ARTÍCULO DE INVESTIGACIÓN / RESEARCH ARTICLE
https://dx.doi.org/10.14482/inde.40.02.622.553

Evaluation of Unsupervised Machine Learning Algorithms with Climate Data

Evaluación de algoritmos de Aprendizaje de Máquina no supervisados con datos climáticos

Juan Sebastián Ramírez*
Néstor Duque-Méndez**

*Universidad Nacional de Colombia.
Departament of Informatics and Computing.
MSc en Computer Systems Administration.
Orcid ID: https://orcid.org/0000-0001-8876-5371.
jsramirezgo@unal.edu.co

**Universidad Nacional de Colombia-
Departament of Informatics and Computing.
PhD en Engineering - Systems.
Orcid ID: https://orcid.org/0000-0002-4608-281X.
ndduqueme@unal.edu.co - Tel 3007876574

ABSTRACT

When using climate data, researchers have difficulty determining the clustering algorithm and the best performing parameters for processing a specific dataset.
We evaluated of the following unsupervised machine learning algorithms: K-means, K-medoids and Linkage-complete, which are applied to three datasets with climatological variables (temperature, rainfall, relative humidity, and solar radiation) for three meteorological stations located in the department of Caldas, Colombia, at different heights above sea level. Five scenarios are defined for 2, 3, and 5 clusters for each of the two partitioned algorithms, and five scenarios for the hierarchical algorithm, in each one of the meteorological stations. Different quantities and groupings of variables are applied for the different scenarios by using Euclidean distance. Davis-Bouldin is the applied method of quality evaluation of clusters. Normalization with techniques such as range-transformation and Z-transformation, as well as some iterations of the algorithm and reduction of dimensionality with PCA. In addition, the computational cost is evaluated. This study can guide researchers on certain decisions in cluster analysis used in meteorological data, as well as identify the most important algorithm and parameters to take into consideration for the best performance, according to particular conditions and requirements.

Keywords: Climate, clustering, machine learning, K-means, K-medoids.

RESUMEN

Al usar datos climáticos, los investigadores tienen dificultades para determinar el algoritmo de agrupamiento y los parámetros de mejor rendimiento para procesar un conjunto de datos específico.
Se realiza la evaluación de algoritmos de aprendizaje automático no supervisados K-means, K-medoids y Linkage-complete, aplicados a tres conjuntos de datos con variables climatológicas (temperatura, lluvia, humedad relativa y radiación solar), para tres estaciones meteorológicas ubicadas en el departamento de Caldas, Colombia, a diferentes alturas sobre el nivel del mar. Se definen 5 escenarios para 2, 3 y 5 clústeres para cada uno de los dos algoritmos particionados y 5 escenarios para el algoritmo jerárquico, para cada una de las estaciones meteorológicas, y aplicando una cantidad y agrupación diferente de variables para los diferentes escenarios y utilizando la distancia euclidiana, Davis-Bouldin como método de evaluación de calidad de clústeres, normalización con técnicas como transformación de rango y transformación Z, varias iteraciones del algoritmo y reducción de dimensionalidad con PCA. Además, se evalúa el costo computacional. Esta investigación puede guiar al investigador sobre ciertas decisiones en el análisis de conglomerados utilizados en datos meteorológicos, así como identificar el algoritmo y los parámetros más importantes a considerar para el mejor desempeño, de acuerdo con las condiciones y requisitos particulares.

Palabras claves: Agrupamiento, aprendizaje de máquina, clima, K-means, K-medoids.

Fecha de recepción: 25 de marzo de 2022
Fecha de aceptación: 16 de agosto de 2022

INTRODUCTION

Climate and atmospheric scenarios have been approached by a variety of researchers to acquire knowledge of interest. Environmental, climatic, and meteorological information has been used to determine behaviors and patterns within the studied area [1]–[12], and air pollutants have been used to understand the formation and impact of natural disasters and greenhouse effect within a region [13]–[16], be it to predict situations, relate causes and effects, and take measurements of the area to finally provide improvements, conclusions, and considerations in favor of the environment.

Currently, many algorithms are used to process records in the analysis of climate data, as shown in Table 1, which compiles more than 30 studies that used clustering algorithms for climate data, using various sources of information, records, timelines, and various objectives.

Table 1. Bibliographic review for clustering algorithms with climatic data.

Clustering	Algorithm used	Method used	Data used	Data timeline	Authors	Year
Hierarchical clustering	Agglomerative	Linkage-Average and Linkage-Complete.	Temperature, wind speed, solar radiation, and atmospheric pressure.	6 years (daily).	[1]	2017
Partitioned	K-medoids.	N/A	Temperature, wind speed, solar radiation, and atmospheric pressure.	6 years (daily).	[1]	2017
Hierarchical clustering	Agglomerative	Linkage-Complete.	Precipitation and evapotranspiration.	49 years (monthly).	[2]	2015
Partitioned	K-means.	N/A	Maximum and Minimum Temperature and Precipitation.	20 years (daily).	[4]	2016
Hierarchical clustering	Agglomerative	Ward.	Temperature and precipitation.	19 years (monthly).	[7]	2017
Partitioned	K-means.	N/A	Temperature and precipitation.	19 years (monthly).	[7]	2017

Continúa...

Clustering	Algorithm used	Method used	Data used	Data timeline	Authors	Year
Others	Stepwise Cluster Analysis (SCA).	N/A	Climatic (temperature and precipitation).	20 years (daily for temperature and monthly for precipitation).	[10]	2013
Hierarchical clustering	Agglomerative	Not defined.	Precipitation.	2 years (periodicity not defined).	[17]	2019
Hierarchical clustering	Agglomerative	Ward.	Wind temperature, humidity, and speed wind.	13 years (hourly).	[18]	2018
Hierarchical clustering	Agglomerative	Not defined.	Temperature, solar radiation, and precipitation.	20 years (daily).	[19]	2016
Partitioned	K-medoids.	N/A	Temperature, solar radiation, and precipitation.	20 years (daily).	[19]	2016
Partitioned	K-means.	N/A	Temperature.	25 years (hourly).	[20]	2019
Partitioned	K-means.	N/A	Solar radiation.	1 year (daily).	[21]	2014
Partitioned	K-means.	N/A	Temperature, wind speed, and relative humidity.	4 years (daily).	[22]	2012
Partitioned	K-means.	N/A	Precipitation.	14 years (daily).	[23]	2018
Partitioned	K-means.	N/A	Wind temperature, speed and direction, precipitation, relative humidity, and solar radiation.	6 years (daily).	[24]	2018
Partitioned	K-means.	N/A	Temperature and solar radiation.	3 years (daily).	[25]	2015
Partitioned	K-means.	N/A	Wind speed.	1 year (daily).	[26]	2019
Partitioned	K-medoids.	N/A	Wind speed.	1 year (daily).	[26]	2019
Partitioned	K-means.	N/A	Solar radiation.	Not defined.	[27]	2019
Partitioned	K-means.	N/A	Solar radiation.	2 years (hourly).	[28]	2019

Continúa...

Clustering	Algorithm used	Method used	Data used	Data timeline	Authors	Year
Partitioned	K-means.	N/A	Temperature, relative humidity, wind speed, and solar radiation.	2 years (daily).	[29]	2018
Partitioned	K-means.	N/A	Satellite imagery of precipitation observations.	2 years (daily).	[30]	2017
Partitioned	K-means.	N/A	Solar radiation.	Not defined.	[31]	2017
Partitioned	K-means.	N/A	Solar radiation, temperature, wind speed and direction.	2 years (hourly).	[32]	2015
Partitioned	K-medoids.	N/A	Maximum temperature.	63 years (daily).	[33]	2015
Partitioned	K-medoids.	N/A	Maximum and minimum temperature.	40 years (daily).	[34]	2018
Others	STPSS (Space-time Permutation Scan Statistics) [52].	N/A	Forest and climatic (temperature, wind speed and relative humidity).	30 years (daily).	[35]	2016
Others	SODCC (Second Order Data Coupled Clustering) [52].	N/A	Air temperature.	70 years (monthly).	[36]	2015
Others	SODCC (Second Order Data Coupled Clustering) [52].	N/A	Wind speed.	35 years (monthly).	[37]	2018
Others	Stepwise Cluster Analysis (SCA).	N/A	Climatic (temperature and precipitation).	13 years (monthly).	[38]	2017

Source: the Authors.

As shown in Table 1, K-means, K-medoids, and hierarchical grouping are the clustering algorithms most used by the authors.

Researchers approached various clustering algorithms (Agglomerative, K-means) [1], with various methods where they applied climate data with a series of metrics to find performance, especially computational performance. Other studies used clustering tools to observe the behavior of the data according to the number of established clusters [22]. Different works recommended some grouping models for specific environmental data by looking for the best projections of particulate matter in the studied region [24]. Also it was demonstrated the proper techniques to view annual temperature trends [34]. Other studies obtained, with proposed partitioned algorithms, the best precipitation estimates in the studied regions [30] and it was used clustering to improve noise reduction in analysis of solar radiation and temperature parameters [19]. Finally, other works used comparisons between unsupervised algorithms using climate data to find higher returns on them [32].

However, there is no clear guidance on which algorithm and parameters specifically serve to obtain the best results with the available data. Therefore, this research focused on working in that gray area to know and understand the behavior of some unsupervised machine learning algorithms applied to various scenarios with climate data.

In order to address this issue from experimentation, the guiding question of the paper is proposed as: How do clustering algorithms behave in different scenarios for climate data processing?

METHODOLOGY

The methodology includes the stages of variable selection, definition of the climatic seasons, obtaining the data set, definition of the scenarios, creation of the scenarios, and choice of tool for the execution of the algorithms. Then, it includes the results and its analysis.

Selection of Climatic Variables

To carry out the investigation, the following four meteorological variables were taken into account: temperature, precipitation, relative humidity, and solar radiation. These variables were selected for being the most reported in the state-of-the-art review in climate research with clustering algorithms [1], [2], [4], [7], [10], [20], [29], [31]–[33], [37]–[39].

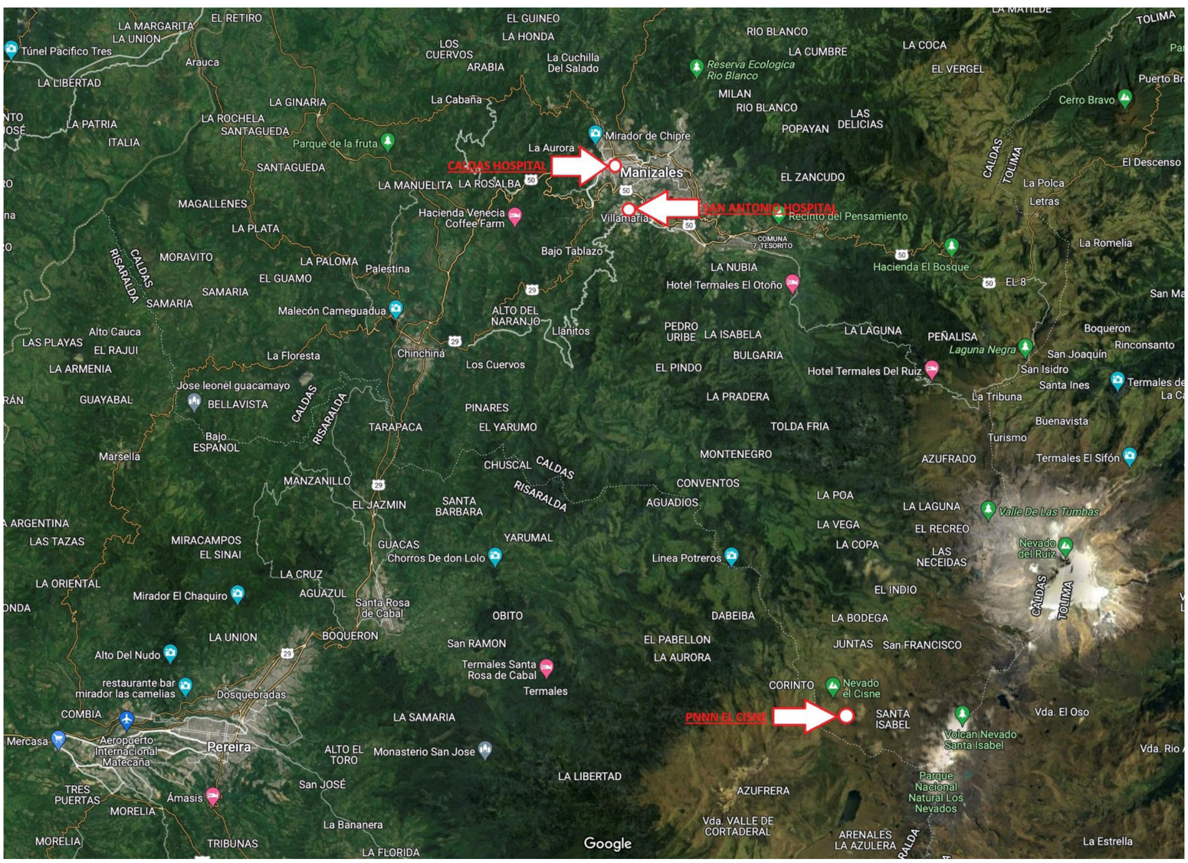
Definition of the Climatic Stations

Data from three meteorological stations called Villamaría Hospital, Caldas Hospital, and Los Nevados National Natural Park (El Cisne) were used, which comprised of 430.635, 530.802, and 248.297 instances, respectively. Table 2 shows the information of the stations.

Table 2. Weather station information

Station name	Typology	Altitude (masl)	Location
Villamaría Hospital	Meteorological	1790	Hospital San Antonio [5.046444567489962, -75.51416420480965]
Caldas Hospital		2183	Hospital de Caldas [5.0630587542854o8, -75.50080152474081]
PNNN El Cisne		4812	Los Nevados National Natural Park [4.830855102285063, -75.42380991366623]

Source: the Authors.



Source: the Authors.

Figure 1. Location of the three meteorological stations on a satellite map of Caldas Department, Colombia.

Obtaining the Data Set

To obtain the records, the data warehouse of the Caldas Environmental Data and Indicators Center (CDIAC) was accessed. The data warehouse is a climate records storage system for the entire department of Caldas, administered and managed by the Adaptive Intelligent Environment Group (GAIA). It is also a project lead by the IDEA (Environmental Studies Institute) of the National University of Colombia, Manizales branch. The data warehouse is a large storage structure implemented in PostgreSQL that houses more than 60 million environmental data, whose information is collected from more than 100 stations, including meteorological stations located in different geographic sectors throughout the department, and whose information can be viewed through http://cdiac.manizales.unal.edu.co. SQL queries were executed to extract the required data from the data warehouse to form the datasets. The records are between April 12, 2012, and August 16, 2017 (time range that contains the whole required data), comprising 64 months (5.3 years) with a data periodicity of every 5 minutes, and for this period information is extracted from the four climatic variables to be analyzed. Table 3 shows the retrieved datasets.

Table 3. Established datasets for the evaluation of unsupervised algorithms

Dataset name	Number of variables	Variable names	Total ecords	Date range
Dataset_HospitalDeVillamaria_64months.xlsx	4	Temperature, precipitation, relative humidity, and solar radiation.	430.635	April 12, 2012, to August 16, 2017
Dataset_HospitalDeCaldas_64months.xlsx			530.802	
Dataset_El CisnePNNN_64months.xlsx			248.297	

Source: the Authors.

Definition of Scenarios

The scenarios are the defined environments where the clustering algorithms are applied, with a diversity of characteristics and parameters to, therefore, analyze and understand how the algorithms behaved in each of these scenarios. The parameters for each scenario are number of variables, type of variable, number of records, missing data, and presence of outliers, which, combined with the characteristics of each station, represent an interesting spectrum for the evaluation of the algorithms.

Different algorithms and modifications in some execution parameters are applied on the data related to the defined scenarios and on the selected stations.

Clustering algorithms: The three clustering algorithms most used by researchers in the analysis of climate data were selected. Agglomerative hierarchical grouping with Linkage-complete, K-means, and K-medoids for partitioned grouping.

Number of clusters (K): The generation of three different groupings whose K value were 2, 3, and 5 was proposed. This selection was based on results from other authors [1], this was corroborated by applying the elbow method in one of the cases, which validated these ranges.

Normalization: For experimentation, normalization with Z-transformation and range-transformation was used as part of the process of reducing value scales in the variables.

Dimensionality reduction: Principal Component Analysis (PCA) was used. The used variables were the four ones: relative humidity, temperature, precipitation, and solar radiation.

Number of algorithm iterations: For experimentation, iteration values of 1, 10, 100 and 1000 were used, where each value is ten times greater than the previous one.

Distance measurement: Euclidean distance was used as the distance function, considered to be the most reliable [1], and being used in a wide variety of jobs in the climate field [2], [7], [18], [19], [30], [33], [40], [41].

Cluster Quality Assessment: This metric consists of evaluating the result of the grouping to determine the quality of the clustering. For experimentation, the Davis-Bouldin index was used as a proposed metric for evaluating cluster quality [1], [22], [42]–[45].

Scenario Creation

Some work scenarios are defined for each one of the three algorithms. These scenarios are configurations to be taken into account in the executions to test each algorithm with different metrics.

Table 4. Work scenarios for the K-means algorithm.

		Clustering strategy for K-means (partitioning clustering)					
		For K=2, K=3 and K=5					
Scenarios		Treatment	Normalization	Dimensionality reduction	Number of iterations	Evaluation criteria	Evaluation criteria
#1	Number of variables: 3 Variable type: temperature, relative humidity, and solar radiation. Number of records: the whole dataset. Missing data: yes. Outliers: yes.	1.1	NO	NO	100	Euclidean	Davis-Bouldin
		1.2	Range-transformation	NO	1000		
		1.3	Z-transformation	PCA with 3 components	10		
#2	Number of variables: 3 Variable type: temperature, relative humidity, and solar radiation. Number of records: the whole dataset. Missing data: no. Outliers: no.	2.1	NO	NO	100	Euclidean	Davis-Bouldin
		2.2	Range-transformation	NO	1000		
		2.3	Z-transformation	PCA with 3 components	10		
#3	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: the whole dataset. Missing data: yes. Outliers: no.	3.1	NO	NO	100	Euclidean	Davis-Bouldin
		3.2	Range-transformation	NO	1000		
		3.3	Z-transformation	PCA with 3 components	10		
#4	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: the whole dataset. Missing data: no. Outliers: yes.	4.1	NO	NO	100	Euclidean	Davis-Bouldin
		4.2	Range-transformation	NO	1000		
		4.3	Z-transformation	PCA with 3 components	10		
#5	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: the whole dataset. Missing data: no. Outliers: no.	5.1	NO	NO	100	Euclidean	Davis-Bouldin
		5.2	Range-transformation	NO	1000		
		5.3	Z-transformation	PCA with 3 components	10		

Source: the Authors [52]

Table 5. Work scenarios for the K-medoids algorithm.

Scenarios		Clustering strategy for K-medoids (partitioning clustering)					
		For K=2, K=3 and K=5					
		Treatment	Normalization	Dimensionality reduction	Number of iterations	Distance criteria	Evaluation criteria
#1	Number of variables: 3 Variable type: temperature, relative humidity, and solar radiation. Number of records: 10.000. Missing data: yes. Outliers: yes.	1.1	NO	NO	100	Euclidean	Davis-Bouldin
		1.2	Range-transformation	NO	1000		
		1.3	Z-transformation	PCA with 3 components	10		
#2	Number of variables: 3 Variable type: temperature, relative humidity, and solar radiation. Number of records: 10.000. Missing data: no. Outliers: no.	2.1	NO	NO	100	Euclidean	Davis-Bouldin
		2.2	Range-transformation	NO	1000		
		2.3	Z-transformation	PCA with 3 components	10		
#3	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: 10.000. Missing data: yes. Outliers: no.	3.1	NO	NO	100	Euclidean	Davis-Bouldin
		3.2	Range-transformation	NO	1000		
		3.3	Z-transformation	PCA with 3 components	10		
#4	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: 10.000. Missing data: no. Outliers: yes.	4.1	NO	NO	100	Euclidean	Davis-Bouldin
		4.2	Range-transformation	NO	1000		
		4.3	Z-transformation	PCA with 3 components	10		
#5	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: 10.000. Missing data: no. Outliers: no.	5.1	NO	NO	100	Euclidean	Davis-Bouldin
		5.2	Range-transformation	NO	1000		
		5.3	Z-ransformation	PCA with 3 components	10		

Source: the Authors [52]

Table 6. Work scenarios for the agglomerative algorithm with the Linkage-Complete method.

Scenarios		Clustering strategy for agglomerative hierarchical grouping (Linkage-complete method)		
		Treatment	Normalization	Distance criteria
#1	Number of variables: 3 Variable type: temperature, relative humidity, and solar radiation. Number of records: 5.000. Missing data: yes. Outliers: yes.	1.1	NO	Euclidean
		1.2	Range-transformation	
		1.3	Z-transformation	
#2	Number of variables: 3 Variable type: temperature, relative humidity, and solar radiation. Number of records: 5.000. Missing data: no. Outliers: no.	2.1	NO	Euclidean
		2.2	Range-transformation	
		2.3	Z-transformation	
#3	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: 5.000. Missing data: yes. Outliers: no.	3.1	NO	Euclidean
		3.2	Range-transformation	
		3.3	Z-transformation	
#4	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: 5.000. Missing data: no. Outliers: yes.	4.1	NO	Euclidean
		4.2	Range-transformation	
		4.3	Z-transformation	
#5	Number of variables: 4 Variable type: temperature, relative humidity, solar radiation, and precipitation. Number of records: 5.000. Missing data: no. Outliers: no.	5.1	NO	Euclidean
		5.2	Range-transformation	
		5.3	Z-transformation	

Choice of Tool to Execute the Algorithms

To create the scenarios and run the algorithms, we chose to use Rapid Miner (version 9.2), a data mining software used for the analysis of a data set using a variety of operators, tools, and functionalities. It has been used by the scientific community in environmental issues [25], [46]–[49] given the versatility and options it allows, as well as the confidence it generates due to its proven effectiveness.

RESULTS

The results obtained from algorithm and hardware performance for K-means, K-medoids, and Hierarchical Grouping for each of the stations are presented in Tables 7, 8, 9, 10, 11, 12, and 13. Each one of the three weather stations of the region (Villamaría Hospital, Caldas Hospital, and Los Nevados National Natural Park) are K = 2, K = 3, and K = 5, respectively.

Table 7. Algorithm and hardware performance results for K-means with K=2 for the Villamaría Hospital, Caldas Hospital and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Algorithm performance					Hardware performance		
				Clustering				Cluster Assessment Criterion: Davis-Bouldin Index			
				Clúster 0	%	Clúster 1	%		Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
Villamaria Hospital	#1	1.1	523.411	523.378	100,0	33	0,0	-0.273	4.764	2.033.904.144	203.125.000
		1.2	523.411	126.466	24,2	396.945	75,8	-0.643	6.884	2.219.269.632	281.250.000
		1.3	523.411	408.926	78,1	114.485	21,9	-0.658	4.481	1.660.873.048	125.000.000
	#2	2.1	430.634	86.671	20,1	343.963	79,9	-0.469	6.048	3.093.277.640	171.875.000
		2.2	430.634	128.778	29,9	301.856	70,1	-0.812	7.783	2.208.655.688	140.625.000
		2.3	430.634	132.287	30,7	298.347	69,3	-0.811	4.869	3.934.525.592	171.875.000
	#3	3.1	520.819	427.583	82,1	93.236	17,9	-0.456	7.059	3.833.508.712	125.000.000
		3.2	520.819	395.606	76,0	125.213	24,0	-0.784	9.977	4.800.722.456	140.625.000
		3.3	520.819	127.177	24,4	393.642	75,6	-0.793	5.061	5.120.613.720	140.625.000
	#4	4.1	430.634	86.671	20,1	343.963	79,9	-0.469	5.726	1.685.900.296	78.125.000
		4.2	430.634	128.778	29,9	301.856	70,1	-0.812	8.940	2.844.027.936	140.625.000
		4.3	430.634	298.402	69,3	132.232	30,7	-0.767	3.976	3.136.620.896	93.750.000
	#5	5.1	430.634	86.671	20,1	343.963	79,9	-0.469	6.222	4.175.675.576	93.750.000
		5.2	430.634	128.778	29,9	301.856	70,1	-0.812	8.886	5.328.579.768	93.750.000
		5.3	430.634	298.402	69,3	132.232	30,7	-0.767	3.904	4.115.358.056	140.625.000

Continúa...

Station	Scenarios	Treatment	Number of items	Algorithm performance					Hardware performance		
				Clustering				Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%				
Caldas Hospital	#1	1.1	535.605	535.604	100,0	1	0,0	-0.000	2.299	2.333.531.112	281.250.000
		1.2	535.605	160.633	30,0	374.972	70,0	-0.715	7.863	1.115.104.408	156.250.000
		1.3	535.605	157.143	29,3	378.462	70,7	-0.717	5.187	1.312.563.032	109.375.000
	#2	2.1	530.801	442.611	83,4	88.190	16,6	-0.514	8.043	2.610.129.464	109.375.000
		2.2	530.801	385.988	72,7	144.813	27,3	-0.932	8.985	2.178.955.064	171.875.000
		2.3	530.801	153.104	28,8	377.697	71,2	-0.928	5.589	1.191.077.624	125.000.000
	#3	3.1	535.585	447.155	83,5	88.430	16,5	-0.513	7.966	2.592.279.952	109.375.000
		3.2	535.585	390.921	73,0	144.664	27,0	-0.934	10.998	2.531.601.632	125.000.000
		3.3	535.585	382.131	71,3	153.454	28,7	-0.892	5.113	3.572.928.056	93.750.000
	#4	4.1	530.801	442.611	83,4	88.190	16,6	-0.514	8.600	2.584.525.768	109.375.000
		4.2	530.801	385.984	72,7	144.817	27,3	-0.933	11.235	4.659.393.576	125.000.000
		4.3	530.801	153.182	28,9	377.619	71,1	-0.891	5.234	3.812.005.904	171.875.000
	#5	5.1	530.801	442.611	83,4	88.190	16,6	-0.514	8.350	4.251.406.208	171.875.000
		5.2	530.801	385.984	72,7	144.817	27,3	-0.933	11.405	4.278.010.648	171.875.000
		5.3	530.801	153.182	28,9	377.619	71,1	-0.891	5.405	4.562.623.736	125.000.000
PNNN El Cisne	#1	1.1	269.755	269.750	100,0	5	0,0	-0.128	1.390	3.390.779.184	234.375.000
		1.2	269.755	221.549	82,1	48.206	17,9	-0.194	1.641	2.667.055.904	78.125.000
		1.3	269.755	269.750	100,0	5	0,0	-0.133	2.453	1.124.003.040	93.750.000
	#2	2.1	248.296	213.494	86,0	34.802	14,0	-0.561	3.469	3.455.568.232	31.250.000
		2.2	248.296	219.896	88,6	28.400	11,4	-0.330	1.907	2.942.050.336	78.125.000
		2.3	248.296	198.722	80,0	49.574	20,0	-1.051	2.656	3.259.417.352	78.125.000
	#3	3.1	269.551	232.749	86,3	36.802	13,7	-0.563	5.032	3.381.520.488	78.125.000
		3.2	269.551	47.948	17,8	221.603	82,2	-0.279	2.549	3.201.284.992	78.125.000
		3.3	269.551	216.348	80,3	53.203	19,7	-0.911	2.813	2.518.188.752	62.500.000
	#4	4.1	248.296	213.494	86,0	34.802	14,0	-0.561	4.548	4.165.905.120	109.375.000
		4.2	248.296	219.896	88,6	28.400	11,4	-0.330	1.938	2.376.130.872	62.500.000
		4.3	248.296	197.787	79,7	50.509	20,3	-0.902	2.657	3.347.225.912	46.875.000
	#5	5.1	248.296	213.494	86,0	34.802	14,0	-0.561	4.173	4.997.481.640	62.500.000
		5.2	248.296	219.896	88,6	28.400	11,4	-0.330	1.923	3.593.737.400	62.500.000
		5.3	248.296	197.787	79,7	50.509	20,3	-0.902	2.485	4.324.097.544	46.875.000

Source: the Authors.

Table 8. Algorithm and hardware performance results for K-means with K=3 for the Villamaría Hospital, Caldas Hospital, and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Algorithm performance							Hardware performance		
				Clustering						Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%				
Villamaria Hospital	#1	1.1	523.411	523.374	100,0	16	0,0	21	0,0	-0.454	4.028	3.127.642.936	109.375.000
		1.2	523.411	241.672	46,2	194.955	37,2	86.784	16,6	-0.650	10.698	3.225.228.720	171.875.000
		1.3	523.411	408.931	78,1	33	0,0	114.447	21,9	-0.530	6.878	3.921.185.408	93.750.000
	#2	2.1	430.634	40.916	9,5	297.303	69,0	92.415	21,5	-0.434	8.186	4.740.428.440	187.500.000
		2.2	430.634	100.969	23,4	276.204	64,1	53.461	12,4	-0.938	17.647	2.323.638.472	171.875.000
		2.3	430.634	87.686	20,4	286.314	66,5	56.634	13,2	-1.004	5.815	4.314.778.744	109.375.000
	#3	3.1	520.819	380.290	73,0	98.728	19,0	41.801	8,0	-0.433	19.436	3.924.269.616	218.750.000
		3.2	520.819	360.192	69,2	50.415	9,7	110.212	21,2	-0.904	24.907	3.998.961.872	203.125.000
		3.3	520.819	393.656	75,6	127.162	24,4	1	0,0	-0.529	6.740	2.151.677.744	109.375.000
	#4	4.1	430.634	40.916	9,5	297.303	69,0	92.415	21,5	-0.434	8.956	4.436.271.224	78.125.000
		4.2	430.634	53.461	12,4	100.969	23,4	276.204	64,1	-0.938	21.793	2.593.872.720	93.750.000
		4.3	430.634	132.239	30,7	298.394	69,3	1	0,0	-0.512	5.905	3.880.938.824	109.375.000
	#5	5.1	430.634	40.916	9,5	297.303	69,0	92.415	21,5	-0.434	8.647	4.036.573.240	109.375.000
		5.2	430.634	53.461	12,4	100.969	23,4	276.204	64,1	-0.938	21.465	4.026.734.904	140.625.000
		5.3	430.634	132.239	30,7	298.394	69,3	1	0,0	-0.512	5.651	5.611.715.328	78.125.000
Caldas Hospital	#1	1.1	535.605	535.603	100,0	1	0,0	1	0,0	-0.000	2.531	1.614.291.064	156.250.000
		1.2	535.605	78.611	14,7	207.695	38,8	249.299	46,5	-0.776	12.653	1.760.767.520	140.625.000
		1.3	535.605	378.129	70,6	157.475	29,4	1	0,0	-0.478	6.504	1.816.672.888	125.000.000
	#2	2.1	530.801	369.704	69,7	37.049	7,0	124.048	23,4	-0.437	8.827	2.378.210.816	125.000.000
		2.2	530.801	335.244	63,2	52.390	9,9	143.167	27,0	-0.954	20.624	3.460.596.528	250.000.000
		2.3	530.801	51.951	9,8	347.142	65,4	131.708	24,8	-1.009	6.649	3.291.468.888	171.875.000
	#3	3.1	535.585	374.110	69,9	37.153	6,9	124.322	23,2	-0.437	9.576	4.338.760.696	156.250.000
		3.2	535.585	142.009	26,5	52.468	9,8	341.108	63,7	-0.957	20.486	2.832.674.096	171.875.000
		3.3	535.585	381.085	71,2	1.281	0,2	153.219	28,6	-0.748	7.180	4.448.009.200	171.875.000
	#4	4.1	530.801	369.704	69,7	37.049	7,0	124.048	23,4	-0.437	9.769	4.049.652.592	140.625.000
		4.2	530.801	335.236	63,2	52.391	9,9	143.174	27,0	-0.955	23.931	4.460.518.968	218.750.000
		4.3	530.801	376.423	70,9	153.321	28,9	1.057	0,2	-0.739	6.900	4.377.875.896	109.375.000
	#5	5.1	530.801	369.704	69,7	37.049	7,0	124.048	23,4	-0.437	9.891	2.681.809.536	156.250.000
		5.2	530.801	335.236	63,2	52.391	9,9	143.174	27,0	-0.955	24.133	4.951.562.096	156.250.000
		5.3	530.801	376.423	70,9	153.321	28,9	1.057	0,2	-0.739	6.532	4.718.480.048	125.000.000

Continúa...

Station	Scenarios	Treatment	Number of items	Algorithm performance							Hardware performance		
				Clustering						Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%				
PNNN El Cisne	#1	1.1	269.755	269.750	100,0	4	0,0	1	0,0	-0.000	1.425	1.239.912.120	93.750.000
		1.2	269.755	45.460	16,9	209.471	77,7	14.824	5,5	-0.598	9.798	3.133.139.616	109.375.000
		1.3	269.755	221.633	82,2	48.117	17,8	5	0,0	-0.428	3.126	1.078.938.952	62.500.000
	#2	2.1	248.296	55.079	22,2	10.753	4,3	182.464	73,5	-0.496	4.407	2.812.463.560	78.125.000
		2.2	248.296	35.825	14,4	27.932	11,2	184.539	74,3	-0.704	7.235	3.520.337.480	109.375.000
		2.3	248.296	25.939	10,4	180.140	72,6	42.217	17,0	-0.780	3.016	2.440.216.008	62.500.000
	#3	3.1	269.551	201.681	74,8	11.108	4,1	56.762	21,1	-0.509	5.751	1.759.418.656	109.375.000
		3.2	269.551	47.455	17,6	186.916	69,3	35.180	13,1	-0.687	8.735	1.729.306.736	62.500.000
		3.3	269.551	180.454	66,9	44.619	16,6	44.478	16,5	-0.598	3.422	3.286.964.312	93.750.000
	#4	4.1	248.296	55.079	22,2	10.753	4,3	182.464	73,5	-0.496	5.097	2.189.603.096	109.375.000
		4.2	248.296	184.541	74,3	27.932	11,2	35.823	14,4	-0.705	8.563	4.538.102.784	62.500.000
		4.3	248.296	180.114	72,5	25.819	10,4	42.363	17,1	-0.625	3.032	2.782.994.696	62.500.000
	#5	5.1	248.296	55.079	22,2	10.753	4,3	182.464	73,5	-0.496	4.689	4.067.016.224	109.375.000
		5.2	248.296	184.541	74,3	27.932	11,2	35.823	14,4	-0.705	7.891	4.329.025.168	78.125.000
		5.3	248.296	180.114	72,5	25.819	10,4	42.363	17,1	-0.625	2.969	4.691.628.080	46.875.000

Table 9. Algorithm and hardware performance results for K-means with K=5 for the Villamaría Hospital, Caldas Hospital, and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Algorithm performance											Hardware performance		
				Clustering										Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%	Clúster 3	%	Clúster 4	%				
Villamaria Hospital	#1	1.1	523.411	523.373	100,0	10	0,0	10	0,0	12	0,0	6	0,0	-0.427	5.562	3.456.148.088	109.375.000
		1.2	523.411	115.491	22,1	38.943	7,4	100.382	19,2	80.348	15,4	188.247	36,0	-0.862	19.577	2.508.987.032	171.875.000
		1.3	523.411	223.656	42,7	37.007	7,1	33	0,0	81.405	15,6	181.310	34,6	-0.623	9.112	2.521.019.120	156.250.000
	#2	2.1	430.634	26.565	6,2	270.357	62,8	61.231	14,2	49.706	11,5	22.775	5,3	-0.497	22.976	2.412.584.056	203.125.000
		2.2	430.634	93.317	21,7	198.221	46,0	44.557	10,3	60.067	13,9	34.472	8,0	-0.890	23.996	4.166.817.320	187.500.000
		2.3	430.634	98.453	22,9	30.048	7,0	46.224	10,7	57.809	13,4	198.100	46,0	-0.916	7.683	3.956.284.576	109.375.000
	#3	3.1	520.819	348.721	67,0	67.293	12,9	23.456	4,5	27.870	5,4	53.479	10,3	-0.490	24.136	2.632.539.208	250.000.000
		3.2	520.819	188.235	36,1	65.537	12,6	49.100	9,4	180.372	34,6	37.575	7,2	-0.823	22.075	3.485.961.280	171.875.000
		3.3	520.819	214.432	41,2	190.419	36,6	63.353	12,2	1	0,0	52.614	10,1	-0.718	11.987	4.946.395.592	171.875.000
	#4	4.1	430.634	22.775	5,3	270.357	62,8	61.231	14,2	49.706	11,5	26.565	6,2	-0.497	25.440	1.725.702.792	187.500.000
		4.2	430.634	60.070	13,9	44.554	10,3	198.221	46,0	93.317	21,7	34.472	8,0	-0.980	28.568	5.104.033.896	109.375.000
		4.3	430.634	48.639	11,3	1	0,0	215.742	50,1	117.486	27,3	48.766	11,3	-0.720	7.484	3.397.838.840	93.750.000
	#5	5.1	430.634	22.775	5,3	270.357	62,8	61.231	14,2	49.706	11,5	26.565	6,2	-0.497	25.802	3.341.342.488	125.000.000
		5.2	430.634	60.070	13,9	44.554	10,3	198.221	46,0	93.317	21,7	34.472	8,0	-0.890	28.403	5.526.369.600	125.000.000
		5.3	430.634	48.639	11,3	1	0,0	215.742	50,1	117.486	27,3	48.766	11,3	-0.720	7.568	2.757.937.944	109.375.000
Caldas Hospital	#1	1.1	535.605	535.598	100,0	1	0,0	1	0,0	1	0,0	4	0,0	-0.129	4.019	2.444.395.896	171.875.000
		1.2	535.605	80.079	15,0	126.697	23,7	163.030	30,4	33.737	6,3	132.062	24,7	-0.935	28.071	1.670.932.016	250.000.000
		1.3	535.605	181.294	33,8	111.806	20,9	1	0,0	196.948	36,8	45.556	8,5	-0.682	8.773	1.847.093.456	125.000.000
	#2	2.1	530.801	329.841	62,1	63.028	11,9	21.093	4,0	89.740	16,9	27.099	5,1	-0.489	19.969	3.249.132.536	109.375.000
		2.2	530.801	143.487	27,0	33.033	6,2	215.770	40,6	57.216	10,8	81.295	15,3	-0.907	29.546	2.841.549.816	234.375.000
		2.3	530.801	145.658	27,4	30.774	5,8	216.746	40,8	57.604	10,9	80.019	15,1	-0.958	8.812	1.616.400.376	171.875.000
	#3	3.1	535.585	27.190	5,1	334.110	62,4	21.143	3,9	63.178	11,8	89.964	16,8	-0.489	22.345	2.579.633.128	156.250.000
		3.2	535.585	81.763	15,3	212.561	39,7	149.977	28,0	33.245	6,2	58.039	10,8	-0.906	24.884	3.217.508.904	203.125.000
		3.3	535.585	182.880	34,1	249.404	46,6	1.281	0,2	58.176	10,9	43.844	8,2	-0.818	9.564	3.837.787.888	156.250.000
	#4	4.1	530.801	329.841	62,1	63.028	11,9	21.093	4,0	89.740	16,9	27.099	5,1	-0.489	22.929	5.412.221.040	171.875.000
		4.2	530.801	57.215	10,8	215.764	40,6	81.295	15,3	143.494	27,0	33.033	6,2	-0.909	34.545	3.526.053.736	187.500.000
		4.3	530.801	182.661	34,4	44.623	8,4	60.274	11,4	1.540	0,3	241.703	45,5	-0.827	9.047	3.091.454.096	156.250.000
	#5	5.1	530.801	329.841	62,1	63.028	11,9	21.093	4,0	89.740	16,9	27.099	5,1	-0.489	22.847	4.849.954.584	203.125.000
		5.2	530.801	57.215	10,8	215.764	40,6	81.295	15,3	143.494	27,0	33.033	6,2	-0.909	34.030	2.817.044.000	203.125.000
		5.3	530.801	182.661	34,4	44.623	8,4	60.274	11,4	1.540	0,3	241.703	45,5	-0.827	9.187	5.172.186.936	156.250.000

Continúa...

Station	Scenarios	Treatment	Number of items	Algorithm performance											Hardware performance		
				Clustering										Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%	Clúster 3	%	Clúster 4	%				
PNNN El Cisne	#1	1.1	269.755	269.733	100,0	4	0,0	3	0,0	1	0,0	14	0,0	-0.170	2.094	3.594.033.696	109.375.000
		1.2	269.755	88.744	32,9	45.131	16,7	10.883	4,0	23.294	8,6	101.703	37,7	-0.735	8.360	2.908.706.544	46.875.000
		1.3	269.755	94.353	35,0	5	0,0	45.233	16,8	107.518	39,9	22.646	8,4	-0.533	4.407	2.881.614.048	109.375.000
	#2	2.1	248.296	27.429	11,0	159.670	64,3	45.762	18,4	10.188	4,1	5.247	2,1	-0.565	10.048	2.827.335.296	125.000.000
		2.2	248.296	25.837	10,4	148.994	60,0	8.492	3,4	55.491	22,3	9.482	3,8	-0.736	13.753	3.406.107.672	109.375.000
		2.3	248.296	45.952	18,5	25.551	10,3	73.825	29,7	91.247	36,7	11.721	4,7	-0.874	4.079	2.995.070.288	46.875.000
	#3	3.1	269.551	148.291	55,0	21.513	8,0	7.149	2,7	53.430	19,8	39.168	14,5	-0.492	11.861	3.338.408.776	78.125.000
		3.2	269.551	8.734	3,2	151.203	56,1	55.159	20,5	9.507	3,5	44.948	16,7	-0.729	12.143	3.040.802.024	125.000.000
		3.3	269.551	146.177	54,2	62.914	23,3	15.876	5,9	44.518	16,5	66	0,0	-0.533	4.657	4.612.584.552	78.125.000
	#4	4.1	248.296	27.865	11,2	5.278	2,1	45.814	18,5	10.280	4,1	159.059	64,1	-0.566	9.986	5.038.083.944	109.375.000
		4.2	248.296	55.491	22,3	25.837	10,4	9.428	3,8	148.994	60,0	8.492	3,4	-0.737	16.588	4.948.851.528	62.500.000
		4.3	248.296	145.979	58,8	25.472	10,3	14.534	5,9	62.278	25,1	33	0,0	-0.567	4.062	3.868.257.872	78.125.000
	#5	5.1	248.296	27.865	11,2	5.278	2,1	45.814	18,5	10.280	4,1	159.059	64,1	-0.566	9.924	2.230.952.336	62.500.000
		5.2	248.296	55.491	22,3	25.837	10,4	9.482	3,8	148.994	60,0	8.492	3,4	-0.737	16.207	5.144.014.328	109.375.000
		5.3	248.296	145.979	58,8	25.472	10,3	14.534	5,9	62.278	25,1	33	0,0	-0.567	4.047	4.092.733.304	62.500.000

Source: the Authors, from previous work [39]

In a global view, for the Hospital de Villamaría station (low altitude), the evaluation indices for all its scenarios and treatments are observed. Regarding the Hospital de Caldas station (intermediate altitude), the evaluation indices are lower than the previous station (gaining quality), and they also remain similar for the rest of its K values. However, for the El Cisne PNNN station (maximum height), quite the opposite happens: The evaluation Davis-Bouldin indices (clustering lose quality) [1] and remain the same for their K values.

Regarding K-means, iterating the algorithm to form clusters by assigning each point to its closest centroid and recalculating the centroid of each cluster is a very efficient and simple process, not only by executing two steps for each iteration, but also by seeing how it is able to process immense quantities of instances very quickly. In its experimentation with the Hospital de Caldas station, it used more than 530.000 instances. This coincides with the contribution of [22], by defining K-means as a simple and efficient algorithm.

The execution times of the K-means algorithm increase as the value of “k” is greater. This is because you must iterate more times due to the need to create more clusters. For stations such as Hospital de Caldas and Villamaría, the execution times are higher since there are datasets of more than 430,000 instances. For the El Cisne PNN station, the execution times are shorter since they comprise a dataset of less than 250,000 instances.

RAM memory consumption is very similar for all k values and for the three weather seasons. Although it differs in certain work scenarios, the average consumption is 2.6 Gb of RAM. This means that regardless of the characteristics of the scenarios and datasets, RAM uses, on average, the same amount of resources because its consumption is given the minimum it needs to run the algorithm's functionality.

The CPU runtime increases as the value of k increases. This is due to the processing it uses for the number of clusters to generate. The same behavior is observed for the three climatic stations.

Table 10. Algorithm and hardware performance results for K-medoids with K=2 for the Villamaría Hospital, Caldas Hospital, and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Algorithm performance					Hardware performance		
				Clustering				Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%				
Villamaria Hospital	#1	1.1	10.000	266	2,7	9.733	97,3	-0.320	193.867	2.200.492.256	193.250.000.000
		1.2	10.000	1.933	19,3	8.066	80,7	-0.545	174.315	1.263.587.808	173.750.000.000
		1.3	10.000	1.764	17,6	8.235	82,4	-0.549	202.442	1.178.103.296	201.796.875.000
	#2	2.1	7.407	6.449	87,1	958	12,9	-0.358	108.984	1.344.279.040	108.359.375.000
		2.2	7.407	1.890	25,5	5.517	74,5	-0.738	108.913	912.989.216	108.468.750.000
		2.3	7.407	1.717	23,2	5.690	76,8	-0.747	105.596	1.710.420.488	105.203.125.000
	#3	3.1	7.407	6.449	87,1	958	12,9	-0.358	155.486	1.633.243.208	149.703.125.000
		3.2	7.407	1.890	25,5	5.517	74,5	-0.740	138.827	1.191.412.048	134.968.750.000
		3.3	7.407	1.683	22,7	5.724	77,3	-0.713	111.392	2.615.978.848	108.312.500.000
	#4	4.1	7.407	6.449	87,1	958	12,9	-0.358	147.896	2.265.506.080	145.234.375.000
		4.2	7.407	1.890	25,5	5.517	74,5	-0.740	233.622	1.661.834.976	229.859.375.000
		4.3	7.407	1.683	22,7	5.724	77,3	-0.713	111.786	2.081.409.144	111.218.750.000
	#5	5.1	7.407	6.449	87,1	958	12,9	-0.358	148.653	1.870.534.696	147.750.000.000
		5.2	7.407	1.890	25,5	5.517	74,5	-0.740	134.652	2.163.714.384	133.968.750.000
		5.3	7.407	1.683	22,7	5.724	77,3	-0.713	112.582	1.952.800.304	111.921.875.000

Continúa...

Station	Scenarios	Treatment	Number of items	Algorithm performance					Hardware performance		
				Clustering				Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%				
Caldas Hospital	#1	1.1	10.000	310	3,1	9.689	96,9	-0.356	221.390	2.251.278.824	220.062.500.000
		1.2	10.000	9.846	98,5	153	1,5	-1.198	243.176	1.452.391.784	241.671.875.000
		1.3	10.000	9.846	98,5	153	1,5	-1.199	247.809	1.855.518.232	246.296.875.000
	#2	2.1	1.859	626	33,7	1.233	66,3	-1.023	9.664	1.743.473.272	8.781.250.000
		2.2	1.859	1.715	92,3	144	7,7	-2.073	11.001	2.605.499.600	10.953.125.000
		2.3	1.859	1.718	92,4	141	7,6	-2.279	11.814	937.706.456	11.687.500.000
	#3	3.1	9.979	8.907	89,3	1.072	10,7	-1.171	293.887	1.015.890.144	293.062.500.000
		3.2	9.979	153	1,5	9.826	98,5	-1.692	379.152	1.114.897.968	377.937.500.000
		3.3	9.979	144	1,4	9.835	98,6	-1.588	287.272	1.783.789.144	285.890.625.000
	#4	4.1	1.859	626	33,7	1.233	66,3	-1.023	10.846	1.987.641.752	10.671.875.000
		4.2	1.859	1.715	92,3	144	7,7	-2.078	13.439	2.446.702.624	13.265.625.000
		4.3	1.859	139	7,5	1.720	92,5	-2.290	11.877	2.295.816.056	11.671.875.000
	#5	5.1	1.859	626	33,7	1.233	66,3	-1.023	11.001	2.526.768.848	10.750.000.000
		5.2	1.859	626	33,7	1.233	66,3	-1.023	10.330	1.941.857.872	10.125.000.000
		5.3	1.859	626	33,7	1.233	66,3	-1.023	8.830	2.320.046.864	8.656.250.000
PNNN El Cisne	#1	1.1	10.000	410	4,1	9.589	95,9	-0.734	243.920	1.804.158.433	242.406.250.000
		1.2	10.000	1.752	17,5	8.247	82,5	-0.947	220.318	2.041.902.144	219.375.000.000
		1.3	10.000	1.752	17,5	8.247	82,5	-0.979	220.707	1.675.429.112	219.703.125.000
	#2	2.1	8.941	8.736	97,7	205	2,3	-0.477	186.900	1.549.659.112	185.843.750.000
		2.2	8.941	8.583	96,0	358	4,0	-0.949	197.158	2.498.252.656	195.921.875.000
		2.3	8.941	8.684	97,1	257	2,9	-0.948	186.041	1.368.465.320	184.875.000.000
	#3	3.1	9.794	9.589	97,9	205	2,1	-0.481	290.582	1.761.889.408	289.000.000.000
		3.2	9.794	9.436	96,3	358	3,7	-0.825	286.000	1.763.412.464	284.609.375.000
		3.3	9.794	9.297	94,9	497	5,1	-0.882	276.227	2.189.905.888	269.609.375.000
	#4	4.1	8.949	213	2,4	8.736	97,6	-0.493	235.194	2.410.475.048	233.750.000.000
		4.2	8.949	377	4,2	8.572	95,8	-0.957	246.335	1.514.266.320	245.265.625.000
		4.3	8.949	498	5,6	8.451	94,4	-0.985	183.660	2.481.649.192	182.609.375.000
	#5	5.1	8.941	8.736	97,7	205	2,3	-0.477	241.157	1.430.077.856	240.156.250.000
		5.2	8.941	8.583	96,0	358	4,0	-0.950	225.707	1.545.609.656	224.531.250.000
		5.3	8.941	8.472	94,8	469	5,2	-0.963	190.510	2.447.676.192	189.171.875.000

Source: the Authors.

Table 11. Algorithm and hardware performance results for K-medoids with K=3 for the Villamaría Hospital, Caldas Hospital, and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Algorithm performance							Hardware performance		
				Clustering						Cluster Assessment Criterion: Davis-Bouldin Index	Execution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%				
Villamaria Hospital	#1	1.1	10.000	4.136	41,4	266	2,7	5.597	56,0	-0.408	241.634	1.972.852.672	241.000.000.000
		1.2	10.000	3.628	36,3	946	9,5	5.415	54,2	-0.610	290.810	735.634.936	289.937.500.000
		1.3	10.000	5.382	53,8	3.697	37,0	920	9,2	-0.665	242.091	1.037.002.353	241.265.625.000
	#2	2.1	7.407	761	10,3	1.791	24,2	4.855	65,5	-0.690	148.747	1.402.463.664	148.265.625.000
		2.2	7.407	1.282	17,3	1.212	16,4	4.913	66,3	-0.697	117.533	1.594.281.192	117.078.125.000
		2.3	7.407	1.210	16,3	1.188	16,0	5.009	67,6	-0.691	129.348	1.099.413.416	128.921.875.000
	#3	3.1	7.407	761	10,3	1.791	24,2	4.855	65,5	-0.690	205.754	1.663.256.125	201.765.625.000
		3.2	7.407	1.282	17,3	1.212	16,4	4.913	66,3	-0.698	151.722	2.010.673.544	149.203.125.000
		3.3	7.407	1.191	16,1	1.156	15,6	5.060	68,3	-0.661	149.358	1.619.011.792	145.984.375.000
	#4	4.1	7.407	761	10,3	1.791	24,2	4.855	65,5	-0.690	202.013	1.623.001.936	197.109.375.000
		4.2	7.407	1.282	17,3	1.212	16,4	4.913	66,3	-0.698	579.674	2.160.207.008	574.062.500.000
		4.3	7.407	1.191	16,1	1.156	15,6	5.060	68,3	-0.661	142.665	1.423.935.424	142.015.625.000
	#5	5.1	7.407	761	10,3	1.791	24,2	4.855	65,5	-0.690	204.839	2.219.246.504	204.234.375.000
		5.2	7.407	1.282	17,3	1.212	16,4	4.913	66,3	-0.698	148.141	1.404.410.704	147.468.750.000
		5.3	7.407	1.191	16,1	1.156	15,6	5.060	68,3	-0.661	145.308	2.468.807.040	144.406.250.000
Caldas Hospital	#1	1.1	10.000	8.907	89,1	864	8,6	228	2,3	-1.204	252.855	1.710.514.624	251.312.500.000
		1.2	10.000	253	2,5	153	1,5	9.593	95,9	-2.614	404.492	1.969.229.920	401.734.375.000
		1.3	10.000	253	2,5	153	1,5	9.593	95,9	-2.615	385.683	2.424.768.752	383.296.875.000
	#2	2.1	1.859	92	4,9	1.233	66,3	534	28,7	-0.549	14.049	1.967.443.312	13.437.500.000
		2.2	1.859	218	11,7	1.508	81,1	133	7,2	-1.538	13.581	1.970.351.064	13.484.375.000
		2.3	1.859	244	13,1	1.492	80,3	123	6,6	-1.1510	14.721	1.992.878.376	14.609.375.000
	#3	3.1	9.979	9.483	95,0	206	2,1	290	2,9	-14.231	436.077	1.129.489.496	434.062.500.000
		3.2	9.979	9.573	95,9	153	1,5	253	2,5	-3.534	543.465	2.656.269.576	541.062.500.000
		3.3	9.979	9.490	95,1	140	1,4	349	3,5	-1.299	429.631	1.996.304.440	427.187.500.000
	#4	4.1	1.859	92	4,9	1.233	66,3	534	28,7	-0.549	15.159	1.796.356.992	14.937.500.000
		4.2	1.859	218	11,7	1.508	81,1	133	7,2	-1.544	16.377	2.637.514.312	16.218.750.000
		4.3	1.859	243	13,1	1.495	80,4	121	6,5	-1.507	14.361	2.007.995.656	14.125.000.000
	#5	5.1	1.859	92	4,9	1.233	66,3	534	28,7	-0.549	15.315	1.657.532.160	15.125.000.000
		5.2	1.859	92	4,9	1.233	66,3	534	28,7	-0.549	14.986	2.236.052.376	14.656.250.000
		5.3	1.859	92	4,9	1.233	66,3	534	28,7	-0.549	12.846	2.175.093.056	12.656.250.000

Continúa...

Station	Scenarios	Treatment	Number of items	Algorithm performance							Hardware performance		
				Clustering						Cluster Assessment Criterion: Davis-Bouldin Index	Exe-cution time (ms)	RAM memory (bytes)	CPU runtime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%				
PNNN El Cisne	#1	1.1	10.000	364	3,6	46	0,5	9.589	95,9	-1.850	301.659	1.866.359.987	299.406.250.000
		1.2	10.000	1.635	16,4	6.612	66,1	1.752	17,5	-5.950	542.491	2.270.397.200	539.953.125.000
		1.3	10.000	1.635	16,4	6.612	66,1	1.752	17,5	-6.225	557.767	1.099.916.248	555.250.000.000
	#2	2.1	8.941	5.552	62,1	205	2,3	3.184	35,6	-0.740	231.495	2.131.149.912	230.046.875.000
		2.2	8.941	5.492	61,4	3.091	34,6	358	4,0	-2.140	343.098	1.552.125.872	340.937.500.000
		2.3	8.941	5.517	61,7	257	2,9	3.167	35,4	-1.883	297.049	2.470.985.984	294.906.250.000
	#3	3.1	9.794	205	2,1	3.185	32,5	6.404	65,4	-0.740	385.413	1.540.575.368	383.218.750.000
		3.2	9.794	358	3,7	6.344	64,8	3.092	31,6	-1.997	536.350	2.232.436.984	533.218.750.000
		3.3	9.794	497	5,1	2.944	30,1	6.353	64,9	-1.784	581.644	2.102.989.992	574.250.000.000
	#4	4.1	8.949	213	2,4	3.184	35,6	5.552	62,0	-0.746	260.682	1.398.742.896	259.484.375.000
		4.2	8.949	5.485	61,3	3.087	34,5	377	4,2	-2.198	519.749	2.333.937.752	516.671.875.000
		4.3	8.949	8.433	94,2	395	4,4	121	1,4	-0.936	307.420	1.773.292.008	305.484.375.000
	#5	5.1	8.941	5.552	62,1	205	2,3	3.184	35,6	-0.740	271.345	1.659.679.344	269.656.250.000
		5.2	8.941	5.492	61,4	3.091	34,6	358	4,0	-2.146	400.576	2.060.661.576	398.312.500.000
		5.3	8.941	5.502	61,5	469	5,2	2.970	33,2	-1.942	291.886	1.456.914.048	289.875.000.000

Source: the Authors.

Table 12. Algorithm and hardware performance results for K-medoids with K=5 for the Villamaría Hospital, Caldas Hospital, and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Algorithm performance											Hardware performance		
				Clustering										Cluster Assessment Criterion: Davis-Bouldin Index			
				Clúster 0	%	Clúster 1	%	Clúster 2	%	Clúster 3	%	Clúster 4	%		Execution time (ms)	RAM memory (bytes)	CPU untime (ns)
Villa-maria Hospital	#1	1.1	10.000	868	8,7	160	1,6	7.493	74,9	1.372	13,7	106	1,1	-1.050	284.949	1.347.942.856	284.000.000.000
		1.2	10.000	3.047	30,5	373	3,7	946	9,5	1.390	13,9	4.243	42,4	-1.228	413.977	1.483.581.032	410.984.375.000
		1.3	10.000	4.146	41,5	920	9,2	444	4,4	3.141	31,4	1.348	13,5	-1.405	526.314	2.104.265.192	524.484.375.000
	#2	2.1	7.407	803	10,8	495	6,7	708	9,6	4.855	65,5	546	7,4	-0.793	215.558	992.556.472	214.406.250.000
		2.2	7.407	1.927	26,0	607	8,2	870	11,7	3.455	46,6	548	7,4	-1.317	190.604	1.960.383.984	189.453.125.000
		2.3	7.407	1.433	19,3	977	13,2	738	10,0	3.712	50,1	547	7,4	-1.222	188.200	1.478.950.816	187.250.000.000
	#3	3.1	7.407	803	10,8	495	6,7	708	9,6	4.855	65,5	546	7,4	-0.793	270.101	1.254.229.584	26.438.125.000
		3.2	7.407	1.927	26,0	607	8,2	870	11,7	3.455	46,6	548	7,4	-1.323	237.861	2.558.802.672	232.625.000.000
		3.3	7.407	1.334	18,0	924	12,5	732	9,9	3.875	52,3	542	7,3	-1.186	232.208	1.899.579.368	226.500.000.000
	#4	4.1	7.407	803	10,8	495	6,7	708	9,6	4.855	65,5	546	7,4	-0.793	273.943	1.754.420.136	268.687.500.000
		4.2	7.407	1.927	26,0	607	8,2	870	11,7	3.455	46,6	548	7,4	-1.323	227.374	2.608.668.536	226.390.625.000
		4.3	7.407	1.334	18,0	924	12,5	732	9,9	3.875	52,3	542	7,3	-1.186	222.429	1.433.790.416	221.046.875.000
	#5	5.1	7.407	803	10,8	495	6,7	708	9,6	4.855	65,5	546	7,4	-0.793	269.616	1.193.032.328	268.328.125.000
		5.2	7.407	1.927	26,0	607	8,2	870	11,7	3.455	46,6	548	7,4	-1.323	235.302	2.709.839.752	234.078.125.000
		5.3	7.407	1.334	18,0	924	12,5	732	9,9	3.875	52,3	542	7,3	-1.186	222.219	2.245.207.584	220.687.500.000
Caldas Hospital	#1	1.1	10.000	8.097	81,0	0	0,0	0	0,0	864	8,6	228	2,3	-1.056	352.985	2.544.728.376	350.156.250.000
		1.2	10.000	522	5,2	9.114	91,1	210	2,1	153	1,5	0	0,0	-1.280	585.297	2.534.837.680	581.109.375.000
		1.3	10.000	522	5,2	9.114	91,1	210	2,1	153	1,5		0,0	-1.296	566.113	1.770.062.624	561.718.750.000
	#2	2.1	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	1.323	21.708	2.661.847.176	21.546.875.000
		2.2	1.859	195	10,5	81	4,4	1.223	65,8	234	12,6	126	6,8	-1.469	22.943	1.540.433.824	22.750.000.000
		2.3	1.859	199	10,7	726	39,1	469	25,2	342	18,4	123	6,6	-1.663	18.425	2.085.368.088	18.234.375.000
	#3	3.1	9.979	8.641	86,6	0	0,0	1.071	10,7	267	2,7	0	0,0	-1.318	538.464	1.647.820.120	535.453.125.000
		3.2	9.979	8.873	88,9	377	3,8	130	1,3	405	4,1	194	1,9	-1.479	590.326	1.982.162.272	587.203.125.000
		3.3	9.979	8.082	81,0	585	5,9	140	1,4	344	3,4	828	8,3	-1.553	426.013	2.158.830.400	423.312.500.000
	#4	4.1	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	-1.323	24.880	2.276.367.176	24.656.250.000
		4.2	1.859	195	10,5	81	4,4	1.223	65,8	234	12,6	126	6,8	-1.482	27.576	1.980.508.816	27.312.500.000
		4.3	1.859	203	10,9	1.190	64,0	6	0,3	339	18,2	121	6,5	-1.668	19.316	1.548.012.072	19.109.375.000
	#5	5.1	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	-1.323	25.254	1.517.812.944	25.062.500.000
		5.2	1.859	1.204	64,8	83	4,5	274	14,7	178	9,6	120	6,5	-1.323	25.249	2.031.422.200	24.953.125.000
		5.3	1.859	1.204	64,8	83	4,5	274	14,7	173	9,3	125	6,7	-1.330	22.191	2.471.411.264	21.906.250.000

Continúa...

Station	Scenarios	Treatment	Number of items	Algorithm performance											Hardware performance		
				Clustering										Cluster Assessment Criterion: Davis-Bouldin Index	Execu-tion time (ms)	RAM memory (bytes)	CPU untime (ns)
				Clúster 0	%	Clúster 1	%	Clúster 2	%	Clúster 3	%	Clúster 4	%				
PNNN El Cisne	#1	1.1	10.000	364	3,6	1.009	10,1	46	0,5	2.556	25,6	6.024	60,2	-1.307	300.303	1.054.879.160	297.968.750.000
		1.2	10.000	1.752	17,5	1.635	16,4	197	2,0	4.554	45,5	1.861	18,6	-1.738	619.253	2.349.652.840	614.953.125.000
		1.3	10.000	1.752	17,5	1.635	16,4	197	2,0	4.554	45,5	1.861	18,6	-1.751	617.469	1.345.886.568	612.843.750.000
	#2	2.1	8.941	149	1,7	1.536	17,2	1.009	11,3	5.172	57,8	1.075	12,0	-0.655	292.231	2.326.999.856	289.875.000.000
		2.2	8.941	264	3,0	1.276	14,3	1.769	19,8	153	1,7	5.479	61,3	-1.940	919.212	2.471.419.288	911.937.500.000
		2.3	8.941	299	3,3	1.428	16,0	1.591	17,8	110	1,2	5.513	61,7	-2.034	759.524	2.064.959.744	754.359.375.000
	#3	3.1	9.794	1.009	10,3	149	1,5	1.537	15,7	6.024	61,5	1.075	11,0	-0.656	447.823	2.024.522.280	444.718.750.000
		3.2	9.794	1.770	18,1	153	1,6	6.331	64,6	264	2,7	1.276	13,0	-1.751	1.397.919	1.479.066.696	1.370.359.375.000
		3.3	9.794	4.258	43,5	4	0,0	2.944	30,1	2.091	21,3	497	5,1	-1.528	457.258	2.101.205.928	444.531.250.000
	#4	4.1	8.949	1.536	17,2	157	1,8	4.682	52,3	1.704	19,0	870	9,7	-7.937	328.065	2.085.002.352	325.968.750.000
		4.2	8.949	259	2,9	1.261	14,1	5.476	61,2	1.784	19,9	169	1,9	-1.972	924.253	1.591.865.848	917.703.125.000
		4.3	8.949	1.240	13,9	2.935	32,8	395	4,4	121	1,4	4.258	47,6	-1.420	339.431	1.649.443.544	336.640.625.000
	#5	5.1	8.941	149	1,7	1.536	17,2	1.009	11,3	5.172	57,8	1.075	12,0	-0.656	340.625	1.960.132.864	338.265.625.000
		5.2	8.941	264	3,0	1.276	14,3	1.769	19,8	153	1,7	5.479	61,3	-1.948	993.772	1.689.836.640	986.203.125.000
		5.3	8.941	1.764	19,7	1.349	15,1	326	3,6	3	0,0	5.499	61,5	-1.349	331.556	1.903.533.320	330.171.875.000

Source: the Authors, from previous work [39]

Regarding K-medoids for the Hospital de Villamaría station (low altitude), the evaluation index becomes lower as the value of K (number of clusters) increases. For the Hospital de Caldas station (intermediate altitude), the evaluation indices are lower than the previous station and the greater the number of groups, the value of these indices is still lower (gaining quality). However, for the El Cisne PNNN station (maximum height), the opposite is true: the evaluation rates are, once again, higher (losing quality).

For previous partitioned algorithms (K-medoids, K-means), standardization and technique type greatly influence the evaluation of cluster quality. The Davis-Bouldin index, when evaluating the quality of the cluster, generates an approach (and visually verifies) the best grouping result. Furthermore, the higher the K value, the more hardware requirements and time requirements will be demanded to execute and process a dataset, and, subsequently, to execute the algorithm.

Also, note that K-means and K-medoids cannot process empty fields. Some authors omit missing and corrupt data from these algorithms [1], therefore the missing data was transformed to an average value of the attribute. This decision was supported by experts in the matter and made to allow the algorithm to run. Also, the average value corresponds to the whole dataset. We did not want to replace it with the lower or higher value because this would generate dragging of clusters, and it would alter the analysis of the results. The research shows a notable difference between clean datasets versus datasets with missing values that are replaced by an average value of the attribute, since, in the results, there is variation in the grouping evaluation index, which is better when the dataset is clean. On the other hand, the datasets with missing data and atypical data (scenario 1) produced the lowest performance results. This is due in part to there being an imbalance in the formation of the clusters and the evaluation index of their treatments not being the best. This signifies that using a raw dataset is not recommended. Furthermore, the outliers did not affect the results, since the clustering evaluation indices given by scenario 4 are very similar. For example, in scenario 5, which uses a clean dataset. This could be due to the fact that the outliers number was small compared to the dataset size (outliers subject to existence within the dataset), or, conversely, normalization allowed for reducing these large distance margins to provide better groupings.

It also corroborates the idea that applying dimensionality reduction with PCA, where three components are obtained, raises the level of abstraction of the results, since it does not allow for direct visualization of the map of the original attributes. As it was mentioned [42], that data transforming from an original space into a new one with a lower dimension, where they cannot be associated with the characteristics of the original, means that an analysis of the new space is very complicated and complex, since there is no physical meaning for the transformed and obtained characteristics.

Therefore, promoting a PCA with two components could determine the behavior of the data in a two-dimensional plane and make its analysis easier. In turn, this brings the reduction of initial attributes (which are four) to only two. In terms of clustering evaluation, PCA did not influence the improvement of the Davis-Bouldin index.

On the other hand, the number of iterations forces the algorithm to form the clusters and recalculate the centroids more times. However, it reaches a point where it finds the calculation it needs without improving with more iterations. As seen in experimentation, a number of iterations in 100 was a balanced value for working with clustering, where computational performance in terms of execution times is not affected for the algorithm. This prevents an investigator from unnecessarily repeating thousands of times. It is verified that iterating with a larger number does not affect the improvement of the evaluation index (recalculating its centroids to find a suitable value).

Table 13. Hardware performance results for Linkage-Complete for the Villamaría Hospital, Caldas Hospital, and PNNN El Cisne stations

Station	Scenarios	Treatment	Number of items	Hardware performance				
				Execution time (milliseconds)	Time (ms)	RAM memory (bytes)	CPU time (ns)	CPU runtime (ns)
Villamaria Hospital	#1	1.1	5.000	81.787	81.787	4.832.249.328	78.296.875.000	78.296.875.000
		1.2	5.000	80.204	80.210	3.018.301.440	75.671.875.000	75.671.875.000
		1.3	5.000	77.671	77.671	3.998.491.928	74.437.500.000	74.437.500.000
	#2	2.1	2.407	12.669	12.670	4.888.031.432	11.343.750.000	11.343.750.000
		2.2	2.407	9.610	9.610	4.660.395.256	8.828.125.000	8.828.125.000
		2.3	2.407	8.981	8.981	4.352.239.560	8.609.375.000	8.609.375.000
	#3	3.1	2.407	9.299	9.405	6.156.410.744	8.718.750.000	8.718.750.000
		3.2	2.407	8.836	8.836	6.820.019.440	8.250.000.000	8.250.000.000
		3.3	2.407	9.192	9.192	7.073.454.304	8.531.250.000	8.531.250.000
	#4	4.1	2.407	9.307	9.308	3.975.729.592	8.906.250.000	8.906.250.000
		4.2	2.407	9.028	9.028	5.122.837.744	8.562.500.000	8.562.500.000
		4.3	2.407	8.634	8.634	6.198.538.304	8.421.875.000	8.421.875.000
	#5	5.1	2.407	9.076	9.076	4.237.062.432	8.843.750.000	8.843.750.000
		5.2	2.407	9.107	9.107	6.249.893.848	8.562.500.000	8.562.500.000
		5.3	2.407	8.991	8.992	3.511.275.344	8.453.125.000	8.453.125.000
Caldas Hospital	#1	1.1	5.000	78.454	78.454	7.021.733.128	76.203.125.000	76.203.125.000
		1.2	5.000	83.444	83.444	7.359.136.688	78.125.000.000	78.125.000.000
		1.3	5.000	101.343	101.492	7.495.140.664	85.109.375.000	85.109.375.000
	#2	2.1	1.002	796	802	7.095.685.424	671.875.000	671.875.000
		2.2	1.002	1.062	1.062	7.493.057.504	953.125.000	953.125.000
		2.3	1.002	961	964	7.727.727.224	859.375.000	859.375.000
	#3	3.1	4.979	81.650	81.666	5.619.392.168	73.671.875.000	73.656.250.000
		3.2	4.979	90.006	90.006	5.448.990.136	80.234.375.000	80.234.375.000
		3.3	4.979	91.647	91.647	3.879.695.680	77.250.000.000	77.250.000.000
	#4	4.1	1.002	962	964	6.222.305.840	812.500.000	812.500.000
		4.2	1.002	1.044	1.045	5.707.073.944	890.625.000	890.625.000
		4.3	1.002	906	917	4.980.822.496	781.250.000	781.250.000
	#5	5.1	1.002	1.148	1.158	4.803.851.432	953.125.000	953.125.000
		5.2	1.002	967	967	4.266.509.624	765.625.000	765.625.000
		5.3	1.002	811	812	7.467.424.808	812.500.000	812.500.000

Continúa...

Station	Scenarios	Treatment	Number of items	Hardware performance				
				Execution time (milliseconds)	Time (ms)	RAM memory (bytes)	CPU time (ns)	CPU runtime (ns)
PNNN El Cisne	#1	1.1	5.000	262.151	262.153	4.881.553.392	244.703.125.000	244.703.125.000
		1.2	5.000	147.247	147.247	4.865.530.976	143.234.375.000	143.234.375.000
		1.3	5.000	74.749	74.749	6.853.886.360	73.578.125.000	73.578.125.000
	#2	2.1	3.942	46.601	46.601	4.056.895.880	41.812.500.000	41.812.500.000
		2.2	3.942	39.821	39.821	6.966.283.928	37.500.000.000	37.500.000.000
		2.3	3.942	37.694	37.694	3.563.307.024	36.593.750.000	36.593.750.000
	#3	3.1	4.795	67.307	67.307	4.302.086.952	65.484.375.000	65.484.375.000
		3.2	4.795	67.606	67.606	5.474.073.176	65.687.500.000	65.687.500.000
		3.3	4.795	67.010	67.010	4.388.374.624	65.796.875.000	65.796.875.000
	#4	4.1	3.950	37.709	37.709	5.342.942.728	36.906.250.000	36.906.250.000
		4.2	3.950	37.740	37.740	4.204.845.088	37.156.250.000	37.156.250.000
		4.3	3.950	37.835	37.835	5.411.607.272	37.296.875.000	37.296.875.000
	#5	5.1	3.942	36.693	36.693	5.784.841.944	35.984.375.000	35.984.375.000
		5.2	3.942	38.366	38.366	5.807.874.168	37.453.125.000	37.453.125.000
		5.3	3.942	37.070	37.070	5.228.633.656	36.390.625.000	36.390.625.000

Source: the Authors.

For the agglomerative clustering algorithm, we decided to process with 20,000 instances to test the previous algorithm operation and determine the subsequent creation of the scenarios. The processing was found to be too slow. This was due in part to the algorithm presenting great computational complexity. Once a distance measurement is determined and used, a dissimilarity matrix is constructed. This process leads to the generation of a 20,000 x 20,000 size matrix (for a dataset of 20,000 instances), which, in hardware terms, requires storage and processing resources. After this, the data sets are merged at each level and the difference matrix is subsequently updated. This has a great impact on computer processing, and execution takes more than 1 hour and 30 minutes (for a dataset of 20,000 instances). That is, it took 72 times more than the previous 5,000 instance scenarios. This conclusion supports the research of [1], where hierarchical grouping is not recommended for a dataset of more than 10,000 instances. Therefore, it was decided to create scenarios with data sets not exceeding 5,000 instances.

Hierarchical grouping cannot process empty fields. With that said, the missing data was transformed to an average value of the attribute.

In terms of attributes, precipitation makes the dendrogram more complex to analyze, not only because it creates an additional agglomeration in the lower levels, but also because it involves increasing the dataset with thousands of more data. This leads to the graph agglomerate creating many instances, as well as becoming narrow for subsequent visualizations and analyzes. Due to the initial dataset being large, it is recommended to use precipitation for a dataset that guarantees a lower number of instances than those used in this experimentation; that is, below 1,000 instances for the agglomerative algorithm.

Based on the above data, a dendrogram of around 3,000 instances (sheets) can allow an investigator to easily see how the instances merge from the intermediate level, and focus the observation on higher levels, despite the lower levels being impossible. To visualize them, a researcher must evaluate from level 0 of the tree. It is suggested to use data sets of less than 100 instances for the dendrograms to be more visible, allowing better analysis from the lowest levels. Hierarchical grouping is preferred for a small dataset [1], [50].

On the other hand, normalization facilitated the construction of dendrograms, helping the dissimilarity and similarity distances (Y axis) to become closer on a scale between zero and one. This allowed the dendrogram to be viewed in a more simple manner. The dimensionality reduction was not transcendental in the results, therefore, it is concluded that it was not useful for the agglomerative algorithm.

In computational terms, the algorithm uses similar machine resources in all the scenarios, regardless of the preset characteristics. However, if high execution times and CPU times are found for scenario 1 (up to eight times greater than the rest of the scenarios, with only 2,000 instances apart), confirming that using datasets with large instance volumes for agglomerative hierarchical grouping can lead to slow processing.

DISCUSSION

To determine, in a preliminary study, the behavior of clustering algorithms on climate data, stations and datasets with different characteristics, scenarios were defined to which variants of the learning algorithms were applied, and the behavior of the metrics was evaluated.

The results, without being conclusive, can guide people who work with these data in the speedy selection of these elements, which we consider the contribution of this work.

For K-means, at the Hospital de Caldas station, there are more clustering evaluations with better quality compared to the other two stations. This is determined by taking a value as a reference to make the count. In this case, the indices are equal to or below -0.700. It could be given by the fact that a dataset whose attribute values do not contain extreme conditions (such as high or low temperatures), is associated to better clustering evaluation indices, with this algorithm.

For K-means, the best clustering evaluation index for the Hospital de Villamaría station had a value of -1,004, as opposed to the Hospital de Caldas station, which had a value of -1,009. These best results are given for the climate dataset extracted from a region that oscillates between 1790 msnm and 2183 msnm (between warm and temperate climates), using K-means with a value of K = 3, performing normalization with transformation Z and a number of 10 iterations.

Regarding the El Cisne PNNN station, a dataset that comes from high altitude sources, such as 4,812 meters above sea level, the best evaluation index was of -1,051, with a value of K = 2, normalization with Z-transformation, and a number of iterations of the algorithm in 10.

On the other hand, for K-medoids, at El Cisne PNNN station, there are more clustering evaluations with better quality compared to the other two stations. This is determined by taking a value as a reference to make the count. In this case, the indices equal to or below -0.700. It could be given by the fact that a dataset whose attribute values contain extreme conditions (high temperatures or relative humidity of the 100%), such as the El Cisne PNNN station, generate an approximation to better clustering evaluation indices for the clusters in K-medoids.

For K-medoids, the best clustering evaluation index for the Villamaría Hospital station had a value of -1,405, these best results are given for a climate dataset extracted from a region that oscillates around 1,790 masl (warm climate), when using K -medoids a value of K = 5 clusters, normalization with Z-transformation, and number of algorithm iterations in 10.

For the Hospital de Caldas station (altitude of 2,183 masl, temperate climate), the best index had a value of 14,231, using a value of K = 3, without any other characteristic. Regarding the El Cisne PNNN station (4,812 meters, extremely cold weather), the best clustering evaluation had a value of -7,937 and used a value of K = 5, without any other characteristics.

Based on the above and the information seen in the Results section, the cluster evaluation indices are observed with very low values for K-medoids, compared to those obtained in K-means. For two partitioned algorithms used in the experimental framework, the algorithm that presented the best performances and results was K-medoids.

For Linkage-Complete agglomerative clustering, dataset processing that contains the fewest instances and has gone through a normalization process with Range-Transformation performs best on dendrograms, in graphic terms. Even though having fewer instances makes the dendrogram easier to visualize and analyze, normalization makes it possible to shorten similarity distances (Y axis). A performance evaluation index or performance cannot be applied to this algorithm because it is hierarchical clustering and researchers must develop external functionalities in software to provide performance evaluations at a mathematical level [51], and to determine at what point they want to cut the tree to obtain a value of clusters (K), and, from there, analyze the results.

The contribution sought with this work is to provide some basic guidelines, so as not to start from scratch, on certain decisions in the analysis of clusters with meteorological data, as well as to help identify the algorithm and the most important parameters to take into account for the best performance, in accordance with the particular conditions and requirements [52].

CONCLUSIONS AND FUTURE WORK

For future work, it is recommended to use other types of scenarios, treatments, algorithms, and other amounts of clusters to see performance evaluations. It would also be important to know how to evaluate hierarchical agglomerative algorithms to determine the quality of dendrograms to break the subjectivity of each researcher and to apply mathematical measurements.

Furthermore, carrying out scenarios with a K value greater than 5 would allow researchers to investigate what happens with clustering and performance for partitioned algorithms (K-medoids, K-means), both at the machine level and in their performance.

On the other hand, evaluating data on a time scale (per day, per week, etc.) using time series would allow for knowing interesting clustering behaviors, as well as the quality of their clusters within a timeline for different seasons, or times of the year (how the performance would be given for cold seasons or summer seasons). Also, it would be interesting to perform processing under different scenarios that comprise a larger data set (of millions of instances) for K-means, in order to better observe the computational behavior on a larger scale. This will help determine how efficient it is for large datasets, to better detect new patterns or relationships.

Based on the results, it is possible to suggest using other normalization methods, such as ratio and interquartile range transformation, to see how clustering behaves with these analyzes.

It is recommended to use techniques, such as Ordinary Kriging, to handle the large amounts of zeros that a variable contains within a dataset.

REFERENCES

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

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

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

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

[5] J. Chen, M. Song, and L. Xu, "Evaluation of environmental efficiency in China using data envelopment analysis," in *Ecol. Indic.*, Vol. 52, 2015, pp. 577–583. Available: https://doi.org/10.1016/j.ecolind.2014.05.008

[6] L. Chen and G. Jia, "Environmental efficiency analysis of China's regional industry : a data envelopment analysis (DEA) based approach," in *J. Clean. Prod.*, Vol. 142, 2017, pp. 846–853. Available: https://doi.org/10.1016/j.jclepro.2016.01.045

[7] R. Falquina and C. Gallardo, "Development and application of a technique for projecting novel and disappearing climates using cluster analysis," in *Atmos. Res.*, Vol. 197, No. July 2017, pp. 224–231. Available: https://doi.org/10.1016/j.atmosres.2017.06.031

[8] A. M. Kalteh, P. Hjorth, and R. Berndtsson, "Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application," in *Environ. Model. Softw.*, Vol. 23, No.7, 2008, pp. 835–845. Available: https://doi.org/http://dx.doi.org/10.1016/j.envsoft.2007.10.001

[9] S. C. Sheridan and C. C. Lee, "The self-organizing map in synoptic climatological research," *Prog. Phys. Geogr.*, Vol. 35, No. 1, 2011, pp. 109–119. Available: https://doi.org/10.1177/0309133310397582

[10] X. Wang *et al.*, "A stepwise cluster analysis approach for downscaled climate projection - A Canadian case study," *Environ. Model. Softw.*, Vol. 49, 2013, pp. 141–151. Available: https://doi.org/10.1016/j.envsoft.2013.08.006

[11] 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," in

Ecol. Indic., No. September 2016, pp. 0–1, 2017. Available: https://doi.org/10.1016/j.ecolind.2017.06.040

[12] X. Zuo, H. Hua, Z. Dong, and C. Hao, "Environmental Performance Index at the Provincial Level for China 2006–2011," in *Ecol. Indic.*, Vol. 75, 2017, pp. 48–56. Available: https://doi.org/10.1016/j.ecolind.2016.12.016

[13] S. A. Cashman *et al.*, "Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing," in *Environ. Sci. Technol.*, Vol. 50, no. 17, 2016, pp. 9013–9025. Available: https://doi.org/10.1021/acs.est.6b02160

[14] C. Gallo, N. Faccilongo, and P. La Sala, "Clustering analysis of environmental emissions: A study on Kyoto Protocol's impact on member countries," *J. Clean. Prod.*, 2017. Available: https://doi.org/10.1016/j.jclepro.2017.07.194

[15] J. Jiang, B. Ye, D. Xie, and J. Tang, "Provincial-level carbon emission drivers and emission reduction strategies in China: Combining multi-layer LMDI decomposition with hierarchical clustering," in *J. Clean. Prod.*, Vol. 169, 2017, pp. 178–190. Available: https://doi.org/10.1016/j.jclepro.2017.03.189

[16] I. Meghea, M. Mihai, I. Lacatusu, and I. Iosub, "Evaluation of Monitoring of Lead Emissions in Bucharest by Statistical Processing," in *J. Environ. Prot. Ecol.*, Vol. 13, No. 2, ,2012, pp. 746–755. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84864251930&partnerID=MN8TOARS

[17] N. Clay and B. King, "Smallholders uneven capacities to adapt to climate change amid Africa's green revolution: Case study of Rwanda's crop intensification program," in *World Dev.*, Vol. 116, 2019, pp. 1–14. Available: https://doi.org/S0305750X18304285

[18] N. D. Abdul Halim *et al.*, "The long-term assessment of air quality on an island in Malaysia," in *Heliyon*, Vol. 4, No. 12, 2018. Available: https://doi.org/10.1016/j.heliyon.2018.e01054

[19] T. Conradt, C. Gornott, and F. Wechsung, "Extending and improving regionalized winter wheat and silage maize yield regression models for Germany: Enhancing the predictive skill by panel definition through cluster analysis," in *Agric. For. Meteorol.*, Vol. 216, 2016, pp. 68–81. Available: https://doi.org/10.1016/j.agrformet.2015.10.003

[20] S. Farah, D. Whaley, W. Saman, and J. Boland, "Integrating Climate Change into Meteorological Weather Data for Building Energy Simulation," in *Energy Build.*, Vol. 183, 2019, pp. 749–760. Available: https://doi.org/S0378778818323296

[21] T. Soubdhan, M. Abadi, and R. Emilion, "Time dependent classification of solar radiation sequences using best information criterion," in *Energy Procedia*, Vol. 57, 2014, pp. 1309–1316. Available: https://doi.org/10.1016/j.egypro.2014.10.121

[22] S. Khedairia and M. T. Khadir, "Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria," in *Atmos. Res.*, Vol. 113, 2012, pp. 89–101. Available: https://doi.org/10.1016/j.atmosres.2012.05.002

[23] T. Schneider, H. Hampel, P. V. Mosquera, W. Tylmann, and M. Grosjean, "Paleo-ENSO revisited: Ecuadorian Lake Pallcacocha does not reveal a conclusive El Niño signal," in *Glob. Planet. Change*, Vol. 168, No. February, 2018, pp. 54–66. Available: https://doi.org/10.1016/j.gloplacha.2018.06.004

[24] F. Franceschi, M. Cobo, and M. Figueredo, "Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering," in *Atmos. Pollut. Res.*, Vol. 9, No. 5, 2018, pp. 912–922. Available: https://doi.org/10.1016/j.apr.2018.02.006

[25] A. K. Yadav, H. Malik, and S. S. Chandel, "Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India," in *Renew. Sustain. Energy Rev.*, Vol. 52, 2015, pp. 1093–1106. Available: https://doi.org/10.1016/j.rser.2015.07.156

[26] Y. Hao, L. Dong, X. Liao, J. Liang, L. Wang, and B. Wang, "A novel clustering algorithm based on mathematical morphology for wind power generation prediction," in *Renew. Energy*, Vol. 136, 2019, pp. 572–585. Available: https://doi.org/10.1016/j.renene.2019.01.018

[27] S. Han *et al.*, "Quantitative evaluation method for the complementarity of wind–solar–hydro power and optimization of wind–solar ratio," in *Appl. Energy*, Vol. 236, No. December 2018, pp. 973–984, 2019. Available: https://doi.org/10.1016/j.apenergy.2018.12.059

[28] M. André, R. Perez, T. Soubdhan, J. Schlemmer, R. Calif, and S. Monjoly, "Preliminary assessment of two spatio-temporal forecasting technics for hourly satellite-derived irradiance in a complex meteorological context," in *Sol. Energy*, Vol. 177, No. December 2018, pp. 703–712, 2019. Available: https://doi.org/10.1016/j.solener.2018.11.010

[29] P. Lin, Z. Peng, Y. Lai, S. Cheng, Z. Chen, and L. Wu, "Short-term power prediction for photovoltaic power plants using a hybrid improved Kmeans-GRA-Elman model based on multivariate meteorological factors and historical power datasets," in *Energy Convers. Manag.*, Vol. 177, No. July, 2018, pp. 704–717. Available: https://doi.org/10.1016/j.enconman.2018.10.015

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

[31] S. Li, H. Ma, and W. Li, "Typical solar radiation year construction using k-means clustering and discrete-time Markov chain," in *Appl. Energy*, Vol. 205, No. May, 2017, pp. 720–731. Available: https://doi.org/10.1016/j.apenergy.2017.08.067

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

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

[34] L. Pokorná, M. Kučerová, and R. Huth, "Annual cycle of temperature trends in Europe, 1961–2000," in *Glob. Planet. Change*, Vol. 170, No. August, 2018, pp. 146–162. Available: https://doi.org/10.1016/j.gloplacha.2018.08.015

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

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

[37] M. I. Chidean, A. J. Caamaño, J. Ramiro-Bargueño, C. Casanova-Mateo, and S. Salcedo-Sanz, "Spatio-temporal analysis of wind resource in the Iberian Peninsula with data-coupled clustering," in *Renew. Sustain. Energy Rev.*, Vol. 81, No. June, 2018, pp. 2684–2694. Available: https://doi.org/10.1016/j.rser.2017.06.075

[38] Y. Zheng *et al.*, "Assessment of global aridity change," *Ecol. Indic.*, Vol. 75, No. September 2016, pp. 151–165, 2016. Available: https://doi.org/10.1016/j.scitotenv.2015.11.063

[39] J.S. Ramirez, N.D. Duque, N. y J.J. Velez, "Normalización en desempeño de k-means sobre datos climáticos," in *Vínculos*, Vol. 16, 201, 9pp. 57–72. Available: https://doi.org/10.14483/2322939X.15550

[40] D. G. de B. Franco and M. T. A. Steiner, "Clustering of solar energy facilities using a hybrid fuzzy c-means algorithm initialized by metaheuristics," in *J. Clean. Prod.*, Vol. 191, 2018, pp. 445–457. Available: https://doi.org/10.1016/j.jclepro.2018.04.207

[41] J. Hidalgo *et al.*, "Comparison between local climate zones maps derived from administrative datasets and satellite observations," in *Urban Clim.*, Vol. 27, No. November 2017, pp. 64–89, 2019. Available: https://doi.org/10.1016/j.uclim.2018.10.004

[42] C. C. Aggarwal and C. K. Reddy, *DATA Custering Algorithms and Applications*, CRC Press, 2013. Available: https://doi.org/10.1201/9781315373515

[43] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, Pennsylvania: SIAM - Society for Industrial and Applied Mathematics, 2007. Available: https://doi.org/10.1137/1.9780898718348

[44] T. T. Nguyen, A. Kawamura, T. N. Tong, N. Nakagawa, H. Amaguchi, and R. Gilbuena, "Clustering spatio-seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta, Vietnam," in *J. Hydrol.*, Vol. 522, 2015, pp. 661–673. Available: https://doi.org/10.1016/j.jhydrol.2015.01.023

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

[46] A. Lausch, A. Schmidt, and L. Tischendorf, "Data mining and linked open data – New perspectives for data analysis in environmental research," in *Ecol. Modell.*, Vol. 295, 2015, pp. 5–17. Available: https://doi.org/10.1016/j.ecolmodel.2014.09.018

[47] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," in *Procedia Comput. Sci.*, Vol. 85, No. Cms, 2016, pp. 662–668. Available: https://doi.org/10.1016/j.procs.2016.05.251

[48] V. Obradovic, D. Bjelica, D. Petrovic, M. Mihic, and M. Todorovic, "Whether We are Still Immature to Assess the Environmental KPIs!," in *Procedia - Soc. Behav. Sci.*, Vol. 226, No. October 2015, pp. 132–139, 2016. Available: https://doi.org/10.1016/j.sbspro.2016.06.171

[49] K. Pitchayadejanant and P. Nakpathom, "Data mining approach for arranging and clustering the agro-tourism activities in orchard," in *Kasetsart J. Soc. Sci.*, 2017. Available: https://doi.org/10.1016/j.kjss.2017.07.004

[50] S. S. Shaukat, T. A. Rao, and M. A. Khan, "Impact of sample size on principal component analysis ordination of an environmental data set: Effects on Eigenstructure," in *Ekol. Bratislava*, Vol. 35, No. 2, 2016, pp. 173–190. Available: https://doi.org/10.1515/eko-2016-0014

[51] N. Erman and J. Suklan, "Performance of selected agglomerative clustering methods," in *Innov. Issues Approaches Soc. Sci.*, Vol. 8, No. January, 2015. Available: https://doi.org/10.12959/issn.1855-0541.IIASS-2015-no1-art11

[52] J. Ramírez, "Evaluación de algoritmos de aprendizaje de máquina no supervisados sobre datos climáticos". Universidad Nacional de Colombia, 2019. Available: https://repositorio.unal.edu.co/bitstream/handle/unal/75848/1053773873.2019.pdf?isAllowed=y&sequence=3