UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Vínculos
https://revistas.udistrital.edu.co/index.php/vinculos/

Revista VÍNCULOS
Ciencia, Tecnología y Sociedad: Un Enlace hacia el Futuro

I+D Investigación y Desarrollo

Normalización en desempeño de k-means sobre datos climáticos

Normalization in k-means performance on climate data

Juan Sebastián Ramírez Gómez[1] Néstor D. Duque Méndez[2] Jorge Julián Vélez Upegui[3]

Para citar este artículo: J. S. Ramírez-Gómez; N. D. Duque-Méndez; J. J. Vélez-Upegui, "Normalización en desempeño de k-means sobre datos climáticos". *Revista Vínculos: Ciencia, Tecnología y Sociedad*, vol 16, n° 1, enero-junio 2019, 57-72. DOI: https://doi.org/10.14483/2322939X.15550.

Recibido: 20-11-2018 / Aprobado: 22-02-2019

Resumen

El análisis de clúster sobre datos climatológicos es usado en diversas investigaciones dado que permite obtener resultados interesantes para cada enfoque propuesto. Por tanto, en este trabajo se presenta la evaluación de desempeño del algoritmo de agrupamiento K-Means a partir del uso de normalización aplicada a un conjunto de datos con cuatro variables climatológicas (temperatura, precipitación, humedad relativa y radiación solar) para una estación ubicada en la ciudad de Manizales, Colombia. Esto con el fin de determinar el efecto de aplicar, o no, la normalización en la calidad de los clústeres y evaluar el costo computacional del algoritmo según las características establecidas. Para ello se definen seis escenarios de ejecución para 2, 3 y 5 clústeres con diferente cantidad y agrupación de variables utilizando distancia euclidiana como medida de alejamiento, Davies-Bouldin como método evaluación de calidad de los clústeres y la aplicación de normalización con Z-transformation y Range transformation. Se concluye que, a través de una comparación con k-medoides y aplicación STFT (Transformada de Fourier de Tiempo Reducido), la normalización mejora los resultados y con Z-transformation se obtienen los mejores desempeños de agrupamiento según el índice de Davis-Bouldin.

Palabras clave: clustering, K-Means, machine learning, normalización, transformada de Fourier a corto plazo.

Abstract

Cluster analysis of climatological data is used in various investigations as it allows interesting results to be obtained for each proposed approach. Therefore, this paper presents the performance evaluation of the K-Means clustering algorithm from the use of standardization applied to a data set with four climatological variables (temperature, precipitation, relative humidity and solar radiation) for a station located in the city of Manizales, Colombia. This in order to determine the

1. MSc (c) en Administración de Sistemas Informáticos. Universidad Nacional de Colombia, Sede Manizales. Correo electrónico: jsramirezgo@unal.edu.co. ORCID: https://orcid.org/0000-0001-8876-5371
2. PhD Ingeniería. Universidad Nacional de Colombia. Profesor Asociado Universidad Nacional de Colombia – Sede Manizales. Director Grupo Investigación GAIA. Correo electrónico: ndduqueme@unal.edu.co. ORCID: https://orcid.org/0000-0002-4608-281X
3. PhD en Planificación y Gestión de Recursos Hidráulicos. Profesor Asociado Universidad Nacional de Colombia – Sede Manizales. Correo electrónico: jjvelezu@unal.edu.co. ORCID: https://orcid.org/0000-0003-3856-1105

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

effect of applying, or not, the normalization in the quality of the clusters and to evaluate the computational cost of the algorithm according to the established characteristics. For this purpose, six execution scenarios are defined for 2, 3 and 5 clusters with differentquantity and grouping of variables using Euclidean distance as a distance measure, Davies-Bouldin as a quality evaluation method of the clusters and the application of normalization with Z-transformation and Range transformation. It is concluded that, through a comparison with k-medoides and STFT application (Fourier Transform of Reduced Time), the normalization improves the results and with Z-transformation the best grouping performances are obtained according to the Davis-Bouldin index.
Keywords: clustering, K-Means, machine learning, normalization, short-time Fourier transform.

1. Introducción

La meteorología ha sido un área de mucho interés y curiosidad no solamente para la sociedad sino también para la comunidad científica en aras de comprender el comportamiento y las condiciones climáticas que se presentan en el planeta tierra. Por esta razón, el clima y los escenarios atmosféricos han sido abordados por diversidad de investigadores para adquirir conocimiento de interés, empleando datos climáticos y meteorológicos junto con algoritmos de agrupamiento buscando determinar o entender comportamientos y patrones dentro del área estudiada, relacionar causas y efectos, comprender la formación e impacto de los desastres naturales y el efecto invernadero dentro de una región, al igual que realizar mediciones de la zona para finalmente proveer mejoras, conclusiones y consideraciones en pro del medio ambiente [1-11].

En dichos análisis de datos climáticos se han empleado diversos algoritmos de agrupamiento para procesar los registros, ya sean de agrupamiento jerárquico [1], [2], [8]; de agrupamiento empleando Stepwise Cluster Analysis o (SCA, por sus siglas en inglés) [11-12], [13]; de agrupamiento usando Space-Time Permutation Scan Statistics o (STPSS, por sus siglas en inglés) [14]; y de Second Order Data Coupled Clustering o (SODCC, por sus siglas en inglés) [6], [7]. Además, dentro del aprendizaje de máquina no supervisado se encuentra también el algoritmo particionado K-Means, muy ampliamente usado por los investigadores en el campo climatológico y uno de los algoritmos más conocidos dentro del *machine learning* para el análisis de variables y datos de meteorología [1], [4], [5], [8–10], [15], debido a que es reconocido como uno de los algoritmos más simples y eficientes dentro del agrupamiento de datos [16].

En la revisión de literatura se encuentra que autores como Ghayekhloo [9] emplearon K-Means en cuatro variables climáticas (radiación solar, temperatura, velocidad y dirección del viento) para un rango de tiempo de dos años, el cual es comparado con GTSOM (Game Theoretic Self-Organizing Map), un modelo que se propone para demostrar quién tiene mejor precisión y así dar un pronóstico sobre K-Means. Sin embargo, la investigación no llevó a cabo análisis de rendimientos computacionales y comportamiento de los algoritmos bajo diferentes escenarios, como tampoco una relación entre las variables, efectos de normalización y calidad de clústeres.

Respecto al algoritmo K-Means, autores como Arroyo [1] realizaron un aporte al llevar a cabo un análisis de algoritmos con datos meteorológicos comparando varios algoritmos de clustering, en donde particularmente emplearon para K-Means el uso de cuatro medidas de distancia (Seuclidean, Cityblock, Cosine y Correlation) y cantidad de clústeres (K=2, 3, 4, 5 y 6) para seis variables climáticas (temperatura máxima y mínima, velocidad del viento, luz solar, presión atmosférica máxima y mínima). No obstante, el estudio no mostró el comportamiento de las variables climáticas ni un completo costo computacional de la ejecución del algoritmo bajo diferentes escenarios que priorizaran el conjunto de datos o información climática, como tampoco una visualización de los efectos de normalizar, o no, y con qué métodos, o cómo se da el manejo y efecto de ejecutar grandes cantidades de ceros en las variables para conocer una evaluación más amplia del desempeño del algoritmo K-Means.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Otros trabajos de investigadores han utilizado K-Means como algoritmo para apoyar las metodologías propuestas por los autores y el posterior procesamiento de sus conjuntos de datos climáticos [5], [8], [10], [15] encontrando satisfacción en sus resultados obtenidos, pero más allá del empleo del algoritmo, los autores no abordaron en sus investigaciones la evaluación del algoritmo K-Means bajo una formulación de criterios diferentes.

Desde esta perspectiva se aprecia que existe un espacio abierto orientado a determinar, con la aplicación de K-Means, qué variables meteorológicas son las que mejor se relacionan entre sí, qué cantidad y cuáles variables usar para la conformación de clústeres, qué cantidad de registros y con qué cantidad de clústeres trabajar, qué métodos de normalización usar y cuáles aplicar, o no, y cómo influye el procesamiento y análisis de resultados con variables que contienen grandes cantidades de ceros. Esto con el fin de construir escenarios de buen desempeño para el análisis de datos, evaluando, también, el costo computacional en cada caso para comprender la exigencia de procesamiento y consumo de hardware. Por estas razones, este artículo se orienta a explorar esta brecha para generar un acercamiento sobre el comportamiento y desempeño del algoritmo de agrupamiento K-Means bajo un conjunto de datos meteorológicos en diferentes escenarios.

2. Metodología

Dentro de las estaciones meteorológicas que conforman el sistema de monitoreo del CDIAC (Centro de Datos e Indicadores Ambientales de Caldas) en la ciudad de Manizales, Colombia se escoge la estación Hospital de Caldas, el principal hospital de este departamento de Colombia, para realizar el estudio propuesto debido a que geográficamente se encuentra en el centro de la ciudad y su ubicación es importante por estar establecido allí.

Para esta estación, mediante el proceso de ETL (Extracción, Transformación y Carga) se genera un conjunto de datos extraído de la bodega de datos ambientales, el cual contiene registros comprendidos entre el 12 de abril de 2012 y el 16 de agosto de 2017 (lo que abarca un margen de tiempo de 64 meses o 5.3 años). Acto seguido, se seleccionan los atributos de temperatura, precipitación, radiación solar y humedad relativa, dado que son los más usados por diferentes autores en sus investigaciones climáticas con algoritmos de clustering [1], [2], [12], [13], [3], [5–11]. Luego, dicho conjunto de datos es procesado con el software de análisis de datos Rapid Miner, versión 7.5.003, utilizando aprendizaje de máquina no supervisado.

En esta medida, el algoritmo seleccionado es K-Means; la Distancia Euclidiana es seleccionada como función de distancia, considerada como la de mayor confianza [1] y una de las más utilizadas en una gran variedad de trabajos [2], [3], [8–10], [15], [17]; y los valores que se determinan para el número de clústeres es de K=2, K=3 y K=5, como parte de la creación de los escenarios en esta investigación. Además, se establece que se aplicará normalización con Z-transformation y Range transformation como parte del proceso de reducir escalas de valores en las variables para algunos escenarios específicos [1], [7], [13], [18–22], y que se utilizará Davis-Bouldin como métrica propuesta de evaluación de calidad de clúster [1], [16], [17], [23], [24]. En total, se ejecutarán 18 escenarios con K-Means donde 6 de ellos corresponden a un valor de clúster (K) diferente, diseñados como se aprecia en la Tabla 1.

Los anteriores escenarios se han propuesto para observar y explicar el comportamiento del agrupamiento aplicando, o no, normalización sobre los datos. Además, en algunos escenarios se toma en cuenta la variable precipitación para ver qué sucede dentro del agrupamiento cuando un atributo contiene una inmensa cantidad de ceros. De este modo, para conocer correlaciones y comportamientos tanto a nivel climatológico como a nivel computacional se aplica el algoritmo en cada uno de los escenarios obteniendo los resultados mostrados en las Tablas 2, 3 y 4, al igual que se exponen los agrupamientos correspondientes en las Figuras 1, 2, 3, 4, 5 y 6.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Tabla 1. Definición de los escenarios y características a utilizar en la experimentación.

Valor de K	Escenarios	Cantidad de variables a utilizar	Variables	Normalizar	Método de normalización	Criterio de distancia	Criterio de evaluación de clúster
2, 3 y 5	#1	3	Temperatura, Humedad relativa y Radiación solar.	NO	No aplica.	Euclidiana	Davis-Bouldin
	#2	4	Temperatura, Humedad relativa, Radiación solar y Precipitación.	NO	No aplica.		
	#3	3	Temperatura, Humedad relativa y Radiación solar.	SI	Range Transformation		
	#4	4	Temperatura, Humedad relativa, Radiación solar y Precipitación.	SI	Range Transformation		
	#5	3	Temperatura, Humedad relativa y Radiación solar.	SI	Z-transformation		
	#6	4	Temperatura, Humedad relativa, Radiación solar y Precipitación.	SI	Z-transformation		

Fuente: elaboración propia.

Tabla 2. Resultados del agrupamiento usando K-Means con K=2 y distancia euclidiana.

Escenarios	Número de ítems	Agrupamiento del modelo				Criterio de evaluación de clúster: Índice de Davis-Bouldin
		Clúster 0	%	Clúster 1	%	
#1	967.289	808.712	83,6%	158.577	16,4%	-0.516
#2		808.712	83,6%	158.577	16,4%	-0.516
#3		257.479	26,6%	709.810	73,4%	-1.091
#4		257.493	26,6%	709.796	73,4%	-1.093
#5		291.239	30,1%	676.050	69,9%	-1.103
#6		293.346	30,3%	673.943	69,7%	-1.158

Fuente: elaboración propia.

Tabla 3. Resultados del agrupamiento usando K-Means con K=3 y distancia euclidiana.

Escenarios	Número de ítems	Agrupamiento del modelo						Criterio de evaluación de clúster: Índice de Davis-Bouldin
		Clúster 0	%	Clúster 1	%	Clúster 2	%	
#1	967.289	676.998	70,0%	65.518	6,8%	224.773	23,2%	-0.441
#2		676.998	70,0%	65.518	6,8%	224.773	23,2%	-0.441
#3		109.549	11,3%	557.118	57,6%	300.622	31,1%	-1.045
#4		300.617	31,1%	109.551	11,3%	557.121	57,6%	-1.047
#5		300.987	31,1%	106.805	11,0%	559.497	57,8%	-1.093
#6		290.496	30,0%	672.767	69,6%	4.026	0,4%	-0.942

Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Tabla 4. Resultados del agrupamiento usando K-Means con K=5 y distancia euclidiana.

Escenarios	Número de ítems	Agrupamiento del modelo										Criterio de evaluación de clúster: Índice de Davis-Bouldin
		Clúster 0	%	Clúster 1	%	Clúster 2	%	Clúster 3	%	Clúster 4	%	
#1	967.289	601.764	62,2%	36.707	3,8%	164.038	17,0%	115.523	11,9%	49.257	5,1%	-0.489
#2		601.764	62,2%	36.707	3,8%	164.038	17,0%	115.523	11,9%	49.257	5,1%	-0.489
#3		144.124	14,9%	391.023	40,4%	272.131	28,1%	100.831	10,4%	59.180	6,1%	-1.007
#4		144.120	14,9%	59.180	6,1%	100.831	10,4%	272.132	28,1%	391.026	40,4%	-1.009
#5		104.521	10,8%	273.682	28,3%	56.918	5,9%	138.148	14,3%	394.020	40,7%	-1.050
#6		110.425	11,4%	90.615	9,4%	412.586	42,7%	350.329	36,2%	3.334	0,3%	-1.016

Fuente: elaboración propia.

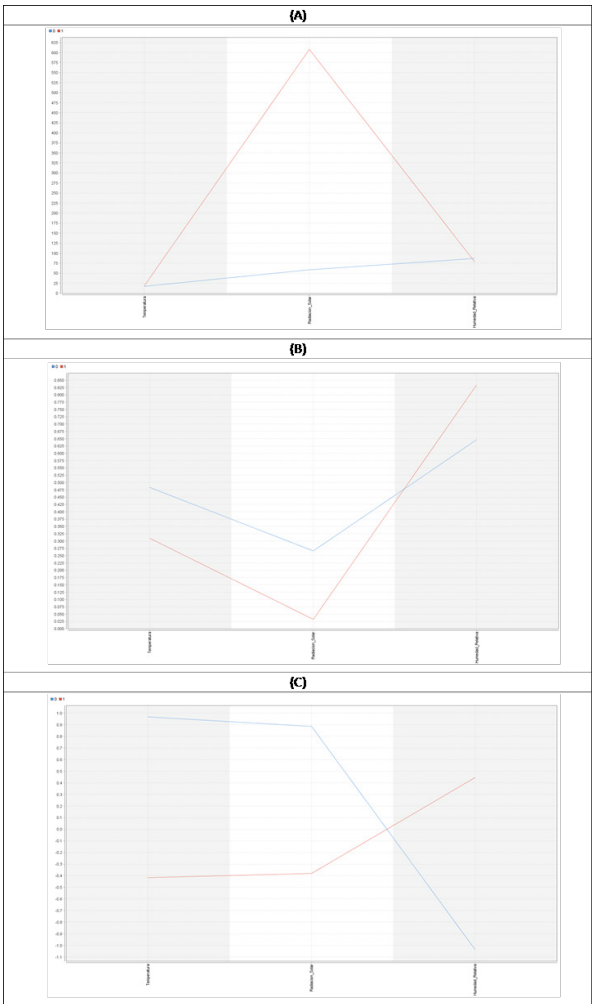

Figura 1. Agrupamiento con K=2 para 3 variables (Temperatura, radiación solar y humedad relativa). (A) Sin normalizar. (B) Normalización con Range Transformation. (C) Normalización con Z-Transformation.
Fuente: elaboración propia.

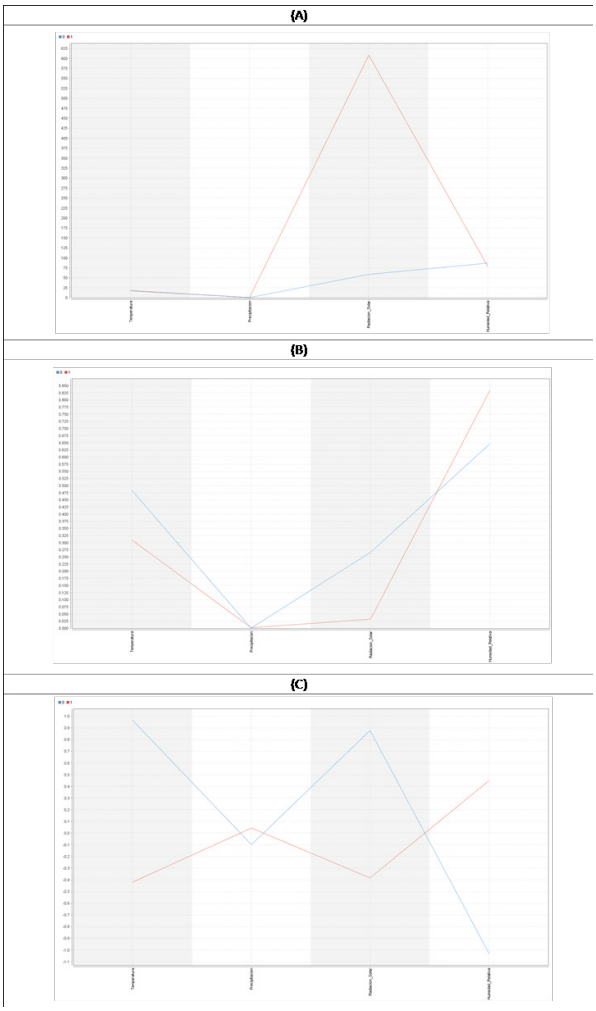

Figura 2. Agrupamiento con K=2 para 4 variables (Temperatura, precipitación, radiación solar y humedad relativa). (A) Sin normalizar. (B) Normalización con Range Transformation. (C) Normalización con Z-Transformation.
Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

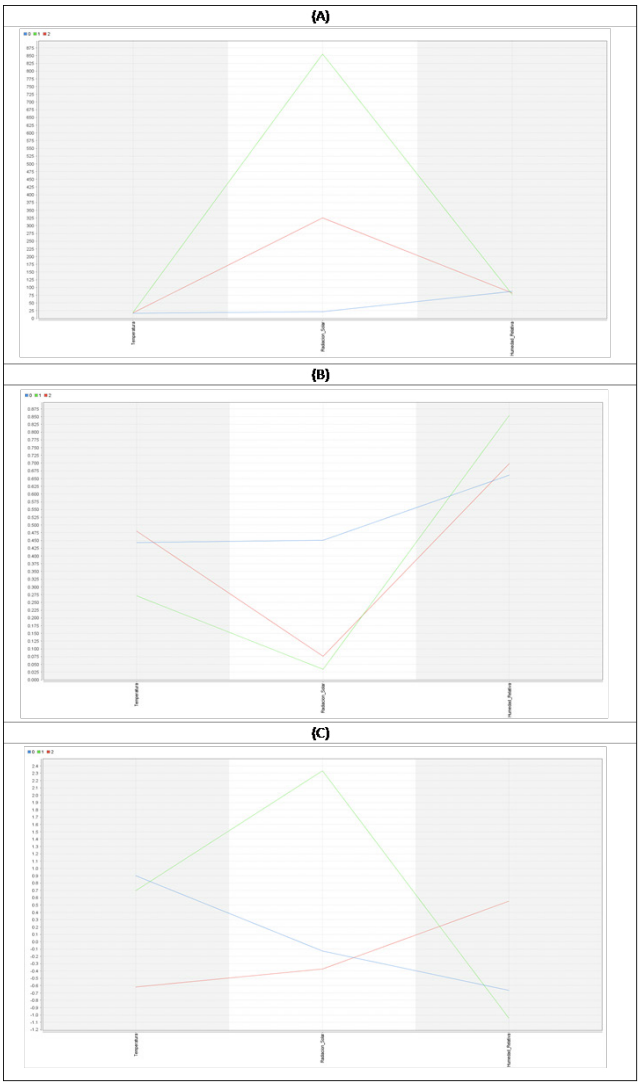

Figura 3. Agrupamiento con K=3 para 3 variables (Temperatura, radiación solar y humedad relativa). (A) Sin normalizar. (B) Normalización con Range Transformation. (C) Normalización con Z-Transformation.
Fuente: elaboración propia.

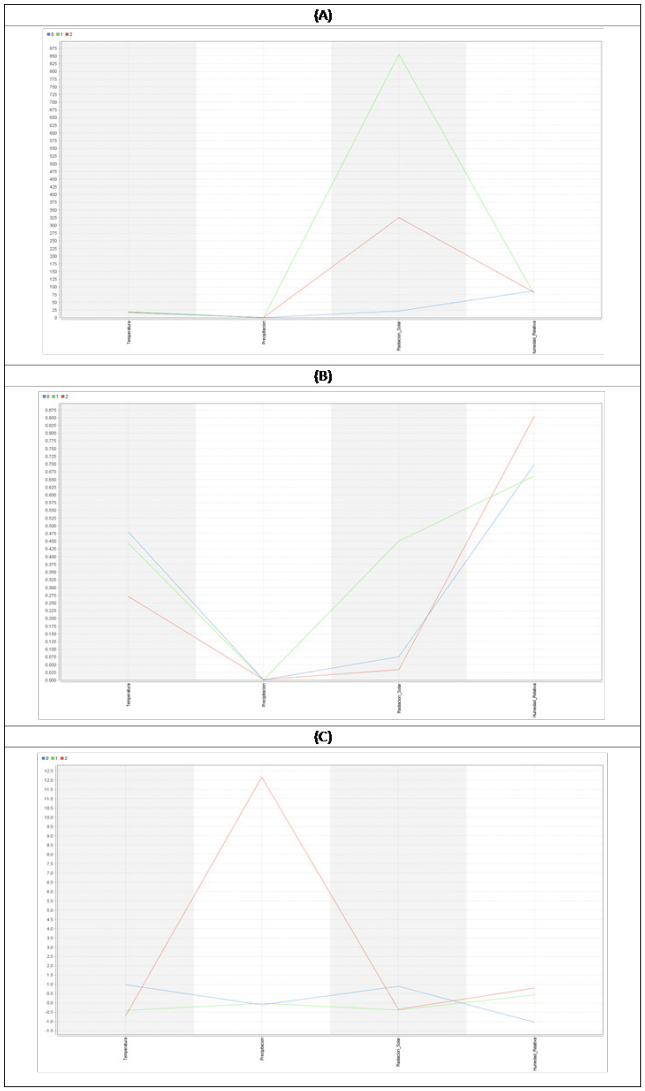

Figura 4. Agrupamiento con K=3 para 4 variables (Temperatura, precipitación, radiación solar y humedad relativa). (A) Sin normalizar. (B) Normalización con Range Transformation. (C) Normalización con Z-Transformation.
Fuente: elaboración propia.

3. Resultados y discusión

Acorde a los escenarios 1, 3 y 5 de la Tabla 2 y según la visualización de la Figura 1, la agrupación se da por la radiación solar. Sin normalizar hay un arrastre por parte del atributo radiación solar con la formación de un clúster muy grande que conforma el 83.6% de los ítems. Con normalización (Range transformation) se disminuye el arrastre de la radiación solar en un 10% y los dos clústeres son un poco más equilibrados en cantidad de ítems (26.6% y 73.4%). El índice de Davis-Bouldin con el que se busca evaluar la calidad de los clústeres ha disminuido hasta -1.091 con respecto al escenario anterior, lo cual quiere decir que la formación del agrupamiento se dio mejor y con mayor calificación, entendiéndose que mientras más bajo sea este índice, mejor es la calidad del clúster [1]. Con

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

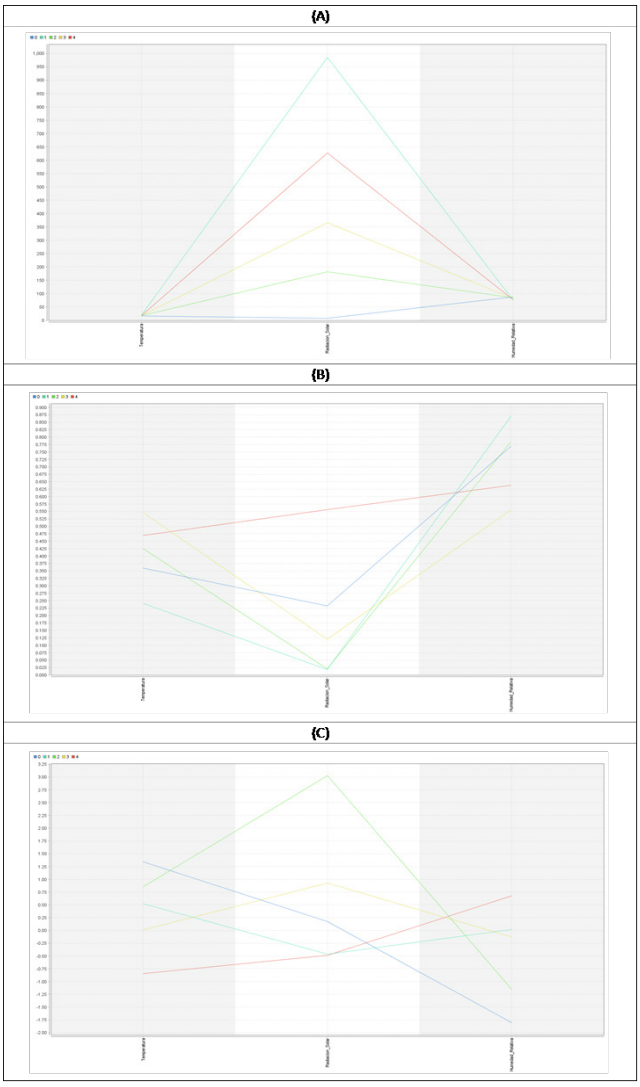

Figura 5. Agrupamiento con K=5 para 3 variables (Temperatura, radiación solar y humedad relativa). (A) Sin normalizar. (B) Normalización con Range Transformation. (C) Normalización con Z-Transformation.
Fuente: elaboración propia.

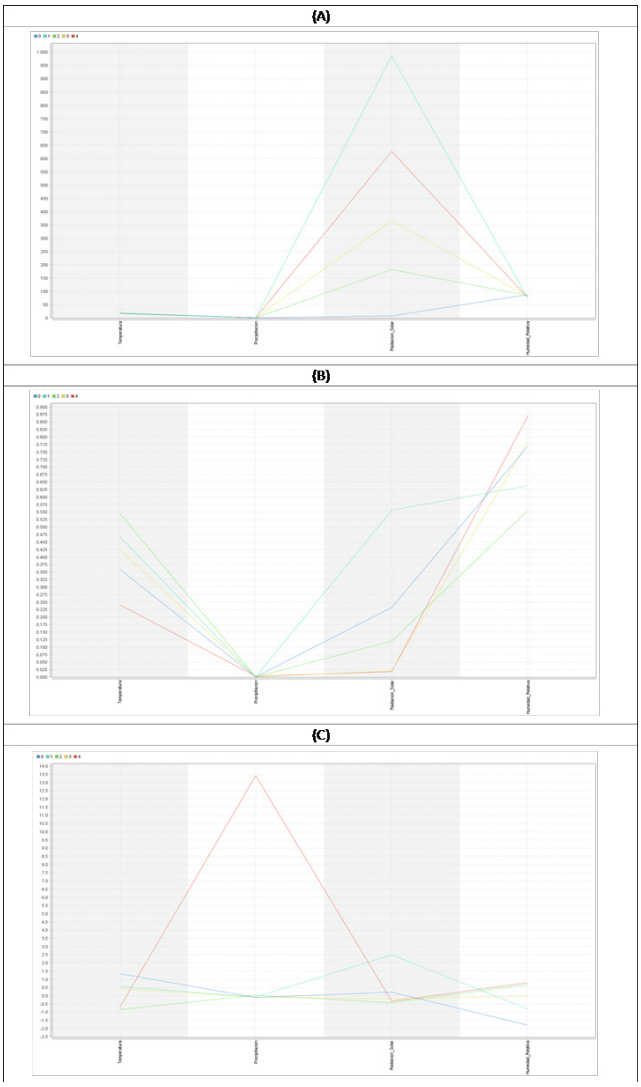

Figura 6. Agrupamiento con K=5 para 4 variables (Temperatura, precipitación, radiación solar y humedad relativa). (A) Sin normalizar. (B) Normalización con Range Transformation. (C) Normalización con Z-Transformation.
Fuente: elaboración propia.

normalización (Z-transformation) también se disminuye el arrastre que genera la radiación solar y los dos clústeres son más equilibrados aún (30.1% y 69.9%) con respecto al escenario anterior. El índice de Davis-Bouldin es menor con -1.103, lo cual refleja qué clúster es mejor.

Con respecto a los escenarios 2, 4 y 6 de la Tabla 2 y según la visualización de la Figura 2, la agrupación se da por la radiación solar. Sin normalizar hay un arrastre muy grande por parte del atributo radiación solar cuyo clúster más grande se conforma por el 83.6% de los ítems, igual que el escenario anterior. Con la normalización (Range transformation) se disminuye el arrastre de la radiación solar en un 10% y los dos clústeres son un poco más equilibrados en cantidad de ítems (26.6% y 73.4%). El índice de Davis-Bouldin ha disminuido hasta -1.093 obteniéndose así un mejor agrupamiento.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Con normalización (Z-transformation) el agrupamiento se equilibra un poco más (30.3% y 69.7%) con respecto a los escenarios anteriores y ahora el índice de Davis-Bouldin es de -1.158, mejorando más la evaluación del agrupamiento.

Para los escenarios 1, 3 y 5 de la Tabla 3 y según la visualización de la Figura 3, la agrupación se da por el atributo radiación solar. Sin normalizar hay un arrastre muy grande por parte de este atributo y una formación de clústeres cuyos ítems se reparten así: 70%, 6.8% y 23.2%. Al ejecutarse normalización con Range transformation, los agrupamientos se balancean mejor de la siguiente forma: 11.3%, 57.6% y 31.1%. El índice de Davis-Bouldin da un valor de -1.045, mejorando considerablemente la calidad de los clústeres. Luego, al ejecutar normalización con Z-transformation los ítems se reparten así: 31.1%, 11% y 57.8%, quedando igual que el escenario anterior a pesar de que el índice de evaluación de clustering mejoró hasta -1.093.

Para los escenarios 2, 4 y 6 de la Tabla 3 y según la visualización de la Figura 4, la agrupación se da por el atributo radiación solar. Sin normalizar hay una formación de clústeres cuyos ítems se reparten así: 70%, 6.8% y 23.2%, donde se evidencia el arrastre del clúster más grande por parte del atributo radiación solar. Al normalizar con Range Transformation se obtienen clústeres del 31.1%, 11.3% y 57.6%, obteniéndose así un mejor balance entre clústeres y cuyo índice Davis-Bouldin mejora hasta obtener -1.047. Luego, en la normalización con Z-transformation se obtienen clústeres muy desequilibrados con el 30%, 69.6% y 0.4%, donde dos de ellos arrastran todos los registros (uno con más del doble de ítems) y cuya evaluación de clustering desmejora notablemente con un valor de -0.942.

Con respecto a los escenarios 1, 3 y 5 de la Tabla 4 y según la visualización de la Figura 5, la agrupación se da por el atributo radiación solar. Sin normalizar hay una formación de clústeres cuyos ítems se reparten así: 62.2%, 3.8%, 17%, 11.9% y 5.1%, evidenciándose el gran arrastre del atributo radiación solar con su gran y más altos valores. Al normalizar con Range transformation mejora la calidad del clúster al obtenerse un índice de Davis-Bouldin en -1.007 y se reparten de forma más balanceada los ítems quedando así los agrupamientos: 14.9%, 40.4%, 28.1%, 10.4% y 6.1%. Luego, normalizando con Z-transformation se obtiene una mejor evaluación de clúster con -1.050 a pesar de que el agrupamiento se dio casi igual al escenario anterior: 10.8%, 28.3%, 5.9%, 14.3% y 40.7%.

Finalmente, para los escenarios 2, 4 y 6 de la Tabla 4 y según la visualización de la Figura 6, la agrupación se da una vez más por el atributo radiación solar. Sin normalizar hay una formación de clústeres cuyos ítems se reparten así: 62.2%, 3.8%, 17%, 11.9% y 5.1%, igual que en el escenario con 3 variables. Al normalizar con Range transformation se obtienen clústeres de 14.9%, 6.1%, 10.4%, 28.1% y 40.4%, igual que en el escenario con 3 variables. La distribución de ítems es más equilibrada y la evaluación de los clústeres obtiene un valor de -1.009 según el índice de Davis-Bouldin, mejorándose así la calidad de estos. Con la normalización empleando Z-transformation se obtiene una mejor evaluación de los clústeres según el índice de Davis-Bouldin al obtenerse un valor de -1.016 aunque la distribución de los ítems es más desproporcionada, lográndose agrupamientos del 11.4%, 9.4%, 42.7%, 36.2% y 0.3%.

Ahora bien, con respecto al costo computacional se obtuvieron los resultados presentados en las Figuras 7, 8, 9 y 10.

Figura 7. Tiempos de ejecución para los 6 escenarios con K=2, K=3 y K=5.
Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

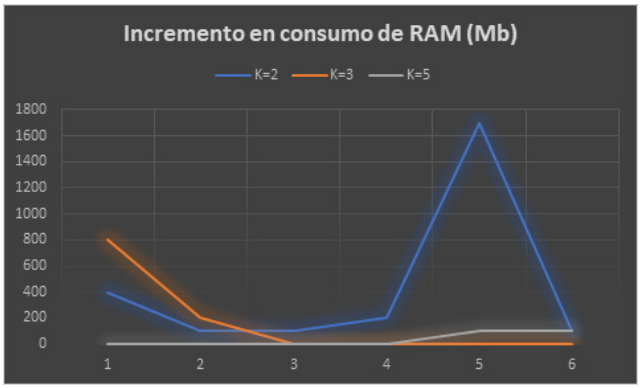

Figura 8. Incremento en consumo de RAM para los 6 escenarios con K=2, K=3 y K=5.
Fuente: elaboración propia.

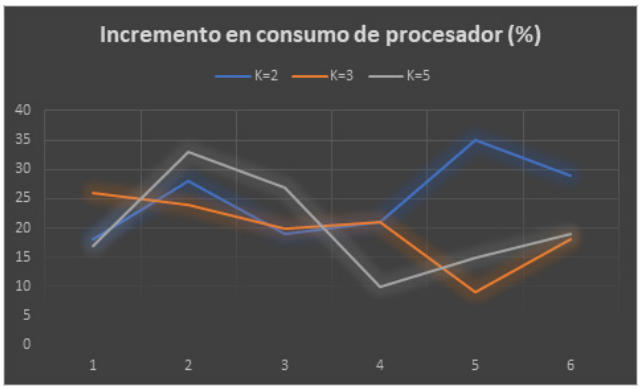

Figura 9. Incremento en consumo de procesador para los 6 escenarios con K=2, K=3 y K=5.
Fuente: elaboración propia.

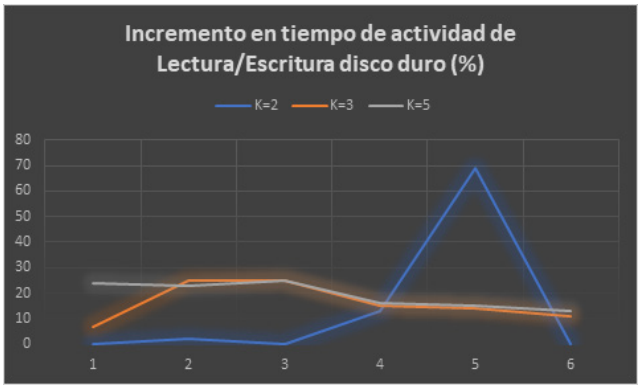

Figura 10. Incremento en tiempo de actividad de Lectura/Escritura para los 6 escenarios con K=2, K=3 y K=5.
Fuente: elaboración propia.

En términos de rendimiento y consumo computacional se encuentra que en tiempos de ejecución se ven incrementos a medida que el algoritmo se va aplicando en los diferentes escenarios, por lo que influye la cantidad de clústeres en los tiempos de ejecución del algoritmo. Con relación al consumo de memoria RAM se observan picos para algunos escenarios con diferentes valores de K, en donde ciertos escenarios tienen un incremento alto, pero en otros es nulo o muy insignificante, por lo que no se evidencia una tendencia o relación proporcional. Con respecto al incremento de consumo de procesador se observan picos elevados para los 6 escenarios con cualquier valor K, lo que muestra que la máquina genera un alto consumo de procesamiento al momento de ejecutar el algoritmo con sus especificaciones. Además, muestran una tendencia común de aumentar para el escenario 2, disminuir para los escenarios 3 y 4 y volver aumentar hacia el escenario 6. Finalmente, para el incremento en tiempo de actividad lectura/escritura de disco duro se evidencian aumentos de actividad con tendencia que se mantienen muy similares para el 66% de los escenarios, lo que muestra que el computador tiene alto nivel de lectura y escritura en el momento de ejecutar el algoritmo en los escenarios establecidos. Cabe mencionar que para el experimento se utilizó un computador portátil Lenovo con las siguientes características: Intel Core i7-6700HQ de 2.60GHz (8 CPU's), 12GB de memoria RAM, sistema operativo Windows 10 Home a 64 bits con disco duro de 1 TB a 7400 rpm.

3.1. Comparación con otros algoritmos

Con el fin de validar si este comportamiento de la normalización en el desempeño se puede aplicar a otros algoritmos de agrupamiento, en la Tabla 5 se exponen los resultados obtenidos empleando el algoritmo K-medoids para un valor de K=5 para el mismo conjunto de datos de la estación Hospital de Caldas. Acorde a los resultados, se evidencia que la normalización provee mejores desempeños de agrupamiento según el índice de Davis-Bouldin.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Asimismo, se observa que, para cada uno de los cinco escenarios, los tratamientos 1.3, 2.3, 3.3, 4.3 y 5.3 tienen el mejor índice de desempeño de agrupamiento, lo que parece confirmar que para los algoritmos de agrupamiento en aprendizaje de máquina no supervisado el método de normalización con Z-transformation provee el mejor desempeño de agrupamiento para el conjunto de datos utilizado. En este sentido, los escenarios y tratamientos establecidos para K-medoids fueron los expuestos en la Tabla 6.

3.2. Transformada de Fourier de Tiempo Reducido STFT sobre variables

Con el fin de fortalecer el análisis de los resultados se aplica STFT (Transformada de Fourier de Tiempo Reducido) para identificar si los comportamientos de las variables por las oscilaciones atmosféricas afectan el análisis de clúster realizado. Los datos son registrados con una frecuencia de muestreo de 0, 2s-1 (1 dato cada 5 minutos). Inicialmente se realiza un gráfico tiempo-frecuencia usando la STFT para verificar que la frecuencia de la señal no cambia a lo largo del tiempo. Para mejorar la visualización de las componentes de interés, el cómputo de los algoritmos para el análisis en frecuencia se realiza restando el valor medio de la señal. Así, se elimina una componente en 0 Hz. Se realiza para diferentes variables.

- **Radiación solar.** Análisis de frecuencia. La STFT realiza un ventaneo de la señal, este se realizó con una ventana de Kaiser con β= 6 y un solape del 50%. La Figura 11 muestra los resultados para la variable radiación solar y con una visualización alternativa se verifica que la señal presenta componentes de baja frecuencia y no cambian en el tiempo.

Esto permite el cálculo de la transformada de Fourier y presentar la señal en el dominio de la frecuencia. En la Figura 12 se presenta el espectro de la señal y se destacan dos componentes, por lo cual en el gráfico se esquematiza el periodo de cada una.

- **Temperatura.** La STFT realiza un ventaneo de la señal, este se realizó con una ventana de Kaiser con β= 20 y un solape del 70%. En la Figura 13 se verifica que la señal presenta componentes de baja frecuencia y no cambian en el tiempo. Esto permite el cálculo de la transformada de Fourier y presentar la señal en el dominio de la frecuencia.

Tabla 5. Resultados del agrupamiento usando K-Medoids con K=5 y distancia euclidiana.

Escenarios	Tratamiento	Número de ítems	Desempeño del algoritmo										
			Agrupamiento										Criterio de evaluación de clúster: Índice de Davis-Bouldin
			Clúster 0	%	Clúster 1	%	Clúster 2	%	Clúster 3	%	Clúster 4	%	
#1	1.1	10.000	8.097	81,0	0	0,0	0	0,0	864	8,6	228	2,3	-1.056
	1.2	10.000	522	5,2	9.114	91,1	210	2,1	153	1,5	0	0,0	-1.280
	1.3	10.000	522	5,2	9.114	91,1	210	2,1	153	1,5		0,0	-1.296
#2	2.1	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	1.323
	2.2	1.859	195	10,5	81	4,4	1.223	65,8	234	12,6	126	6,8	-1.469
	2.3	1.859	199	10,7	726	39,1	469	25,2	342	18,4	123	6,6	-1.663
#3	3.1	9.979	8.641	86,6	0	0,0	1.071	10,7	267	2,7	0	0,0	-1.318
	3.2	9.979	8.873	88,9	377	3,8	130	1,3	405	4,1	194	1,9	-1.479
	3.3	9.979	8.082	81,0	585	5,9	140	1,4	344	3,4	828	8,3	-1.553
#4	4.1	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	-1.323
	4.2	1.859	195	10,5	81	4,4	1.223	65,8	234	12,6	126	6,8	-1.482
	4.3	1.859	203	10,9	1.190	64,0	6	0,3	339	18,2	121	6,5	-1.668
#5	5.1	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	-1.323
	5.2	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	-1.323
	5.3	1.859	1.204	64,8	83	4,5	274	14,7	173	9,3	125	6,7	-1.330

Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Tabla 6. Escenarios y tratamientos establecidos para K-Medoids para un conjunto de datos.

Escenarios		K=5			
		Tratamiento	Normalización	Criterio de distancia	Criterio de evaluación
#1	Cantidad de variables: 3 Tipo de variables: temperatura, humedad relativa y radiación solar. Cantidad de registros: 10.000. Datos faltantes (missing): sí. Valores atípicos (outliers): sí.	1.1	NO	Euclidiana	Davis-Bouldin
		1.2	Range-transformation		
		1.3	Z-transformation		
#2	Cantidad de variables: 3 Tipo de variables: temperatura, humedad relativa y radiación solar. Cantidad de registros: 10.000. Datos faltantes (missing): no. Valores atípicos (outliers): no.	2.1	NO		
		2.2	Range-transformation		
		2.3	Z-transformation		
#3	Cantidad de variables: 4 Tipo de variables: temperatura, humedad relativa, radiación solar y precipitación. Cantidad de registros: 10.000. Datos faltantes (missing): sí. Valores atípicos (outliers): no.	3.1	NO		
		3.2	Range-transformation		
		3.3	Z-transformation		
#4	Cantidad de variables: 4 Tipo de variables: temperatura, humedad relativa, radiación solar y precipitación. Cantidad de registros: 10.000. Datos faltantes (missing): no. Valores atípicos (outliers): sí.	4.1	NO		
		4.2	Range-transformation		
		4.3	Z-transformation		
#5	Cantidad de variables: 4 Tipo de variables: temperatura, humedad relativa, radiación solar y precipitación. Cantidad de registros: 10.000. Datos faltantes (missing): no. Valores atípicos (outliers): no.	5.1	NO		
		5.2	Range-transformation		
		5.3	Z-transformation		

Fuente: elaboración propia.

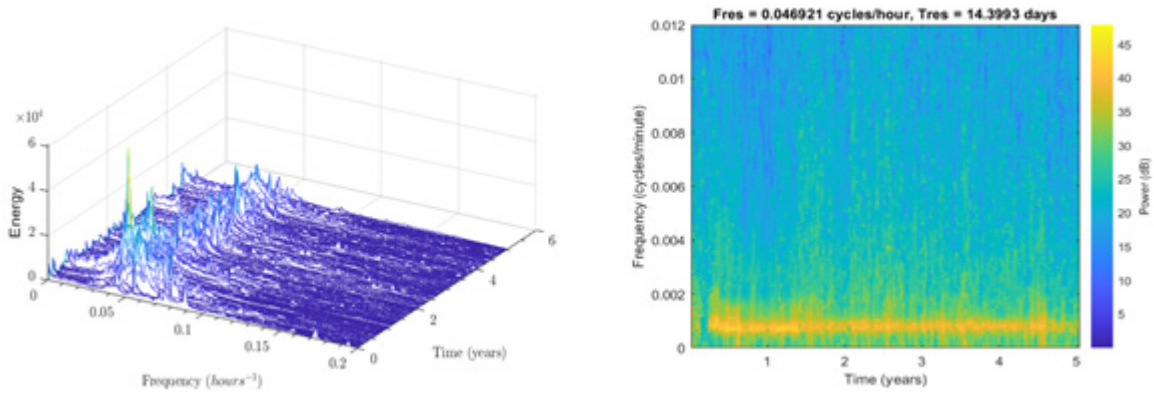

Figura 11. Análisis de frecuencia para la variable radiación solar.
Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

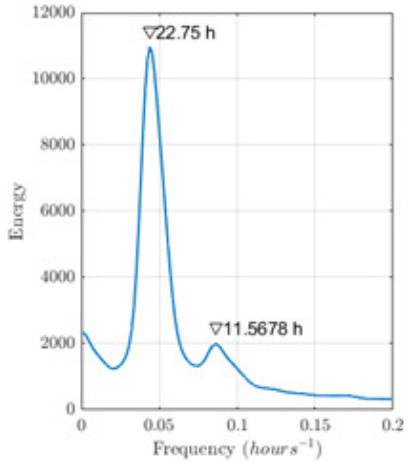

Figura 12. Transformada de Fourier para la variable radiación solar.
Fuente: elaboración propia.

- **Humedad relativa.** La STFT realiza un ventaneo de la señal, este se realizó con una ventana de Kaiser con β= 20 y un solape del 70%. En la Figura 14 se verifica que la señal presenta componentes de baja frecuencia y no cambian en el tiempo. Esto permite el cálculo de la transformada de Fourier y presentar la señal en el dominio de la frecuencia.

Los análisis adicionales realizados permiten, como se sospechaba, determinar que el uso de la normalización en los escenarios propuestos mejora el desempeño de los algoritmos de clustering sin depender del tipo de algoritmo ni del comportamiento de las variables bajo análisis.

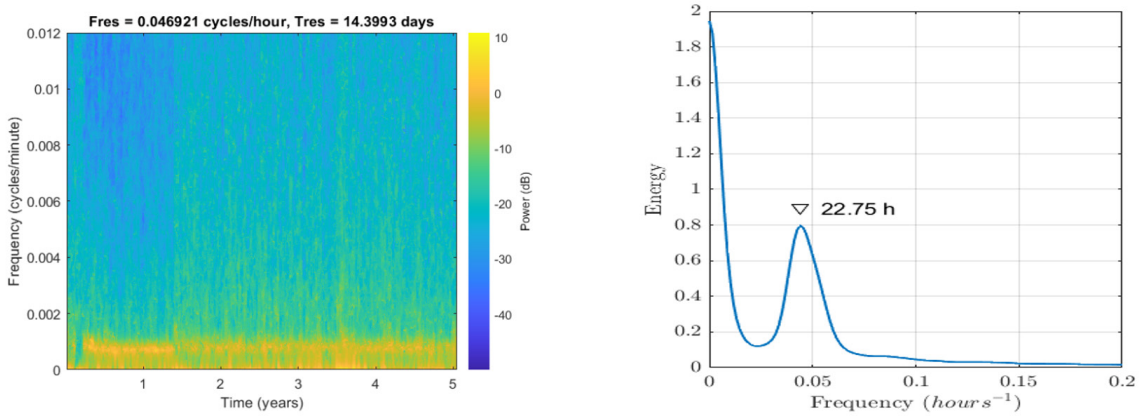

Figura 13. Análisis de frecuencia y Transformada de Fourier para la variable temperatura.
Fuente: elaboración propia.

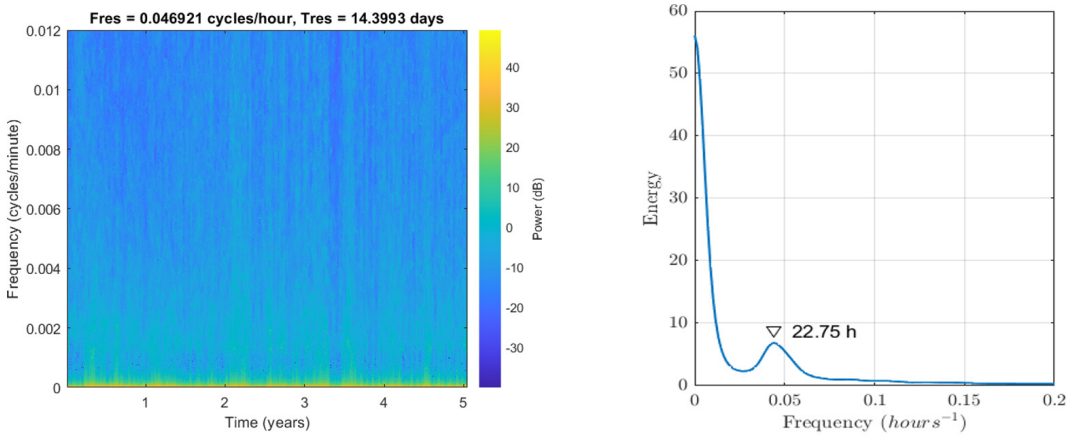

Figura 14. Análisis de frecuencia y Transformada de Fourier para la variable humedad relativa
Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

4. Conclusiones y trabajo futuro

De acuerdo con los resultados obtenidos con respecto a las características de los escenarios se encuentra que la radiación solar es un atributo a partir del cual se forman los agrupamientos, ya sea que se normalice, o no, para la ejecución del algoritmo K-Means. El rango de este atributo oscila entre 0 y 1500, y el tener cientos de miles de registros ubicados en una escala muy alta conllevan a que se aglomeren a partir de la distancia del centroide lo que posteriormente forma un clúster de gran proporción de ítems a comparación de los otros. No obstante, llevar a cabo normalización demuestra que el atributo radiación solar pierde un porcentaje de arrastre grande y que los otros clústeres se balancean con más ítems, por lo que se sugiere hacer uso de normalización cuando atributos como radiación solar hagan parte de un conjunto de datos y cuya escala difiera enormemente en tamaño a comparación de los otros atributos, sin mencionar que la evaluación de calidad del clúster obtendrá un mejor resultado.

Como segunda conclusión, se encuentra que para K=2 los efectos de normalizar ayudan notablemente en la mejora del clustering, no solo porque visualmente se balancean mejor los agrupamientos, sino también porque se obtiene un índice de evaluación de clúster mucho mejor. Para estos escenarios la normalización con Z-transformation mostró mejores resultados, al igual que para K=3 y K=5, según lo arrojado por el índice de Davis-Bouldin, a pesar de que algunos agrupamientos se observaron visualmente más desproporcionados, por lo que se sugiere normalizar con Z-transformation para la obtención de mejores resultados.

Como tercera conclusión, se observa que el atributo precipitación, el cual contiene cientos de miles de registros con valor cero, influye notablemente en los agrupamientos al cambiar el gráfico de los clústeres y generar un inmenso arrastre hacia un valor que no aporta nada útil a los resultados. Si bien no llueve todos los días y las precipitaciones son esporádicas y por corto tiempo, tener un atributo cuyo valor cero está en más del 90% de los registros podría ocasionar malas interpretaciones de resultados o dañar los agrupamientos. Además, este atributo sería el que formaría los clústeres si el atributo radiación solar no estuviera incluido en el conjunto de datos debido a su enorme cantidad de registros bajo un mismo valor, lo que podría causar que los problemas anteriormente mencionados fueran más grandes.

Por lo tanto, se determina con base al índice de Davis-Bouldin que el escenario con mejor desempeño para K=2 fue el escenario 6 donde se aplican las cuatro variables (temperatura, precipitación, humedad relativa y radiación solar) normalizando con Z-transformation. Para K=3 fue el escenario 5 donde se aplican las tres variables (temperatura, humedad relativa y radiación solar) normalizando con Z-transformation. Para K=5 fue igualmente el escenario 5 donde se aplican las mismas tres variables con la misma técnica de normalización, lo cual concluye el buen desempeño que tiene la normalización con Z-transformation en donde el 66% de estos escenarios no tienen en cuenta el atributo precipitación.

Con relación a lo climatológico, se encuentra para todos los escenarios que con el aumento del valor de la precipitación disminuye el valor de la radiación solar. Con temperaturas altas disminuye la concentración de humedad relativa; si la radiación solar aumenta disminuye la humedad relativa; a temperaturas bajas, el valor de la precipitación aumenta. Si la precipitación aumenta, la humedad relativa va desapareciendo, de lo contrario crece su valor. La temperatura mostró igual tendencia que la radiación solar, en términos de visualización y formación de clústeres, por lo que ambas variables climáticas tienen un comportamiento muy similar, posiblemente dado por su misma fuente de origen.

Ahora, respecto a la ciudad de Manizales, se comprueba que para temperaturas frías o muy calientes la radiación solar es muy baja o escasa, pero en temperaturas intermedias que oscilan entre los 15°C y 23°C la concentración de radiación solar es elevada. Esto muestra que en Manizales los valores de radiación solar son altos debido a que su

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

temperatura se encuentra dentro de ese rango en la mayor parte del año, por lo que sus habitantes podrían ser propensos a altos riesgos de afectaciones en la piel.

Con relación al costo computacional, lo que requiere un algoritmo como K-Means para procesar un conjunto de datos compuesto por 3.8 millones de registros es poco. Por otro lado, en tiempos de procesamiento se evidencia que mientras el valor K sea más alto, más tiempo de ejecución le tomará al algoritmo su proceso. En consumo de memoria RAM no existe una tendencia o relación directa con respecto a los escenarios de trabajo establecidos, por lo que los incrementos se dan de forma muy aleatoria. En el consumo de procesador se evidencia que en un escenario con todas las variables y sin normalizar se genera un alto incremento, pero para escenarios donde se normaliza con máximos y mínimos (Range Transformation) se obtiene los incrementos más bajos, por lo que puede que el algoritmo que normaliza con escalas de 0 y 1 sea más corto o eficiente que el Z-transformation. Por otro lado, en consumo de disco duro no se observa relación entre los incrementos de actividad y los escenarios de trabajo establecidos. Por lo tanto, se concluye de manera general , solo para los casos presentados, que sólo en los tiempos de ejecución y consumo de procesador se ven relaciones o tendencias entre los escenarios, mientras que para consumo de RAM y actividad de disco duro no hay una correlación directa entre hardware y la cantidad de registros procesados, cantidad y tipo de variables utilizadas, aplicación de técnicas de normalización y evaluación de calidad de agrupamiento entre los diferentes escenarios de trabajo.

Para futuros trabajos se recomienda emplear técnicas como *Ordinary Kriging* para el manejo de las grandes cantidades de ceros que contiene una variable dentro de un conjunto de datos y observar qué ocurre con los resultados y análisis. Además, llevar a cabo escenarios con un valor K superior a 5 permitiría a los investigadores indagar qué ocurre con el agrupamiento y el rendimiento, tanto a nivel de máquina como en el desempeño del algoritmo.

También se sugiere emplear otros métodos de normalización como transformación de proporción y rango inter-cuartil para ver cómo se comportan los clustering con estos análisis. Asimismo, llevar a cabo pruebas de homocedasticidad e independencia para determinar criterios de normalización podría ser interesante para trabajos futuros. Por otro lado, evaluar datos dentro de una escala temporal (por día, por jornada, etc.) permitiría conocer interesantes conductas climáticas en determinados momentos del día para observar comportamientos meteorológicos en una línea del tiempo.

Otro importante aporte a futuro sería considerar para el conjunto de datos algunos metadatos externos como cercanía entre estaciones meteorológicas, altura sobre el nivel del mar, entre otros, de modo que eso contribuya a mejorar los resultados en cuanto a calidad e interpretabilidad de las clases resultantes. De igual modo, sería interesante realizar procesamientos bajo diferentes escenarios que comprendan un conjunto de datos más grande (por encima de los 35 millones de registros), es decir 10 veces más que el conjunto de datos utilizado para observar el comportamiento computacional a mayor escala con el fin de detectar algún patrón o relación.

Finalmente, con los análisis adicionales, haciendo uso de otro algoritmo de clustering para la validación, y con la aplicación de STFT (Transformada de Fourier de Tiempo Reducido) para identificar si los comportamientos de las variables por las oscilaciones atmosféricas afectan el análisis de clúster realizado, se determina, sin ser concluyentes, que el uso de la normalización en los escenarios propuestos mejora el desempeño sin depender del tipo de algoritmo ni del comportamiento de las variables bajo análisis.

Referencias

[1] Á. Arroyo, Á. Herrero, V. Tricio y E. Corchado, “Analysis of meteorological conditions in Spain by means of clustering techniques,” *J. Appl. Log.*, vol. 24, pp. 76–89, 2017. https://doi.org/10.1016/j.jal.2016.11.026

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

[2] M. A. Asadi Zarch, B. Sivakumar y A. Sharma, "Assessment of global aridity change," *J. Hydrol.*, vol. 520, pp. 300–313, 2015. https://doi.org/10.1016/j.jhydrol.2014.11.033

[3] M. Bador, P. Naveau, E. Gilleland, M. Castellà y T. Arivelo, "Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe," *Weather Clim. Extrem.*, vol. 9, pp. 17–24, 2015. https://doi.org/10.1016/j.wace.2015.05.003

[4] L. Carro-Calvo, C. Ordóñez, R. García-Herrera y J. L. Schnell, "Spatial clustering and meteorological drivers of summer ozone in Europe," *Atmos. Environ.*, vol. 167, pp. 496–510, 2017. https://doi.org/10.1016/j.atmosenv.2017.08.050

[5] M. J. Carvalho, P. Melo-Gonçalves, J. C. Teixeira y A. Rocha, "Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation," *Phys. Chem. Earth*, vol. 94, pp. 22–28, 2016. https://doi.org/10.1016/j.pce.2016.05.001

[6] M. I. Chidean, A. J. Caamaño, J. Ramiro-Bargueño, C. Casanova-Mateo y S. Salcedo-Sanz, "Spatio-temporal analysis of wind resource in the Iberian Peninsula with data-coupled clustering," *Renew. Sustain. Energy Rev.*, vol. 81, June, pp. 2684–2694, 2018. https://doi.org/10.1016/j.rser.2017.06.075

[7] M. I. Chidean, J. Muñoz-Bulnes, J. Ramiro-Bargueño, A. J. Caamaño y S. Salcedo-Sanz, "Spatio-temporal trend analysis of air temperature in Europe and Western Asia using data-coupled clustering," *Glob. Planet. Change*, vol. 129, pp. 45–55, 2015. https://doi.org/10.1016/j.gloplacha.2015.03.006

[8] R. Falquina y C. Gallardo, "Development and application of a technique for projecting novel and disappearing climates using cluster analysis," *Atmos. Res.*, vol. 197, July, pp. 224–231, 2017. https://doi.org/10.1016/j.atmosres.2017.06.031

[9] M. Ghayekhloo, M. Ghofrani, M. B. Menhaj y R. Azimi, "A novel clustering approach for short-term solar radiation forecasting," *Sol. Energy*, vol. 122, pp. 1371–1383, 2015. https://doi.org/10.1016/j.solener.2015.10.053

[10] S. Li, H. Ma, y W. Li, "Typical solar radiation year construction using k-Means clustering and discrete-time Markov chain," *Appl. Energy*, vol. 205, May, pp. 720–731, 2017. https://doi.org/10.1016/j.apenergy.2017.08.067

[11] X. Wang *et al.*, "A stepwise cluster analysis approach for downscaled climate projection - A Canadian case study," *Environ. Model. Softw.*, vol. 49, pp. 141–151, 2013.

[12] Y. Zheng *et al.*, "Assessment of global aridity change," *Ecol. Indic.*, vol. 75, no. September 2016, pp. 151–165, 2016.

[13] Y. Zheng *et al.*, "Vegetation response to climate conditions based on NDVI simulations using stepwise cluster analysis for the Three-River Headwaters region of China," *Ecol. Indic.*,. September 2016, pp. 0–1, 2017. https://doi.org/10.1016/j.ecolind.2017.06.040

[14] J. Parente, M. G. Pereira y M. Tonini, "Space-time clustering analysis of wildfires: The influence of dataset characteristics, fire prevention policy decisions, weather and climate," *Sci. Total Environ.*, vol. 559, pp. 151–165, 2016. https://doi.org/10.1016/j.scitotenv.2016.03.129

[15] F. Mokdad y B. Haddad, "Improved infrared precipitation estimation approaches based on k-means clustering: Application to north Algeria using MSG-SEVIRI satellite data," *Adv. Sp. Res.*, vol. 59, no. 12, pp. 2880–2900, 2017. https://doi.org/10.1016/j.asr.2017.03.027

[16] C. C. Aggarwal y C. K. Reddy, "DATA Custering Algorithms and Applications". CRC Press, 2013.

[17] T. T. Nguyen, A. Kawamura, T. N. Tong, N. Nakagawa, H. Amaguchi y R. Gilbuena, "Clustering spatio-seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta, Vietnam," *J. Hydrol.*, vol. 522,

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

pp. 661–673, 2015. https://doi.org/10.1016/j.jhydrol.2015.01.023

[18] Y. Chen *et al.*, "Air quality data clustering using EPLS method," *Inf. Fusion*, vol. 36, pp. 225–232, 2017.

[19] A. Ruzmaikin y A. Guillaume, "Clustering of atmospheric data by the deterministic annealing," *J. Atmos. Solar-Terrestrial Phys.*, vol. 120, pp. 121–131, 2014. https://doi.org/10.1016/j.jastp.2014.09.009

[20] C. Li, L. Sun, J. Jia, Y. Cai y X. Wang, "Risk assessment of water pollution sources based on an integrated k-means clustering and set pair analysis method in the region of Shiyan, China," *Sci. Total Environ.*, vol. 557–558, pp. 307–316, 2016. https://doi.org/10.1016/j.scitotenv.2016.03.069

[21] T. R. Sivaramakrishnan y S. Meganathan, "Point rainfall prediction using data mining technique," *Res. J. Appl. Sci. Eng. Technol.*, vol. 4, no. 13, pp. 1899–1902, 2012.

[22] C. Marzban y S. Sandgathe, "Cluster Analysis for Verification of Precipitation Fields," *Weather Forecast.*, vol. 21, no. 5, pp. 824–838, 2006. https://doi.org/10.1175/waf948.1

[23] H. Yahyaoui y H. S. Own, "Unsupervised clustering of service performance behaviors," *Inf. Sci. (Ny).*, vol. 422, pp. 558–571, 2018. https://doi.org/10.1016/j.ins.2017.08.065

[24] G. Gan, C. Ma y J. Wu, "Data Clustering: Theory, Algorithms, and Applications". SIAM - Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania. 2007.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, N° 1 (enero-junio 2019). pp. 57-72. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.