución y la cantidad de memoria de almacenamiento que utiliza el programa R para conocer las medidas experimentales de un análisis de complejidad algorítmica en los métodos clústeres utilizados en el análisis estadístico implicativo y en los métodos clústeres utilizados en el análisis multivariado. Para elegir el mejor método con base al tiempo se tiene en cuenta que no existe diferencia significativa entre los métodos analizados. Para seleccionar el método más óptimo con base en la memoria, se considera que existe diferencia significativa entre los métodos con base al criterio de memoria.

Descripción del caso de estudio

Las bases de datos utilizadas en este estudio fueron proporcionadas por el Centro de Investigación y Desarrollo Ecuador (CIDE). Para su creación, se empleó una computadora equipada con un microprocesador Intel® Core™ i7-CPU a 2.2 GHz y 8 GB de memoria RAM. El sistema operativo

utilizado fue Windows 8 de 64 bits. En cuanto al software, se trabajó con R, una plataforma estadística libre, en su versión 3.6.1. Además, se utilizó el entorno de desarrollo integrado RStudio en su versión 1.0.143, que también es una herramienta libre. Para llevar a cabo los análisis, se empleó el paquete RCHIC en su versión 0.27.

Las bases de datos se generaron de forma aleatoria utilizando la función runif(), perteneciente al paquete estándar de R. La mayoría de los datos utilizados eran de tipo categórico, con un total de 10 categorías consideradas.

Se tomaron como variables:

- **Variable independiente:** métodos de análisis clúster.
- **Variable dependiente:** tiempo de ejecución (en segundos) y espacio de memoria (en megabytes)

En esta investigación para cumplir con el objetivo principal de realizar un análisis comparativo de las principales técnicas multivariadas usadas en datos cualitativos, se plantearon los siguientes objetivos específicos: comparar bajo el criterio de tiempo de ejecución, los métodos MHCPC, MSIM, Hrarchy y Hsimy para datos cualitativos, y comparar bajo el criterio de cantidad de memoria, los métodos MHCPC, MSIM, Hrarchy y Hsimy para datos cualitativos.

Metodología

Esta investigación según el enfoque de investigación es cuantitativa, debido a que se compara a través de los resultados numéricos los métodos multivariantes. Tiene un diseño experimental – pre experimental por la manipulación de variables las cuales no tienen un grupo de control son de tipo aleatorio. El tipo de inferencia es inductiva porque los resultados serán generalizados en la población mediante pruebas de hipótesis. Además, esta investigación es de corte transversal, debido a que se considera el periodo académico de abril – 2020 a septiembre – 2020.

La hipótesis de investigación plantea que existe diferencia significativa entre los resultados de las técnicas multivariadas en los criterios de memoria y tiempo de ejecución, para el caso de una población que estuvo compuesta por un total de 60000000 bases de datos. Cada una de estas bases de datos estaba formada por un máximo de 100,000 observaciones y 600 variables. Se priorizó el uso de datos de tipo categórico. La información sobre la población incluye el nombre del archivo, el número de filas y columnas, el tamaño total de los datos, el tiempo de ejecución, la memoria utilizada y el

sistema operativo. El análisis de los datos se realizó mediante el software IBM SPSS Statistics versión 21 y R versión 3.0 para demostrar la hipótesis la prueba paramétrica t para proporciones.

Para la comprobación de los supuestos se utilizaron pruebas no paramétricas, el test de lillie.test para comprobar la existencia de una distribución normal, el supuesto de normalidad mediante el test de Levene, la autocorrelación se comprobó por el test de Durbin-Watson. Adicionalmente, la prueba estadística utilizada fue de Kruskall-Wallis para encontrar diferencias significativas entre los 4 métodos y una post-prueba de Tukey para medir la homogeneidad.

Resultados

Se diseñó el algoritmo para calcular el tiempo de procesamiento y el uso de memoria por método, después de realizar la agrupación basada en el criterio de la cantidad de datos (ndatos) y verificar la homogeneidad dentro de los grupos se obtuvieron cinco grupos, se analizó la heterogeneidad entre los grupos, se procedió a eliminar los valores atípicos dentro de cada grupo. Se llevaron a cabo las pruebas no paramétricas donde se verificaron los supuestos de normalidad, homocedasticidad e independencia. Respecto a los ndatos máximos y mínimos por grupo se obtuvo la tabla 1 que permite identificar el rango en el que se encuentra la base de datos que se va a analizar, en la cual se considera dentro del grupo 1 a las bases cuyos ndatos se encuentran entre [1 - 4054200], para el grupo 2 las observaciones que están entre [4054201 – 10322400], en el grupo 3 entre [10322401 – 20100600] datos, para el grupo 4 entre el rango de [20100601 – 28033000], y en el grupo 5 entre [28033001 – 60000000] datos.

Tabla 1. *Mínimos y máximos por grupos para la población.*

Grupo	Mínimo	Máximo
1	1	4054200
2	4054201	10322400
3	10322401	20100600
4	20100601	28032000
5	28032001	6000000

Nota. Elaborado por Batallas (2022)

El test no paramétrico de Kruskal-Wallis para evaluar las diferencias entre los grupos y los métodos utilizados. Los resultados por grupos de los cuatro métodos utilizados, muestra que los p-valor por grupo tienen valores superiores al 0,05 indicando que no existe diferencia significativa en

la aplicación de dichos métodos. En cuanto a los p-valor por método poseen valores inferiores al 0,05 esto muestra que existe una diferencia significativa entre los 5 grupos en relación a timeMHCPC, timeMSIM, timeHrarchy y timeHsimy. Para analizar la homogeneidad, se aplicó una post-prueba de Tukey, en la cual para los métodos MHCPC, MSIM y Hrarchy no se presentan diferencias significativas entre los grupos 4 y 5 por lo tanto son similares entre ellos. Además, en el método Hsimy entre los grupos 3 y 4, y los grupos 4 y 5 tampoco se presentan diferencias significativas, por lo que son similares entre ellos.

Tabla 2*. Resultados del test de Kruskall Wallis en base al tiempo.*

Grupo	timeMHCPC	timeMSIM	time Hrarchy	timeHsimy	**p-valor**
1	1,30E+24	8,55E+22	1,99E+24	1,52E+23	**0,06827**
2	2,62E+23	1,56E+23	1,49E+23	2,37E+23	**0,99919**
3	2,19E+23	1,88E+24	1,12E+24	5,61E+23	**0,38754**
4	1,13E+24	2,05E+24	2,17E+24	1,93E+24	**0,70941**
5	1,97E+24	7,57E+23	1,60E+24	1,28E+24	**0,82842**
p-valor	**1,23E-39**	**1,20E-38**	**1,13E-38**	**9,50E-45**	

Nota. Elaborado por Batallas (2022)

Entonces, con base en el tiempo de ejecución, cuando se trabaja con bases de grandes magnitudes de datos, entre [1- 60000000], es factible aplicar cualquiera de los cuatro métodos de agrupación debido a que se encontró que no existen diferencias significativas en la aplicación de los cuatro métodos por ello se recomienda utilizar el método que posee menor valor en su varianza.

La tabla 3, indica los resultados en base a la memoria, en los cuales el p-valor por grupo muestra que solamente el grupo 1 no presenta diferencia significativa entre métodos, mientras que en los grupos 2, 3, 4 y 5 sí existe una diferencia significativa. A su vez, entre los cuatro métodos existe una diferencia significativa por lo que se aplicó la post-prueba de homogeneidad de Kruskall Wallis, Tukey`s range test. En los resultados de la post-prueba de homogeneidad entre métodos para el grupo 2 en la comparación dos a dos, se comprueba que los cuatro métodos no presentan diferencias significativas por tanto se podría utilizar cualquiera de los métodos para este grupo. Mientras que para los grupos 3, 4 y 5 los métodos MHCPC y MSIM no presentan diferencias significativas, pero Hrarchy y Hsimy son similares entre sí y ocupan un poco más de memoria que los otros métodos.

Adicionalmente, se realizó la post-prueba de homogeneidad para la comparación de dos a dos por métodos entre los cinco grupos, donde se obtuvo que de acuerdo a los resultados de los cuatro

métodos (MHCPC, MSIM, Hrarchy y Hsimy) los grupos 2, 3 y 4 no presentan diferencias significativas.

Tabla 3. *Resultados del test de Kruskall Wallis en base a la memoria.*

Grupo	Mbytes MHCPC	MbytesMSIM	MbyteHrarchy	MbyteHsimy	**p-valor**
1	327,3	327,3	335	335,1	**0,17869**
2	314,9	314,9	318,1	323,6	**0,0447**
3	321,9	321,9	332,9	329,2	**8,6E+01**
4	314,9	314,9	320,2	328	**3,05E+01**
5	319,6	319,6	325,4	332,2	**2,61E+01**
p-valor	**5,53E-15**	**5,53E-15**	**1,35E-16**	**1,47E-17**	

Nota. Elaborado por Batallas (2022)

En los cuatro métodos, se observa que el grupo 5 presenta el menor consumo de memoria, mientras que el grupo 1 requiere la mayor cantidad de memoria. Tras realizar pruebas de homogeneidad, se encontraron diferencias significativas entre los cinco grupos. Se determinó que los grupos 2, 3 y 4 son similares entre sí y requieren una menor cantidad de memoria en comparación con el grupo 1.

Tabla 4. *Post-pruebas de homogeneidad por métodos.*

MbytesMHCPC					MbytesMSIM				
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
3					3				
	2	2	2			2	2	2	
				1					1

MbyteHrarchy					MbyteHsimy				
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
3					3				
	2	2	2			2	2	2	
				1					1

Nota. Elaborado por Batallas (2022)

En cuanto al uso de memoria, en el caso del grupo 1, donde el número total de datos de la base se encuentra entre [1 - 4054200], la elección del método resulta irrelevante, pero para seleccionar el modelo más apropiado, se sugiere considerar aquel que presente la menor varianza en el uso de memoria. Por otro lado, si el número total de datos se encuentra en el intervalo de [4054200 -

10322401], como en el caso del grupo 2, o si las observaciones caen en los intervalos [10322400 - 20100600] del grupo 3, o en los intervalos [20100600 - 28033000] del grupo 4, o entre [28033000 - 60000000] del grupo 5, se recomienda utilizar los métodos más eficientes, como MHCPC o MSIM, debido a su menor consumo de memoria.

Conclusiones

El proceso realizado permitió determinar los métodos óptimos en términos de complejidad algorítmica, resultando en la formación de cinco grupos homogéneos. Los resultados obtenidos al analizar los cuatro métodos utilizados (MHCPC, MSIM, Hrarchy y Hsimy) en relación al tiempo de ejecución revelaron que no hay diferencias significativas entre ellos. La prueba de homogeneidad aplicada a los métodos MHCPC, MSIM y Hrarchy mostró que los grupos 4 y 5 son similares entre sí. En el caso del método Hsimy, se encontraron similitudes entre los grupos 3 y 4, así como entre los grupos 4 y 5. Dado que no se encontraron diferencias significativas entre los métodos, se recomienda seleccionar aquel que presente la menor varianza, debido a que esto indica un menor tiempo de ejecución.

En relación al análisis por uso de memoria, al no existir diferencias significativas en el grupo 1 se podría aplicar cualquiera de los cuatro métodos. Sin embargo, un análisis más detallado revela que en el grupo 2 no importa el método utilizado, mientras que en los grupos 3, 4 y 5, los métodos MHCPC o MSIM presentan mejores resultados. Además, se observó que el grupo 5 ocupa menos memoria en comparación con el grupo 1, que requiere una mayor cantidad de memoria. En definitiva, se concluye que para la cantidad de observaciones en el grupo 1, es posible utilizar cualquiera de los cuatro métodos, a pesar de que ocupan más memoria. Para los grupos 2, 3, 4 y 5, se recomienda aplicar el método MHCPC o el método MSIM.

Referencias

Álvarez-Perdomo. P., Tamayo, M., y Govea, J., (2022). Técnicas multivariadas: una contribución al análisis económico financiero en PYMES bananeras ecuatorianas. *Revista Universidad y Sociedad. 14*(4). 475-485. Epub 30 de agosto de 2022. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202022000400475&lng=es&tlng=pt.

Batallas, N. (2022). Análisis de las características de la Educación Superior Ecuatoriana en tiempos de la pandemia: Comparación de técnicas multivariadas. [Tesis de grado, Riobamba: Escuela Superior Politécnica de Chimborazo (ESPOCH)].

Chan, D., y Galli, M., (2020). Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior. *Revista Argentina de Educación Superior (RAES). 12*(20). 123-136. https://dialnet.unirioja.es/servlet/articulo?codigo=7592065

Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C., (2019). Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM (JACM). 66*(4). 1-42. https://doi.org/10.1145/3321386.

Fotiadis, T. A., & Anastasiadou, S. (2019). Contemporary advanced statistical methods for the science of marketing: Implicative Statistical Analysis vs Principal Components Analysis. *International Journal of Entrepreneurship and Innovative Competitiveness – IJEIC 1*(1) http://hephaestus.nup.ac.cy/handle/11728/11393.

Govender, P., & Sivakumar, V., (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research. 11*(1). 40-56. https://doi.org/10.1016/j.apr.2019.09.009

Hidalgo, A., (2019). Técnicas estadísticas en el análisis cuantitativo de datos. Revista sigma. 15(1). 28-44. http://coes.udenar.edu.co/revistasigma/articulosXV/1.pdf

Koh, K., Y., Ahmad, S., Lee, J., I., Suh, G., H., & Lee, C., M., (2022). Hierarchical clustering on principal components analysis to detect clusters of highly pathogenic avian influenza subtype H5N6 epidemic across South Korean Poultry Farms. *Symmetry. 14*(3). 598. https://doi.org/10.3390/sym14030598

Martín del Campo, C., (2019). *Atribución de autoría con aprendizaje automático*. [Tesis de maestría. México: Instituto Politécnico Nacional] https://tesis.ipn.mx/bitstream/handle/123456789/27768/T2006.pdf?sequence=1&isAllowed=y.

Maugeri, A., Barchitta, M., Basile, G., & Agodi, A., (2021). Applying a hierarchical clustering on principal components approach to identify different patterns of the SARS-CoV-2 epidemic across Italian regions. *Scientific reports. 11*(1). 7082. https://doi.org/10.1038/s41598-021-86703-3.

Naranjo, M., Pazmiño, R., Conde, M. y Peñalvo, F., (2018, junio). Métodos de agrupamiento LA & SIA: Comparación computacional. En *Congreso de Ciencia y Tecnología ESPE 13*(1). https://doi.org/10.24133/cctespe.v13i1.817

Pazmiño-Maji, R. A., García-Peñalvo, F. J., & Conde-González, M. A. (2017, October). Comparing hierarchical trees in statistical implicative analysis & hierarchical cluster in learning analytics. In *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 1-7). https://doi.org/10.1145/3144826.3145399

Pulido, M. (2020). Aportes del análisis multivariante en planificación por escenarios. *@ limentech. Ciencia y Tecnología Alimentaria*. *18*(2). 25-51. https://revistas.unipamplona.edu.co/ojs_viceinves/index.php/ALIMEN/article/view/4152/2639.

Sagaró del Campo, N. M., y Zamora, L. (2019a). Evolución histórica de las técnicas estadísticas y las metodologías para el estudio de la causalidad en ciencias médicas. *Medisan. 23*(3). 534-556. https://medisan.sld.cu/index.php/san/article/view/2434

Sagaró del Campo, N. M., y Zamora-Matamoros, L. (2019b). Métodos gráficos en la investigación biomédica de causalidad. *Revista Electrónica Dr. Zoilo E. Marinello Vidaurreta, 44*(4). https://revzoilomarinello.sld.cu/index.php/zmv/article/view/1846.

Sagaró del Campo, N. M., y Zamora, L. (2020). ¿Cómo interpretar los resultados del análisis estadístico implicativo en los estudios de causalidad en Salud? *MediSur, 18*(2), 292-306. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1727-897X2020000200292&lng=es&tlng=es.

Vargas, J. J., Muñoz, J. J., Paba, N. A., y Ordoñez, N. (2020). Aplicación de la técnica multivariada Manova a dos variables de control provenientes de tres modelos de simulación estocásticos

de un proceso productivo. *Entre Ciencia e Ingeniería*. *14*(28). 66-75. https://doi.org/10.31908/19098367.2056