Region - Water Conservancy, 2023, Volume6, Issue1
DOI: 10.32629/rwc.v6i1.1203
ISSN Online: 2630-4902

Data mining to aquifer vulnerability assessment

Rosa María Valcarce Ortega*, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño

José Antonio Echeverría, Havana Institute of Technology, Cuba
*Corresponding author.
E-mail address: rosy@tesla.cujae.edu.cu

Abstract: The vulnerability map of aquifer pollution is a part of the early warning system to prevent the deterioration of groundwater quality. The weighted index overlay methods are commonly used to map aquifer vulnerability, but they have a series of drawbacks, indicating the need to apply alternative methods that introduce the least number of a priori considerations in the parameters processing and allow a more accurate interpretation of the final results. The purpose of this study is to use data mining techniques for cluster analysis to evaluate the vulnerability of groundwater pollution in the Almondarez Vento Karst Basin in Havana Province, Cuba, and compare the results with those obtained using the RISK method, which is a weighted index overlay method to study karstic aquifers. The variables selected to apply this unsupervised classification technique were: aquifer lithology, topographic slope of the terrain, soil attenuation index to pollutants, fault density per km^2 and presence of direct infiltration zones. The cluster analysis achieved greater spatial discrimination and definition of areas with different degrees of vulnerability, demonstrating its high resolution power.

Key words: aquifer vulnerability; data mining; K-means; Almendares-Vento basin

1 Introduction

Currently, a large amount of data have been generated in various fields of science. Storing and accessing these data in an orderly and fast manner enables the analysis and extraction of useful information, which has led to the birth and development of technologies such as data mining. According to Hernández-Orallo et al. in 2004, data mining is the process of extracting useful and understandable previously unknown knowledge from a large amount of data stored in different formats.

Data mining typically solves four types of tasks: classification, clustering, regression, and association rules. In this study, cluster analysis was used to assess the inherent vulnerability of groundwater pollution in the karst basin, which is crucial for Cuba-- the Almondares Vento Basin.

In Cuba, current environmental legislation recognizes the importance of comprehensive and sustainable management of terrestrial water bodies to achieve the necessary harmony between socio-economic development and environmental protection, and promote the appropriate application of science and technology for this purpose (ANPP, 2017). The vulnerability map of aquifer pollution is part of an early warning system to prevent the deterioration of groundwater quality and is a useful tool for establishing territorial management that is conducive to protecting this resource. Developing methods that can draw these maps more effectively and efficiently is a very important task.

French hydrogeologist J. Margat began studying the inherent pollution vulnerability of groundwater in the late 1960s,

Copyright © 2023 by author(s) and Frontier Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0

based on the fact that the physical environment protects aquifers to some extent from pollutants that may seep from the surface (Margat, 1968). Since then, different methods have been developed to evaluate the inherent vulnerability of aquifers. The most commonly used method is the weighted index overlay method, which assigns points to each parameter based on its range of change and assigns weights to it based on its relative importance in protecting the aquifer. For the method of "n" parameter and "n" weighting factor, the vulnerability index of each unit on the map is calculated based on a linear combination of index and weighting parameters. The vulnerability map is generated by dividing the calculated index values into different ranges and assigning varying degrees of groundwater pollution vulnerability to each range (Olumuyiva and Osakpolor, 2020).

The disadvantage of all these methods is the level of subjectivity of the research team, as separating the range of changes for each parameter, the scores assigned to each range, and the determined weighting factors depend on the experience of the researchers and their prior knowledge of the studied aquifer. Another drawback is that they often use redundant variables to influence the results by providing homogeneous and low resolution vulnerability maps. Another issue is that applying different weighted parameter methods in the same aquifer often yields very different results. On the other hand, these methods are developed for different types of aquifers in various countries and are often imported for different hydrogeological conditions, so modifications are needed. These modifications once again depend on the experience of researchers and often subjective standards.

(Miranda et al., 2015) studied the vulnerability of aquifers in the Coxim River basin in the state of Mato Grosso, South Brazil. They used a weighted exponential superposition method called GOD and EKv. The former takes the depth of groundwater, the type of aquifer (free, enclosed, or monolayer), and the lithology of the aquifer as parameters. The second method considers two parameters: the depth of groundwater level and hydraulic conductivity in the unsaturated zone above the aquifer. The vulnerability maps obtained by the two methods differ greatly. GOD reported three vulnerability categories (low, medium, and high), while EKv identified two vulnerability categories (medium and high) in aquifers.

The DRASTIC and GOD weighted parameter methods were used to study the natural vulnerability of aquifers in the northern part of the state of Sierra, Brazil (Moura et al., 2016). The so-called DRASTIC method uses the following indicators: groundwater level depth, aquifer recharge, aquifer lithology, soil type, terrain slope, unsaturated zone lithology and aquifer hydraulic conductivity. The results of the two methods differ greatly, and DRASTIC provides more robust and accurate results by identifying five vulnerability categories in aquifers. GOD only discriminated against two regions, one with moderate vulnerability accounting for 51% of the area, and the other with high vulnerability accounting for 41% of the total research area.

Fonseca et al. (2019) developed a new method to evaluate the natural pollution vulnerability of free aquifers with intergranular pores. This new weighted index superposition method takes the longitudinal conductivity of the geological layer above the aquifer, the slope of the surface terrain, and the recharge of the aquifer as parameters, studying the Rio Claro aquifer in Sao Paulo, Brazil. The advantage of this method is that it uses a small number of parameters, but the obtained vulnerability map does not provide sufficient resolution because the entire aquifer is classified as highly fragile, which seems unreasonable considering the variability of the parameters used.

In order to minimize the inherent drawbacks of these weighted parameter methods to the greatest extent possible, alternative strategies need to be developed to introduce the least prior considerations when dealing with the parameters to be used, and to allow a more accurate interpretation of the final results. In recent years, there has been a trend towards using data mining techniques to map the vulnerability of aquifers, in order to obtain more objective results, and even in many cases, to have greater spatial discrimination ability.

K-means is a very popular clustering analysis method in data mining. Its purpose is to divide N observations into K groups and assign each observation to the group with the closest average. It is usually suitable for large datasets with many attributes. Hamdan and Emad (2017) analyzed the advantages of this algorithm and pointed out that due to its simplicity, reliability, and high efficiency, it was widely used in different fields and could quickly provide the final results of iterations. In addition, Narang et al. (2016) emphasized the advantages of this clustering technique and suggested combining it with other data mining techniques such as genetic algorithms, neural networks, decision trees, and induction. Other authors have reported on the use of this clustering technique in solving different tasks, such as describing the agronomic behavior of corn varieties and selecting the most pest resistant genotype (Chávez et al., 2010); market segmentation research to improve business management and more effective promotion strategies (Pascal et al., 2015); clustering of hepatitis B virus DNA sequences (Bustamam et al., 2016).

However, despite the effective application of data mining techniques in various scientific fields in recent years, their application in the study of groundwater pollution vulnerability is still limited internationally, with the commonly used weighted index superposition method. The author of this study did not find any publications on the use of data mining techniques to study the vulnerability of Cuban aquifer pollution.

(Javadi and Hashemy, 2016) applied the K-means algorithm to assess the pollution vulnerability of aquifers located in Alborz Province, Iran. They compared the results obtained using the clustering technique and using the DRASTIC method. They concluded that the vulnerability map obtained with the K-means algorithm showed higher accuracy by achieving a Pearson correlation coefficient equal to 0.72 between nitrate concentration in groundwater and the defined vulnerability category.

Vulnerability to salinization of coastal aquifers caused by saline intrusion in Mazandaran province, northern Iran (Motevalli et al., 2019) was studied by applying generalized additive model, generalized linear model and support vector machines as data mining techniques. The map obtained by applying these techniques accurately revealed the zones of low, moderate, high and very high vulnerability of the studied coastal aquifers to salt intrusion.

The purpose of this study is to use data mining clustering analysis technology to evaluate the vulnerability of groundwater pollution in the Almondares Vento karst basin, and compare the results with the weighted index superposition RISK method. Through this approach, the aim is to evaluate the problem-solving ability of cluster analysis in solving these tasks

2 Methods

47% of the total water consumption in Havana City comes from the Almendares Vento Basin (Herrera et al., 2004), indicating the importance of conducting research that helps protect its groundwater from potential pollution sources. The northern end of the basin is bordered by the San Francisco de Paula Mountains and the Santa Maria Del Rosario Romeo Mountains, the eastern end is bordered by the Haruko Staircase, the southern end is bordered by the Behukar Maduga Colosseum Heights, the western end is bordered by the mouth of the Almondares River, as well as an ocean terrace appearing on the beach and the northern coastal border of Revolution Square City (Fig. 1). The estimated exploitable resources of this hydrogeological basin are 287 million cubic meters per year. The hydrological network is mainly composed of the Almondares River and its tributaries, which pass through its central area. There is a significant development of urban and industrial activities within the watershed. The chemical, pharmaceutical, food and steel industries, as well as agricultural and livestock activities and multiple services to the population stand out with several hospitals, educational centers, parks, hotels, cultural centers and an important development of transportation including

several highways and the country's main international airport, which causes an increase in the potential pollution of the watershed and the need for its protection.

(Valcarce et al. 2020) evaluated the natural vulnerability of groundwater contamination in the Almendares-Vento basin by applying the RISK method, a weighted rank superposition method that takes its name from the acronym formed by the name of the variables it uses: aquifer rock (R), aquifer infiltration conditions (I), soil properties (S), and development of the karst network or karstification (K). The behavior of these parameters defines the greater or lesser protection of the physical environment from aquifer contamination (Dörfliger et al., 2004).

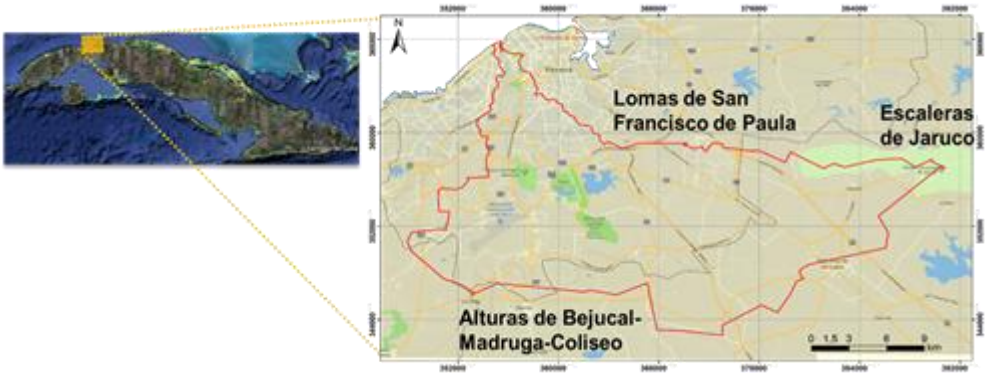

Fig. 1. The geographical location of the Almendares Vento Basin

The aquifer rock parameter (R) reflects the nature and degree of fracture of the tectonics, which has a great impact on the type of underground circulation and therefore on the transfer speed of pollutants in the aquifer. The assessment was carried out according to the 1:100,000 scale geological map of the Republic of Cuba (PGI, 2016). Parameter 1 takes into account the slope of the terrain and topography, as the larger the slope of the terrain, the greater the acceleration of surface runoff and the smaller the permeability to the aquifer. It also takes into account the existence of karst forms that directly connect surface water flow and groundwater flow, and evaluates them based on information from the 10×10 digital elevation model (Geocuba, 2010). The soil parameters represent the first protective barrier of the aquifer. Its thickness, texture (pebbles, matrix, etc.), and composition (clay, silt) affect the greater or lesser vulnerability of the aquifer to pollution. The characteristic of this standard is based on a 1:25,000 scale soil map (Soil Research Institute, 1990). Finally, parameter K describes the impact of the underground karst network, which facilitates the transport of pollutants in aquifers. This standard was also evaluated based on a 1:100,000 scale geological map of the Republic of Cuba (PGI, 2016).

Each of these standards is evaluated and divided into different ranges, with scores ranging from 0 to 4 (from less to more) for each range. We also defined a weighting factor and ultimately calculated the risk vulnerability index based on the following equation:

$$RISK = 0.15R + 0.41I + 0.25S + 0.20K \quad (1)$$

The risk index is divided into several ranges, each of which is assigned a vulnerability category. This method addresses the limitations of the above analysis, namely the subjectivity level of the research team in dividing the range of each parameter, assigning scores to each range, and setting weight factors.

Clustering analysis, one of the techniques of data mining, has shown that it can provide more objective results in the assessment of aquifer vulnerability than weighted rank overlay methods, provided that the variables used in the analysis are sufficiently effective. For this reason, one of the clustering techniques, the K-means algorithm, was applied to assess vulnerability of groundwater contamination in this basin.

Clustering analysis is an unsupervised pattern recognition technique that allows groups to be obtained from a large amount of data, so that the elements of each group are very similar to each other while being very different from the elements of other groups. There are different clustering or clustering algorithms. In this study, the K-means algorithm was used to assign each object to the nearest group, which is achieved by calculating the similarity measure between the centroids of the object and each cluster.

There are different similarity measures, which are more or less effective considering the quantity, qualitative or binary properties of variables representing the object to be classified, as well as the lack of data and the level of correlation between different variables. The appropriate selection of similarity measures plays an important role, as it largely defines the maximum reliability of the final result. In this study, the proposed Euclidean distance coefficient was used as a similarity measure when the variables used were quantitatively represented and their statistical correlation was low (Alfonso, 1989; Núñez-Colín y Escobedo-López, 2011).

The number of groups and their centroids were initially randomly calculated, and after the object was first assigned to each group, the centroids were recalculated as the average variable of the points assigned to them. Once the centroid of each cluster is updated, the objects will be reassigned to the nearest cluster. This process is repeated until convergence is reached, that is, until the allocation of points remains unchanged, or until the predetermined number of iterations is reached. This final result represents the adjustment that maximizes the distance between the different groups and minimizes the intra-group distance. The main advantage of the method is its simplicity and speed, but it is an algorithm that is significantly sensitive to the centroids that are initially selected randomly. This effect can be reduced by increasing the number of iterations of the procedure (Hernández-Orallo, et al., 2004).

This algorithm is more efficient when the variables used are not redundant, and is very sensitive to the fact that variables have different ranges of change. Therefore, it is recommended to standardize them between 0 and 1 by subtracting their minimum value from each variable and dividing it by its range.

It is important to emphasize that the algorithm will always group objects. The knowledge of researchers will make it possible to determine which groups are important and which groups are not. The software used is free software WEKA developed by Waikato University, so its name is Waikato Knowledge Analysis Environment (Martínez, 2018).

The variables to be used must be selected based on the objectives of the clustering analysis to be conducted. In this study, the aim was to identify areas with varying degrees of vulnerability in aquifers. The most common variables in traditional methods were used to evaluate the vulnerability of karst aquifers: aquifer lithology (Lit), terrain slope (PenTop), soil attenuation index (IAS) for pollutants, fault density per square kilometer (Df), and the presence of direct infiltration zones (Zi).

The lithology of an aquifer affects the transfer rate of pollutants in the aquifer. The K-means algorithm requires all variables to be quantitative. Due to the qualitative nature of lithology, values ranging from 1 to 4 are assigned to different types of rocks, indicating that the higher the value, the greater the vulnerability. It is determined that 1 is carbonate rock with high clay and marl, 2 is limestone with low clay material content and interlayer of marl, 3 is limestone with low fracture and massive dolomite, and 4 is limestone and massive dolomite with high fracture strength.

The topographic slope of the terrain is calculated based on the Digital Elevation Model (MDE). As mentioned above, on larger terrain slopes, due to the dominance of surface runoff, the infiltration of pollutants into the aquifer is relatively small.

The soil attenuation index is a parameter created based on the sum of three soil characteristics (thickness, organic matter content, and viscosity). The increase of these characteristics enhances the ability of soil to delay the vertical migration of potential pollutants deposited on the surface, thereby reducing its vulnerability to groundwater pollution.

The core density estimation tool of QGIS software is used to calculate the fault density per square kilometer. The higher the density of faults, the higher the porosity of cracks, and the higher the permeability, which helps to penetrate potential pollutants.

The direct infiltration zone is determined by replacing the Fill Sinks grid obtained through hydrological expansion using the Terrain Analyst tool of the QGIS software with a digital elevation model to identify areas with negative terrain forms that may be related to karst landscape performance such as Dolinas and Uwalas. (Pardo-Iguzquiza et al., 2014). Then, overlap the surface drainage network to detect direct infiltration areas, which correspond to terrain depressions where surface water flow loses continuity. These areas are highly susceptible to groundwater pollution because they allow direct connection between the surface and underground karst networks, so any pollutant can immediately communicate with the aquifer.

Table 1 summarizes the data sources and their variation ranges for extracting selected variables from them. All maps of these variables were drawn in Raster format at a scale of 1:100,000. In order to establish a database, the values of each attribute were collected every 25 meters in the watershed area, with a total of 719,131 instances.

Table 1. The data source and range of variation for each selected variable

Variable	Rango de variación	Fuente de datos
Litología (Lit)	1–4	Mapa Geológico de la República de Cuba a escala 1:100 000 (IGP, 2016)
Pendiente Topográfica (PenTop)	0% – 118%	Modelo digital de elevación 10X10 (GEOCUBA, 2010)
Índice de atenuación del suelo (Ias)	41 –143	Mapa de Suelos a escala 1:25 000 Instituto de Suelos (1990).
Densidad de fallas por km^2 (Df)	0–2	Mapa Geológico de la República de Cuba a escala 1:100 000 (IGP, 2016)
Zonas de infiltración directa (Zi)	0–1	Modelo digital de elevación 10X10 (GEOCUBA, 2010)

In order to link the degree of pollution vulnerability corresponding to each cluster, the concept of ideal points was used (Vías et al., 2003) as a reference for the most favorable scenario, which is to strengthen the protection of aquifers. This point is located in rocks with low permeability, with a large terrain slope, high soil attenuation rate, low fault density, and no direct infiltration zone. For the aquifer being studied, the coordinates of ideal points in n-dimensional Euclidean space are: Lit=1; Pentop=118%; Ias=143; Df=0, Zi=0. The distance from each group to an ideal point in n-dimensional Euclidean space is calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - p_j)^2} \quad (2)$$

Among them:

D_{ij}: Euclidean distance from the centroid of cluster i to the ideal point Pj.

X_{ik}: The centroid of cluster i, that is, the point whose coordinates are the average of n variables that make up the points of cluster i.

P_j: The coordinates of the ideal point.

3 Results and discussion

Then, the RISK method and K-means algorithm were used to analyze the results of groundwater pollution vulnerability assessment in the Almondares Vento Basin.

Fig. 3 shows a map of the watershed R, I, S, and K standards at a scale of 1:100,000. According to the criteria of aquifer rock (R), aquifer permeability (I), and karst network development (K), it can be seen that the basin has a high and very high degree of vulnerability. Soil parameters have greater protection capabilities, reflecting the advantages of medium to high vulnerability.

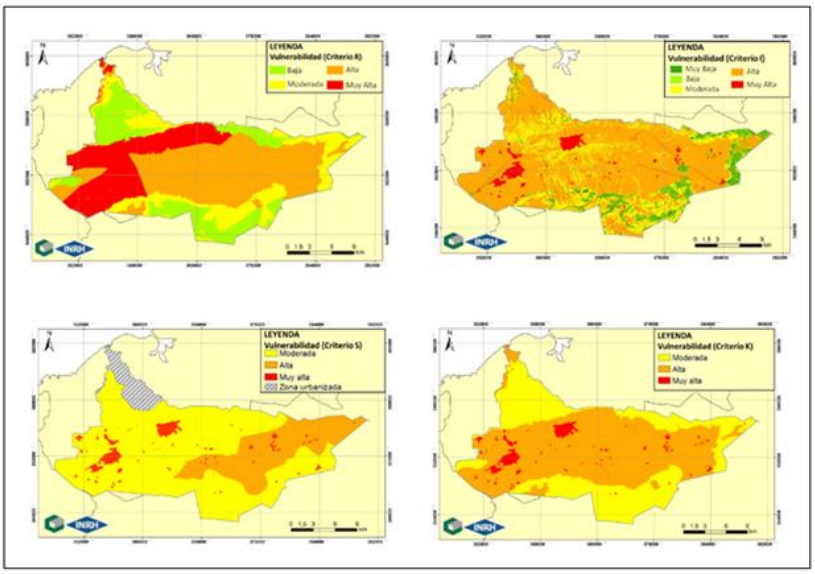

Fig. 3. R, I, S, and K standard maps of the Almondares Vento Basin. Scale 1:100,000

Once the risk index is calculated, it is divided into different ranges and classified as shown in Table 2.

Table 2. Vulnerability index classification based on RISK approach

División en rangos	Puntuación	Clase de vulnerabilidad
3,2 - 4	4	Muy alta
2,4 – 3,19	3	Alta
1,6 – 2,39	2	Moderada

(Modified by Dörfliger et al., 2004)

Fig. 4 shows the vulnerability map generated by the spatial representation of the RISK index.

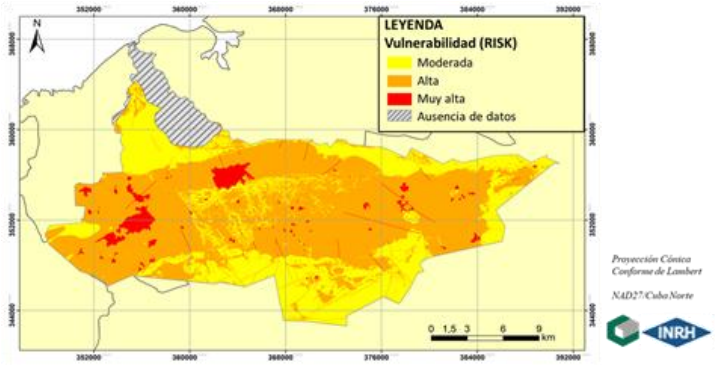

Fig. 4. The inherent vulnerability map of groundwater pollution in the Almondares Vento Basin obtained through RISK analysis

The groundwater in the central basin is highly vulnerable to the vertical migration of potential pollutants deposited on its surface. In 62.4% of the region, the development level of karst carbonate rocks is high, and compared to other areas where clay rocks dominate, the fragility of karst carbonate rocks is higher. 4% of the areas are classified as highly vulnerable, consistent with direct infiltration areas. The remaining 33.6% belongs to moderately fragile areas, mainly composed of rocks with high clay content and highland slopes.

To apply the K-means algorithm, low correlation is required between the selected variables. Table 3 shows this premise.

Table 3. Select the linear correlation matrix between variables using the K-means method

	Lit	PenTop	las	Df	Dzi
Lit	1	-0,34	0,27	-0,41	0,2
PenTop	-0,34	1	-0.15	0,23	-0,09
las	0,27	-0,15	1	-0.19	0,11
Df	-0,41	0,23	-0,19	1	-0,1
Zi	0,2	-0,09	0,11	-0,1	1

The centroid, the distance from each centroid to the ideal point, and the degree of vulnerability assigned to each group are shown in Table 4.

Table 4. The centroid of each cluster, the distance to the ideal point, and the degree of fragility of each cluster

		CENTROIDES					Distancia al punto ideal	Grado de vulnerabilidad
CLÚSTER	TOTAL DE INTANCIAS	Lit	PenTop	las	Df	Zi		
1	27 146	3,7	2,0	41	0,2	1	154	Muy alta
2	203 914	3,5	3,0	89	0,2	0	127	Alta
3	138 044	2,9	4,0	91	0,3	0	125	Moderada
4	116 703	3,4	3,4	117	0,3	0	118	Baja
5	233 324	1,5	9,0	143	0,6	0	109	Muy baja

By analyzing the distance from each centroid to the ideal point, the vulnerability level of groundwater pollution related to each cluster can be directly explained.

The group identified as highly vulnerable is the group furthest from the ideal point, characterized by direct infiltration of potential pollutants into aquifers, small terrain slopes, low soil attenuation index, and rocks with high karst development. The group classified as very low vulnerability is the group without direct infiltration. The attenuation rate of soil is the

highest, the slope is the largest, the lithology is characterized by the predominance of very clayey carbonate rocks and marls, and its centroid is closer to the ideal point.

Fig. 5 shows the vulnerability map obtained using the cluster analysis application. In the central region of the basin, according to the RISK method, the vulnerability is high, and cluster analysis has successfully identified three categories. To the west, it is classified as highly fragile, corresponding to the existence of tectonics, where the surface karst shows a high degree of development, the terrain slope is small, and the soil attenuation rate is low. A moderately fragile area related to the Columbus Formation was identified eastward, characterized by the presence of more clay rocks and less developed outer karst forms. In the periphery of the central region, due to the presence of thicker soil with higher organic matter and clay content, i.e. stronger protective capacity, as well as the presence of more clay lithology and larger terrain slopes, the vulnerability is low and very low. The areas that are highly susceptible to groundwater pollution are consistent with the maps obtained by the RISK method and K-means algorithm, and are areas where potential pollutants directly penetrate the surface.

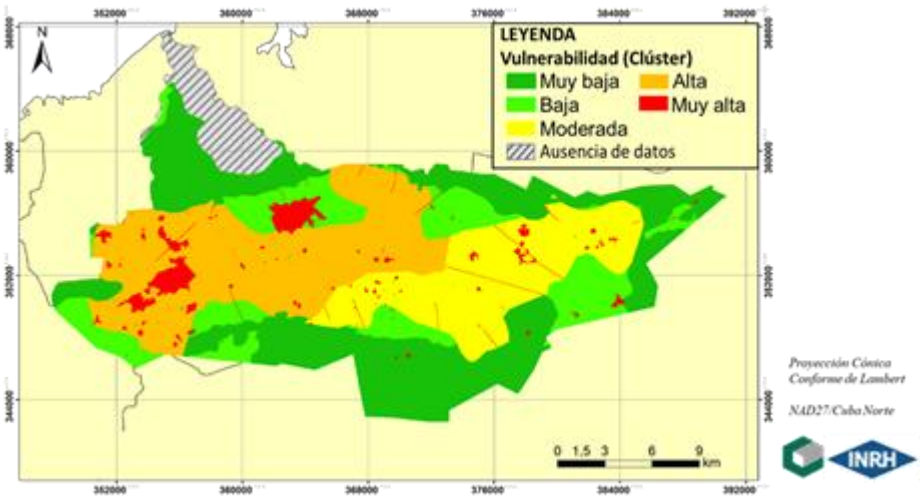

Fig. 5 The K-means algorithm was applied to obtain the inherent vulnerability map of groundwater pollution in the Almendares Vento Basin

In order to integrate the information provided by the RISK method and K-means algorithm, as shown in Fig. 6, it is clear that by providing a model with five types of groundwater degradation sensitivity, using unsupervised classification can improve the resolution of drawing the natural vulnerability of aquifer pollution.

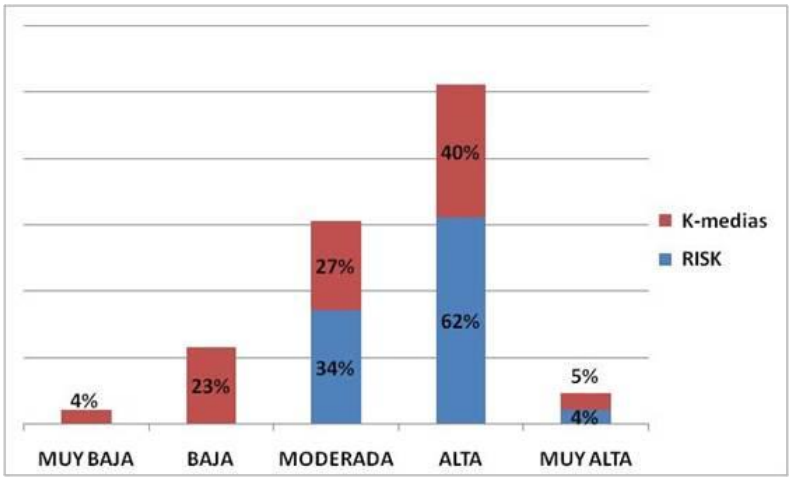

Fig. 6. Compare the inherent vulnerability level of groundwater pollution in the Almondares Vento Basin using the RISK method and K-means algorithm

From the economic and environmental perspective, the contribution of such research is valuable, as the supply and quality of water are the main challenges faced by any country, and Cuba is no exception. In particular, the Almondarez Vento Basin is crucial for Cuba's economic and social development, and the pollution caused by irresponsible human actions may be irreversible or require significant resources and time to take effective remedial action. It is crucial to guide correct policies in the territorial management of the basin and ensure the protection of its groundwater resources. Therefore, assessing the natural vulnerability of aquifers is the first step. The cost of these studies is not high, as the information required to successfully carry out these tasks can be found in the archives and databases of companies related to geological activities in the country, and there are human resources, technical and professional capabilities to develop these tasks.

4 Conclusion

A classification model was obtained, which included five clusters related to varying degrees of groundwater pollution sensitivity. The model successfully distinguished very high, high, medium, low, and very low vulnerability areas in the Almondares Vento Karst Basin. This result shows a higher resolution than the RISK method that only identified three types of vulnerabilities. Research has shown that this unsupervised statistical classification technique does not have the drawbacks of the weighted index superposition method commonly used to evaluate the vulnerability of aquifers, and allows a more objective and accurate mapping of the vulnerability of aquifers to pollution.

The research conducted contributes to research on the protection of aquifers in Cuba. Due to its efficiency and resolution, it is recommended to use the K-means algorithm to evaluate other important watersheds on the national territory.

Conflicts of interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Asamblea Nacional del Poder Popular, ANPP. Ley No. 124 DE LAS AGUAS TERRESTRES. [En línea]. Gaceta Oficial No. 51 Extraordinaria. 2017. [Consultado el: 11 de enero de 2018]. p. 985-1047 Disponible en Disponible en http://www.gacetaoficial.cu/

[2] Alfonso, J. R. Estadísticas en las Ciencias Geológicas, Tomo 2. La Habana, ISPJAE, 1989. 308 p.

[3] Bustamam, A.; Tasman, H.; Yuniarti, N.; Mursidah, I. Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV). [En línea] International Symposium on Current Progress in Mathematics and Sciences 2016. AIP Conference Proceedings 1862, 030134. [Consultado el: 22 de noviembre de 2020]. p. 1-8. Disponible en: Disponible en: https://doi.org/10.1063/1.4991238

[4] Chávez, D.; Miranda, I.; Varela, M.; Fernández, L. Utilización del análisis de cluster con variables mixtas en la selección de genotipos de maíz (Zea mays). Revista Investigación Operacional, 2010, 30 (3): p. 209-216.

[5] Dörfliger, N.; Jauffret, D.; ET Loubier, S. Cartographie de la vulnerabilité des aquiferes karstiques en Franche-Compté. Francia. [En línea]. BRGM RP-53576-FR, 2004 [Consultado el: 29 de septiembre de 2018]. p. 547-571. Disponible en: Disponible en: https://www.google.com.cu/search?q=Cartographie+de+la+vulnerabilit%C3%A9+des+aquiferes+karstiques+en+Franche+%E2%80%93+Comt%C3%A9&spell=1&sa=X&ved=2ahUKEwj176iprLqAhVHUt8KHUsLBpQQkeECKAB6BAgLECo&biw=1999&bih=979

[6] Hamdam, H., Emad, L. K-means clustering algorithm applications in data mining and pattern recognition. International Journal of Science and Research, 2017, 6(8): p. 1577-1584.

[7] Hernández-Orallo, J.; Ramirez Quintana, M. J.; Ferri Ramírez, C. Introducción a la Minería de datos. Pearson Educación, 2004. 680 p.

[8] Herrera J.; Fonseca, C.; Goicochea, D. Perspectivas del medio ambiente urbano GEO La Habana. La Habana, SI-MAR S.A., 2004. 190 p.

[9] Instituto de Geología Y Paleontología, IGP. Mapa Geológico de la República de Cuba a escala 1:100 000. La Habana: Servicio Geológico de Cuba, 2016.

[10] Instituto de Suelos. Mapa de los suelos de Cuba a escala 1:25 000. La Habana: Ministerio de la Agricultura. 1990.

[11] Javadi, S.; Hashemy, S. M.; Mohammadi, K.; Howard, K. W.; Neshat, A. Classification of aquifer vulnerability using K-means cluster analysis. Journal of Hydrology, 2017, (549): p. 27-37.

[12] Margat, J. Vulnerabilité des nappes d'eau souterraine a la pollution. Francia, BRGM, 1968. 68 p.

[13] Martínez, A. F. Aplicación de técnicas de minería de datos con software Weka. [En línea]. II Semana Doctoral Formación de la Sociedad del Conocimiento, Universidad de Salamanca, 2018. [Consultado el: 29 de septiembre de 2018]. 17 p. Disponible en: Disponible en: http://dx.doi.org/10.1007/s10115-003-0128-3

[14] Motevalli, A.; Reza, H.; Hashemi, H.; Gholami, V. Assessing the vulnerability of groundwater to salinization using GIS - based data mining techniques in a coastal aquifer. [En línea]. Spatial Modeling in GIS and R for Earth and Environmental Sciences, 2019. [Consultado el: 11 de enero de 2019] p. 547-571. Disponible en: Disponible en: https://doi.org/10.1016.B978-0-12-815226-3.00025-9

[15] Miranda, C.; Mioto, C.; Lastoria, G.; Garcïa, S.; Paranhos, F. Uso de Sistemas de Informação Geográfica (SIG) na modelagem da vulnerabilidade de aquífero livre: comparação entre os métodos GOD e Ekv na bacia do rio Coxim, São Gabriel do Oeste, MS, Brasil. Geociencias, 2015, 34(2): p. 312-322.

[16] Moura, P.; Sabadia, J.A.; Cavalcante, I. Mapeamento de vulnerabilidade dos aquíferos Dunas, Barreiras e Fissural na porção norte do complexo industrial e portuário do Pecém, estado do Ceará. Geociencias, 2016, 35(1): p. 77-89.

[17] Narang, B., Verma, P., Kochar, P. Application based, advantageous K-means Clustering Algorithm in Data Mining - A Review. International Journal of Latest Trends in Engineering and Technology, 2016, 7(2): p. 121-126.

[18] Núñez-Colín, C.; Escobedo-López, D. Uso correcto del análisis clúster en la caracterización de germoplasma vegetal. Agronomía Mesoamericana, 2011, 22(2): p. 415-427.

[19] Olumuyiwa, F.; Osakpolor, O. Groundwater vulnerability mapping and quality assessment around coastal environment of Ilaje Local government area, southwestern Nigeria. International JournalofEarthSciencesKnowledge and Applications, 2020, 2(2): p. 74-91.

[20] Pardo-Iguzquiza, E., Durán, J., Luque-Espinar, J., martos-ROSILLO, S. Análisis del relieve kárstico mediante el modelo digital de elevaciones. Aplicación a la Sierra de las Nieves (Provincia de Málaga). Boletín Geológico y Minero, 2014, 125(3): p. 381-389.

[21] Pascal, CH.; Ozuomba, S.; Kalu, C. Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. International Journal of Advanced Research in Artificial Intelligence, 2015, 4(10): p. 40-44.

[22] Valcarce, R. M.; Vega, M.; Rodríguez, W.; Suárez, O. Vulnerabilidad intrínseca de las aguas subterráneas en la cuenca Almendares-Vento. Ingeniería Hidráulica y Ambiental, 2020, 41(2): p. 33-47.

[23] Vías, J. M.; Perles, M. J.; Andreo, B. Aplicación de un análisis clúster para la evaluación de la vulnerabilidad a la contaminación de los acuíferos. Revista Internacional de Ciencia y Tecnología de la Información Geográfica, 2003, (3): p. 199-215.