Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

Tipo de artículo: Artículo original

Temática: Reconocimiento de patrones

Recibido: 21/12/2021 | Aceptado: 06/03/2021

Aplicación de la minería de datos a la evaluación de la vulnerabilidad de acuíferos

Data mining to aquifer vulnerability assessment

Rosa María Valcarce Ortega[1*] https://orcid.org/0000-0001-9981-6832

Oscar Suárez González[2] https://orcid.org/0000-0003-1617-5262

Willy Rodríguez Miranda[3] https://orcid.org/0000-0003-2938-6472

Marina Vega Carreño[4] https://orcid.org/0000-0001-6745-5282

[1]Ingeniera Geofísica, Profesora Titular. Universidad Tecnológica de La Habana "José Antonio Echeverría", CUJAE. Calle 114 # 11901 e/ Rotonda y Ciclovía, Marianao, La Habana, Cuba. rosy@tesla.cujae.edu.cu

[2]Ingeniero Geofísico, Universidad Tecnológica de La Habana "José Antonio Echeverría", CUJAE. Calle 114 # 11901 e/ Rotonda y Ciclovía, Marianao, La Habana, Cuba. oasuarezg@gmail.com

[3]Ingeniero Geofísico, Profesor Titular. Universidad Tecnológica de La Habana "José Antonio Echeverría", CUJAE, Calle 114 # 11901 e/ Rotonda y Ciclovía, Marianao, La Habana, Cuba. willy@civil.cujae.edu.cu

[4]Ingeniera Geofísica, Profesora Titular, Profesora Consultante. Universidad Tecnológica de La Habana "José Antonio Echeverría", CUJAE. Calle 114 # 11901 e/ Rotonda y Ciclovía, Marianao, La Habana, Cuba. mvega@civil.cujae.edu.cu

[*] Autor para correspondencia. (rosy@tesla.cujae.edu.cu)

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

RESUMEN

Los mapas de vulnerabilidad a la contaminación de los acuíferos forman parte de un sistema de alerta temprana para prevenir el deterioro de la calidad de las aguas subterráneas. Los métodos de superposición de índices ponderados son comúnmente empleados para realizar la cartografía de la vulnerabilidad de los acuíferos, pero presentan un conjunto de desventajas que indican la necesidad de aplicar métodos alternativos que introduzcan el menor número de consideraciones *a priori* en el procesamiento de los parámetros a utilizar y permitan una interpretación más precisa de los resultados finales. El objetivo de esta investigación fue evaluar la vulnerabilidad a la contaminación de las aguas subterráneas de la cuenca kárstica Almendares–Vento en la provincia La Habana, Cuba, al emplear la técnica de minería de datos análisis de agrupamiento, y comparar los resultados con los obtenidos al aplicar el método RISK, que es un método de superposición de índices ponderados para el estudio de acuíferos kársticos. Las variables seleccionadas para aplicar esta técnica de clasificación no supervisada fueron: litología del acuífero, pendiente topográfica del terreno, índice de atenuación del suelo a los contaminantes, densidad de fallas por km^2 y presencia de zonas de infiltración directa. El análisis de agrupamiento logró una mejor discriminación espacial y definición de zonas con diferentes grados de vulnerabilidad, demostrando su mayor poder resolutivo.

Palabras clave: vulnerabilidad de acuíferos; minería de datos; algoritmo K-medias; cuenca Almendares–Vento.

ABSTRACT

The maps of vulnerability to contamination of the aquifers are part of an early warning system to avoid the deterioration groundwater quality. Weighted index overlay methods are commonly used to map aquifer vulnerability. These methods have disadvantages that indicate the need to apply alternative methods that introduce the least number of *a priori* considerations in the parameters processing and allow a more precise interpretation of the final results. The objective of this research was to evaluate the vulnerability to contamination of groundwater in the Almendares-Vento karstic basin, Havana, Cuba, by using the data mining technique, and to compare the results obtained by applying the RISK methods, which is a weighted index overlay method to study karstic aquifers. The variables selected to apply this unsupervised classification technique was: 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.

Keywords: aquifer vulnerability; data mining; K-means; Almendares–Vento basin.

Introducción

En la actualidad se generan enormes cantidades de datos en todos los campos de la ciencia. Almacenar y acceder a estos datos de forma ordenada y rápida, con la capacidad de analizar y extraer de ellos información útil, ha motivado el nacimiento y desarrollo de técnicas conocidas como minería de datos. Según (Hernández–Orallo *et al.* 2004) la minería de datos es “el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos”.

Usualmente la minería de datos resuelve cuatro clases de tareas: clasificación, agrupamiento (*clustering*), regresión y reglas de asociación. En esta investigación se emplea el agrupamiento

para evaluar la vulnerabilidad intrínseca a la contaminación de las aguas subterráneas en una cuenca kárstica de gran importancia para Cuba; la cuenca Almendares–Vento.

En Cuba la legislación ambiental vigente reconoce la importancia de la gestión integrada y sostenible de las aguas terrestres para lograr la necesaria armonía entre el desarrollo socio económico y la protección del medio ambiente, y se promueve la adecuada aplicación de la ciencia y la tecnología con este propósito (ANPP, 2017). Los mapas de vulnerabilidad a la contaminación de los acuíferos forman parte de un sistema de alerta temprana para prevenir el deterioro de la calidad de las aguas subterráneas, son herramientas útiles para establecer un ordenamiento territorial que favorezca la protección de este recurso. El desarrollo de métodos que permitan elaborar estos mapas de manera cada vez más eficaz y eficiente constituye tarea de gran importancia.

Los estudios de vulnerabilidad intrínseca del agua subterránea a la contaminación comienzan a desarrollarse a finales de la década del 60 del pasado siglo por el hidrogeólogo francés J. Margat, basándose en el hecho de que, en cierta medida, el medio físico protege al acuífero de contaminantes que pueden infiltrarse desde la superficie (Margat, 1968). A partir de ese momento se han desarrollado diferentes métodos para evaluar la vulnerabilidad intrínseca de los acuíferos, los más comúnmente empleados son los métodos de superposición de índices ponderados, que consisten en asignar puntos a cada parámetro según su rango de variación, y además, asignarle un factor de ponderación según su importancia relativa en la protección del acuífero. Para un método de “n” parámetros y “n” factores de ponderación, el índice de vulnerabilidad para cada celda del mapa se calcula a partir de la combinación lineal de los parámetros indexados y ponderados. El mapa de vulnerabilidad se genera al separar en rangos los valores del índice calculado y asignar a cada rango diferente grado de vulnerabilidad a la contaminación de las aguas subterráneas (Olumuyiwa y Osakpolor, 2020).

Todos estos métodos presentan como desventaja el nivel de subjetividad que le imprime el equipo de investigadores, pues la separación en rangos de variación de cada parámetro, la puntuación asignada a cada rango y el factor de ponderación establecido, dependen de la experiencia de los investigadores y del conocimiento *a priori* que posean del acuífero bajo

estudio. Otra desventaja es que frecuentemente emplean variables redundantes afectando los resultados al brindar mapas de vulnerabilidad homogéneos y poco resolutivos. Otro problema radica en el hecho de que la aplicación en un mismo acuífero, de diferentes métodos paramétricos ponderados generalmente reporta resultados muy disímiles. Por otra parte, estos métodos han sido desarrollados para diferentes tipos de acuíferos en diversos países, y con frecuencia se importan para ser aplicados en condiciones hidrogeológicas diferentes, por lo que se requiere hacer modificaciones que otra vez dependen de la experiencia de los investigadores y de criterios muchas veces subjetivos.

(Miranda *et al*. 2015) estudiaron la vulnerabilidad de los acuíferos de la cuenca del río Coxim en Mato Grosso del Sur, Brasil. Emplearon los métodos de superposición de índices ponderados denominados GOD y EKv. El primero emplea como parámetros la profundidad del agua subterránea, el tipo de acuífero (libre, confinado o semiconfinado) y la litología del acuífero. El segundo método considera dos parámetros; profundidad del nivel freático y conductividad hidráulica de la zona no saturada que sobreyace al acuífero. Los mapas de vulnerabilidad obtenidos con ambos métodos difieren en gran medida, GOD reportó tres clases de vulnerabilidad (baja, media y alta) mientras que EKv identificó dos clases de vulnerabilidad en el acuífero (media y alta).

El estudio de la vulnerabilidad natural de los acuíferos al norte del estado de Ceará, en Brasil, fue abordado por los métodos paramétricos ponderados DRASTIC y GOD (Moura *et al*., 2016). El método denominado DRASTIC emplea los índices: profundidad del nivel freático, recarga del acuífero, litología del acuífero, tipo de suelo, pendiente topográfica del terreno, litología de la zona no saturada y conductividad hidráulica del acuífero. Los resultados por ambos métodos difieren en gran medida, DRASTIC ofreció resultados más robustos y precisos al identificar cinco clases de vulnerabilidad en el acuífero; GOD solo discriminó dos zonas, una de vulnerabilidad moderada que alcanza el 51% del área, y otra de alta vulnerabilidad que representa el 41% del área total estudiada.

Fonseca *et al*. (2019) desarrollaron un nuevo método para evaluar la vulnerabilidad natural a la contaminación de acuíferos libres con porosidad intergranular. Este nuevo método de

superposición de índices ponderados emplea como parámetros: la conductividad eléctrica longitudinal de las capas geológicas que sobreyacen al acuífero, la pendiente topográfica de la superficie del terreno y la recarga del acuífero. Fue estudiado el acuífero Río Claro, en Sao Paulo, Brasil. El método tiene la ventaja de emplear un reducido número de parámetros, no obstante, el mapa de vulnerabilidad obtenido no ofreció adecuado poder resolutivo porque todo el acuífero fue clasificado de alta vulnerabilidad lo que no parece razonable atendiendo a la variabilidad que presentan los parámetros empleados.

Para minimizar las desventajas propias de estos métodos paramétricos ponderados se requiere desarrollar estrategias alternativas que introduzcan el menor número de consideraciones *a priori* en el procesamiento de los parámetros a utilizar y permitan una interpretación más precisa de los resultados finales. En los últimos años existe la tendencia a emplear técnicas de minería de datos para cartografiar la vulnerabilidad de los acuíferos, obteniéndose de esta forma resultados más objetivos e incluso, en muchas ocasiones, con una mayor capacidad de discriminación espacial.

K-medias es un método muy popular para el análisis de conglomerados en la minería de datos. Tiene como objetivo dividir n observaciones en k grupos, asignando cada observación al grupo con la media más cercana. Generalmente se aplica a grandes conjuntos de datos con muchos atributos. (Hamdan y Emad, 2017) analizan las ventajas de este algoritmo señalando: su amplia aplicación en diferentes campos, su capacidad de brindar el resultado final de sus iteraciones de manera rápida debido a la simplicidad del algoritmo que emplea, su alta confiabilidad y eficiencia. También (Narang *et al*., 2016) destacan las ventajas de esta técnica de agrupamiento y sugiere su integración con otras técnicas de la minería de datos, como los algoritmos genéticos, redes neuronales, árboles de decisión y métodos de inducción. Otros autores reportan la utilización de esta técnica de agrupamiento en la solución de diferentes tareas, por ejemplo, para caracterizar el comportamiento agronómico de variedades de maíz y seleccionar los genotipos más resistentes a las plagas (Chávez *et al*., 2010); estudios de segmentación del mercado para una mejor gestión de los negocios y más eficientes estrategias de promoción (Pascal et al., 2015); agrupamiento de las secuencias de ADN del virus de la hepatitis B (Bustamam *et al*., 2016).

No obstante, aunque las técnicas de minería de datos han sido empleadas en los últimos años con demostrada efectividad en diversos campos de la ciencia, todavía es escasa su aplicación a nivel internacional en estudios de vulnerabilidad de las aguas subterráneas a la contaminación, donde prevalece el uso de los métodos de superposición de índices ponderados. Los autores de la presente investigación no han encontrado publicaciones sobre el empleo de técnicas de minería de datos para el estudio de la vulnerabilidad a la contaminación en acuíferos cubanos.

(Javadi y Hashemy, 2016) aplican el algoritmo K- medias para evaluar la vulnerabilidad a la contaminación de acuíferos ubicados en la Provincia Alborz, en Irán. Comparan los resultados obtenidos con esta técnica de agrupamiento y con el empleo del método DRASTIC. Concluyen que el mapa de vulnerabilidad obtenido con el algoritmo K-medias presentó mayor precisión al lograr un coeficiente de correlación de Pearson igual a 0,72 entre la concentración de nitrato en las aguas subterráneas y las clases de vulnerabilidad definidas.

Fue estudiada la vulnerabilidad a la salinización de acuíferos costeros provocada por la intrusión salina en la provincia de Mazandaran, al norte de Irán (Motevalli *et al*., 2019) aplicando como técnicas de minería de datos el modelo aditivo generalizado, el modelo lineal generalizado y máquinas de soporte vectorial. El mapa obtenido con la aplicación de estas técnicas reveló de manera precisa las zonas de baja, moderada, alta y muy alta vulnerabilidad a la intrusión salina del acuífero costero estudiado.

El objetivo de esta investigación es evaluar la vulnerabilidad a la contaminación de las aguas subterráneas de la cuenca kárstica Almendares–Vento empleando la técnica de minería de datos análisis de agrupamiento, y comparar los resultados con los obtenidos al aplicar el método RISK, que es un método de superposición de índices ponderados. De esta manera se pretende valorar el poder resolutivo del análisis clúster en la solución de estas tareas.

Métodos

El 47% del volumen total del agua que consume la ciudad de La Habana proviene de la cuenca Almendares–Vento (Herrera *et al.*, 2004) lo que demuestra la importancia del desarrollo de investigaciones que contribuyan a la protección de sus aguas subterráneas ante el efecto de potenciales fuentes contaminantes. Esta cuenca limita por su extremo norte con las Lomas de San Francisco de Paula y el Lomerío de Santa María del Rosario, por su extremo este con las Escaleras de Jaruco, por el sur con las Alturas de Bejucal-Madruga-Coliseo, y por el oeste con las cercanías de la desembocadura del río Almendares y las terrazas marinas emergidas del límite costero norte de los municipios Playa y Plaza de la Revolución (Fig. 1). Los recursos explotables de esta cuenca hidrogeológica están calculados en 287 millones de m^3/año. La red hidrográfica está compuesta principalmente por el río Almendares y sus afluentes, los que transitan por su zona central. Sobre la cuenca existe gran actividad urbana y desarrollo de actividades industriales. Se destaca la industria química, farmacéutica, alimentaria y siderúrgica, así como actividades agropecuarias y múltiples servicios a la población; varios hospitales, centros educacionales, parques, hoteles, centros culturales y un importante desarrollo del transporte que incluye varias carreteras y al principal aeropuerto internacional del país. Esta situación provoca el incremento de la contaminación potencial de la cuenca y la necesidad de su protección.

(Valcarce *et al*. 2020) evaluaron la vulnerabilidad natural a la contaminación de las aguas subterráneas de la cuenca Almendares–Vento aplicando el método RISK, un método de superposición de rangos ponderados que toma su denominación a partir del acrónimo formado por el nombre de las variables que emplea: roca del acuífero (**R**), condiciones de infiltración al acuífero (**I**), propiedades del suelo (**S**), y desarrollo de la red kárstica o karstificación (**K**). El comportamiento de estos parámetros define la mayor o menor protección del medio físico a la contaminación del acuífero (Dörfliger *et al*., 2004).

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

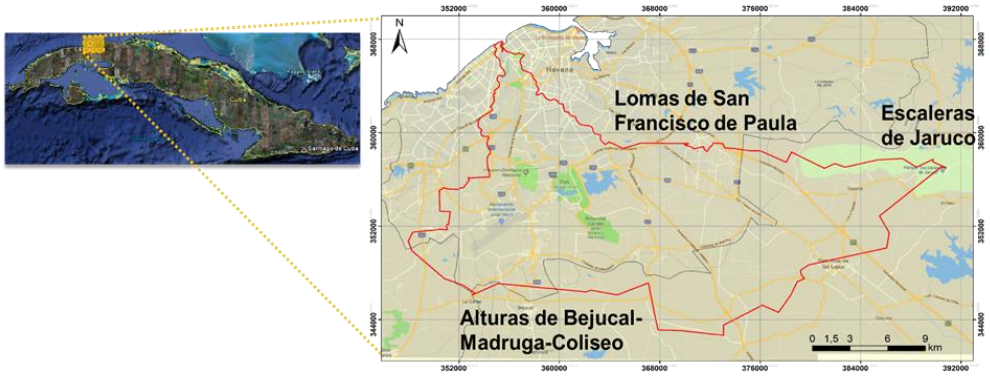

Fig. 1 – Ubicación geográfica de la cuenca Almendares–Vento.

El parámetro roca del acuífero (R) refleja la naturaleza y grado de fracturación de las formaciones geológicas, lo cual tiene gran influencia en el tipo de circulación subterránea y, por lo tanto, en la velocidad de transferencia de un contaminante en el acuífero, fue evaluado a partir del Mapa Geológico de la República de Cuba a escala 1:100 000 (IGP, 2016). El parámetro I tiene en cuenta la pendiente del relieve topográfico porque a mayor pendiente topográfica mayor aceleración de la escorrentía superficial y menor infiltración al acuífero, también considera la presencia de formas kársticas que comunican directamente los flujos de aguas superficiales y el flujo subterráneo, fue evaluado considerando la información del Modelo digital de elevación 10X10 (GEOCUBA, 2010). El parámetro suelo (S) caracteriza a la primera barrera protectora del acuífero; su espesor, textura (guijarros, matriz entre otros) y composición (arcillas, limo) influyen en la mayor o menor vulnerabilidad del acuífero a la contaminación. Este criterio se caracterizó atendiendo al Mapa de Suelos a escala 1:25 000 (Instituto de Suelos, 1990). Por último, el parámetro K describe la influencia de la red kárstica subterránea, la que facilita el transporte del contaminante en el acuífero, criterio que también fue evaluado a partir del Mapa Geológico de la República de Cuba a escala 1:100 000 (IGP, 2016).

Cada uno de estos criterios fueron evaluados y divididos en rangos, y a cada rango fue asignada puntuación entre 0 y 4 (de menos a más vulnerable). Se definió también un factor de peso y finalmente fue calculado el índice de vulnerabilidad *RISK* según la ecuación:

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

El índice *RISK* se separa en rangos y a cada uno se asigna una clase de vulnerabilidad. Este método presenta las limitaciones ya analizadas anteriormente, referidas al nivel de subjetividad que imprime el equipo de investigadores a la separación en rangos de cada parámetro, a la puntuación asignada a cada rango y al factor de ponderación establecido.

El análisis de agrupamiento o análisis clúster, una de las técnicas de la minería de datos, ha demostrado que puede brindar resultados más objetivos en la evaluación de la vulnerabilidad de acuíferos que los métodos de superposición de rangos ponderados, siempre que las variables empleadas en el análisis sean lo suficientemente efectivas. Por esta razón se aplicó una de las técnicas de agrupamiento, el algoritmo K-medias, para evaluar la vulnerabilidad a la contaminación de las aguas subterráneas en esta cuenca.

El análisis clúster es una técnica de reconocimiento de patrones no supervisada, que permite obtener grupos a partir de gran cantidad de datos, de tal manera que los elementos de cada grupo sean muy similares entre sí y, a la vez, sean muy diferentes a los elementos de los otros grupos. Existen disímiles algoritmos de agrupamiento o *clustering*. En esta investigación se empleó el algoritmo K-medias que consiste en asignar cada objeto al grupo más cercano, lo que se logra calculando una medida de similitud entre el objeto y el centroide de cada clúster.

Existen diferentes medidas de similitud, las que resultan más o menos efectivas teniendo en cuenta la naturaleza cuantitativa, cualitativa o binaria de las variables que caracterizan a los objetos a clasificar, así como la ausencia parcial de datos y el nivel de correlación entre las

diferentes variables. La selección adecuada de la medida de similitud juega un rol importante porque en gran medida define que los resultados finales tengan la mayor confiabilidad posible. En esta investigación se empleó como medida de similitud el coeficiente de distancia euclidiano recomendado cuando las variables empleadas se expresan de forma cuantitativa y existe baja correlación estadística entre ellas (Alfonso, 1989; Núñez-Colín y Escobedo-López, 2011).

El número de grupos y sus centroides se calculan inicialmente de forma aleatoria, y después de la primera asignación de objetos a cada grupo, se recalculan los centroides como la media de las variables de los puntos que le fueron asignados. Una vez actualizado el centroide de cada clúster se vuelven a reasignar los objetos al grupo más cercano. Este procedimiento se repite hasta lograr la convergencia, o sea, hasta que las asignaciones de los puntos no cambien, o hasta alcanzar el número de iteraciones prefijado. Este resultado final representa el ajuste que maximiza la distancia entre los distintos grupos y minimiza la distancia intragrupo. La principal ventaja del método es su sencillez y rapidez, pero es un algoritmo significativamente sensible a los centroides que se seleccionan inicialmente de manera aleatoria. Este efecto se puede reducir incrementando el número de iteraciones del procedimiento (Hernández–Orallo, *et al.*, 2004).

El algoritmo es más eficiente en la medida que las variables empleadas no sean redundantes, y es muy sensible al hecho de que las variables posean diferente rango de variación, por lo que se recomienda estandarizarlas entre 0 y 1 sustrayendo a cada variable su valor mínimo y dividiendo por su rango.

Es importante destacar que el algoritmo siempre separará los objetos en grupos. El conocimiento del investigador hará posible identificar qué grupos son significativos y cuáles no. El software utilizado fue WEKA, software libre desarrollado por la Universidad de Waikato, de ahí su nombre: *Waikato Environment for Knowledge Analysis* (Martínez, 2018).

Las variables a emplear deben ser seleccionadas en función de los objetivos del análisis clúster a realizar. En esta investigación, donde el objetivo es identificar en el acuífero las zonas con diferente grado de vulnerabilidad, se han empleado las variables más comúnmente utilizadas en los métodos convencionales para evaluar la vulnerabilidad de acuíferos kársticos: litología del acuífero (Lit), pendiente topográfica del terreno (PenTop), índice de atenuación del suelo a los contaminantes (Ias), densidad de fallas por km^2 (Df), presencia de zonas de infiltración directa (Zi).

La litología del acuífero se emplea por su influencia en la velocidad de transferencia de un contaminante en el acuífero. El algoritmo K-medias requiere que todas las variables sean cuantitativas; al ser la litología un atributo cualitativo se asignaron valores de 1 a 4 a los diferentes tipos de rocas, indicando los mayores valores mayor vulnerabilidad. Se estableció: valor 1 a rocas carbonatadas muy arcillosas y margas, valor 2 a rocas calizas con intercalaciones de margas y bajo contenido de materiales de arcilla, valor 3 a calizas y dolomías masivas poco fracturadas y valor 4 a calizas y dolomías masivas con alta intensidad de fracturación.

La pendiente topográfica del terreno fue calculada a partir del modelo digital de elevaciones (MDE). Como ya fue mencionado, a mayor pendiente topográfica menor infiltración de contaminantes en el acuífero al predominar la escorrentía superficial.

El índice de atenuación del suelo fue un parámetro creado a partir de la suma de tres propiedades del suelo: espesor, contenido de materia orgánica y arcillosidad. El incremento de estas propiedades eleva la capacidad del suelo para retardar la migración vertical de potenciales contaminantes depositados en la superficie del terreno, y por tanto, disminuye la vulnerabilidad a la contaminación del agua subterránea.

La densidad de fallas por km^2 fue calculada empleando la herramienta *Kernel density estimation* del software QGIS. A mayor densidad de fallas, mayor porosidad de fracturas y mayor permeabilidad, lo que facilita la infiltración de los potenciales contaminantes.

Las zonas de infiltración directa se determinaron sustrayendo la malla *Fill sinks,* obtenida utilizando la extensión *Hydrology* de la herramienta *Terrain Analyst* del software QGIS, al modelo digital de elevaciones, para identificar las áreas donde existen formas negativas del relieve, las que pueden estar relacionadas con manifestaciones del paisaje kárstico como dolinas, úvalas, etc. (Pardo-Iguzquiza *et al.*, 2014). Luego, se superpuso la red de drenaje superficial para detectar las zonas de infiltración directa, que se corresponden con las depresiones del relieve donde el flujo de agua superficial pierde su continuidad. Estas constituyen áreas de muy alta vulnerabilidad a la contaminación del agua subterránea porque permiten la conexión directa de la superficie con la red kárstica subterránea y, por tanto, la comunicación inmediata de cualquier contaminante con el acuífero.

La tabla 1 resume las fuentes de datos de donde fueron extraídas las variables seleccionadas y su rango de variación. Todos los mapas de estas variables fueron elaborados a escala 1:100 000 en formato *ráster*. Para construir la base de datos se tomaron los valores de cada atributo cada 25 metros en el área de la cuenca conformándose un total de 719 131 instancias.

Tabla 1 - Fuente de datos y rango de variación de cada variable seleccionada.

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	0% – 118%	Modelo digital de elevación

(PenTop)		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)

Para asociar a cada cluster el grado de vulnerabilidad a la contaminación que le corresponde, se empleó el concepto de punto ideal (Vías *et al.*, 2003), el cual actúa como referencia de la situación más favorable, o sea, de mayor protección del acuífero. Ese punto se encuentra en la roca menos permeable, con mayor pendiente topográfica, mayor índice de atenuación del suelo, menor densidad de fallas y ausencia de zonas de infiltración directa. Para el acuífero bajo estudio las coordenadas del punto ideal en el espacio euclidiano n dimensional son: Lit = 1; PenTop= 118%; Ias = 143; Df = 0, Zi = 0. La distancia de cada grupo al punto ideal en el espacio euclidiano n dimensional se calculó como:

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

donde:

d_{ij}: distancia euclidiana del centroide del clúster i al punto ideal p_j

x_{ik}: centroide del clúster i, o sea, punto que tiene como coordenadas el valor medio de las n variables de los puntos que integran el clúster i.

p_j: coordenadas del punto ideal.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

Resultados y discusión

A continuación, se analizan los resultados obtenidos al evaluar la vulnerabilidad a la contaminación de las aguas subterráneas en la cuenca Almendares-Vento empleando el método RISK y el algoritmo K-medias.

La figura 2 muestra la cartografía de los criterios R, I, S y K en la cuenca, a escala 1: 100 000. Se aprecia que, según los criterios roca del acuífero (R), infiltración al acuífero (I) y desarrollo de la red kárstica (K), en la cuenca existe predominio de alta y muy alta vulnerabilidad. El parámetro suelo (S) presenta mayor capacidad de protección y refleja predominio de vulnerabilidad moderada y alta.

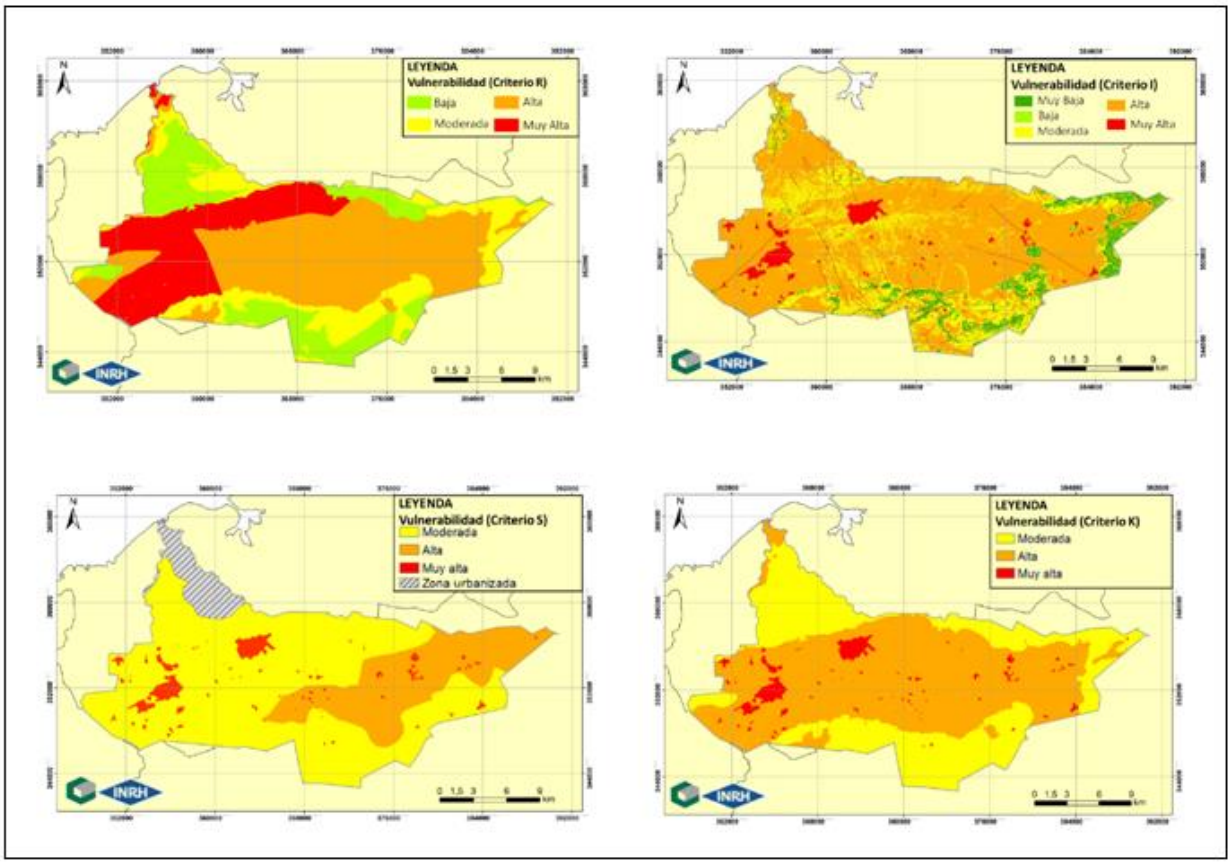

Fig. 3 – Mapa de los criterios R, I, S y K en la cuenca Almendares–Vento. Escala 1:100 000

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

Una vez calculado el índice RISK, fue separado en rangos y quedó clasificado como se presenta en la tabla 2.

Tabla 2 - Clasificación del índice de vulnerabilidad según el método RISK

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

(Modificado de Dörfliger *et al*., 2004)

La figura 4 presenta el mapa de vulnerabilidad resultante de la representación espacial del índice RISK.

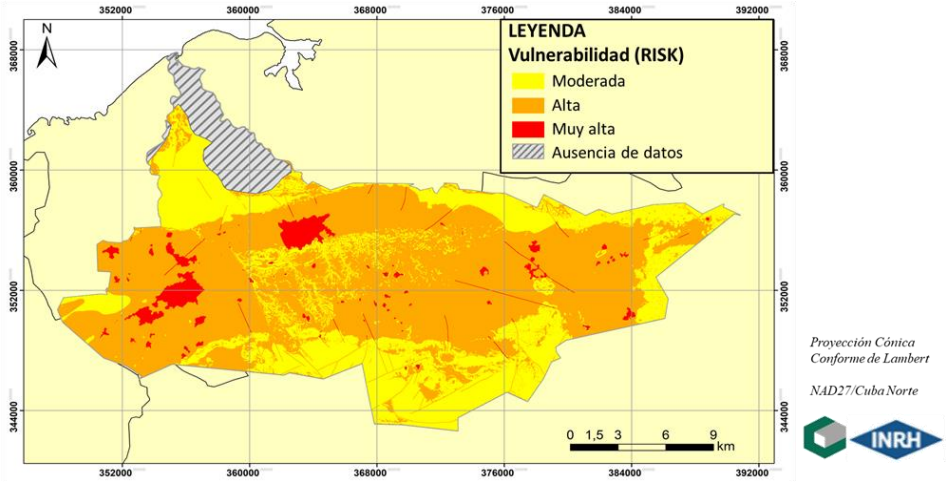

Fig. 4 – Mapa de vulnerabilidad intrínseca a la contaminación del agua subterránea en la cuenca Almendares–Vento obtenido por el método RISK.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

La parte central de la cuenca presenta alta vulnerabilidad del agua subterránea a la migración vertical de potenciales contaminantes depositados en su superficie. En esta zona, que representa el 62,4% del área, existe un gran desarrollo de rocas carbonatadas karstificadas lo que le confiere un grado de vulnerabilidad mayor con respecto a otras zonas donde predominan rocas más arcillosas. El 4% del área se clasifica de muy alta vulnerabilidad y coincide con las zonas de infiltración directa. El 33,6% restante corresponde a zonas de vulnerabilidad moderada, donde predominan rocas con mayor grado de arcillosidad, y pendientes topográficas elevadas.

Para aplicar el algoritmo K-medias es necesario que exista baja correlación entre las variables seleccionadas. La tabla 3 muestra que se cumple esta premisa.

Tabla 3 - Matriz de correlación lineal entre las variables seleccionadas para aplicar el método K-medias.

	Lit	**PenTop**	**Ias**	**Df**	**Dzi**
Lit	1	-0,34	0,27	-0,41	0,2
PenTop	-0,34	1	-0.15	0,23	-0,09
Ias	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

Los centroides, la distancia de cada centroide al punto ideal y el grado de vulnerabilidad asignado a cada grupo se presentan en la tabla 4.

Tabla 4 - Centroides de cada grupo, distancia al punto ideal y grado de vulnerabilidad de cada clúster.

	CENTROIDES	**Distanci**	**Grado de**

CLÚSTER	TOTAL DE INTANCIAS	Lit	PenTop	las	Df	Zi	a al punto ideal	vulnerabilidad
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

Analizando la distancia de cada centroide al punto ideal fue posible interpretar de manera directa el nivel de vulnerabilidad a la contaminación de las aguas subterráneas asociado a cada cluster.

El grupo identificado como muy alta vulnerabilidad es el más alejado del punto ideal y se caracteriza por presentar zonas de infiltración directa de potenciales contaminantes al acuífero, menor pendiente topográfica, menor índice de atenuación del suelo y rocas con alto desarrollo kárstico. El grupo clasificado como de muy baja vulnerabilidad es aquel donde no existe infiltración directa, el suelo posee el mayor índice de atenuación, se encuentran las mayores pendientes y la litología se caracteriza por el predominio de rocas carbonatadas muy arcillosas y margas, y su centroide se encuentra más cerca del punto ideal.

La figura 5 muestra el mapa de vulnerabilidad obtenido con la aplicación del análisis clúster. En la zona central de la cuenca, donde según el método RISK predomina la alta vulnerabilidad, el análisis clúster logra identificar tres categorías. Hacia el oeste clasifica como alta vulnerabilidad y corresponde con la presencia de la Formación Geológica Güines, donde existe un mayor desarrollo de manifestaciones kársticas superficiales, menor pendiente topográfica y menor

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 2, Abril-Junio, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 1-23

índice de atenuación del suelo; hacia el este se identifica una zona con vulnerabilidad moderada asociada a la Formación Colón, que se caracteriza por la presencia de rocas más arcillosas y formas exokársticas con menor desarrollo; en la periferia de la zona central predomina la baja y muy baja vulnerabilidad debido a la presencia de suelos con mayor espesor, contenido de materia orgánica y arcillosidad, o sea, con mayor capacidad protectora, así como la presencia de litologías más arcillosas y mayor pendiente topográfica. Las zonas de muy alta vulnerabilidad a la contaminación del agua subterránea coinciden en los mapas obtenidos por el método RISK y por el algoritmo K-medias, y son aquellas donde existe infiltración directa de los potenciales contaminantes depositados en la superficie.

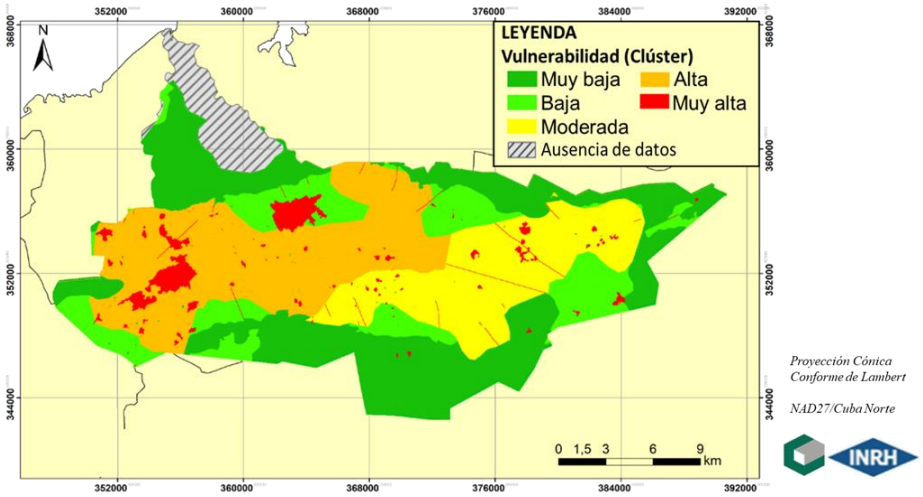

Fig. 5 – Mapa de vulnerabilidad intrínseca a la contaminación del agua subterránea en la cuenca Almendares–Vento obtenido aplicando el algoritmo K-medias.

Para integrar la información aportada por el método RISK y por el algoritmo K-medias, se presenta la figura 6, donde puede apreciarse claramente que el empleo de clasificación no supervisada permitió un mayor poder resolutivo para cartografiar la vulnerabilidad natural a la contaminación del acuífero, al brindar un modelo con cinco clases de susceptibilidad a la degradación del agua subterránea.

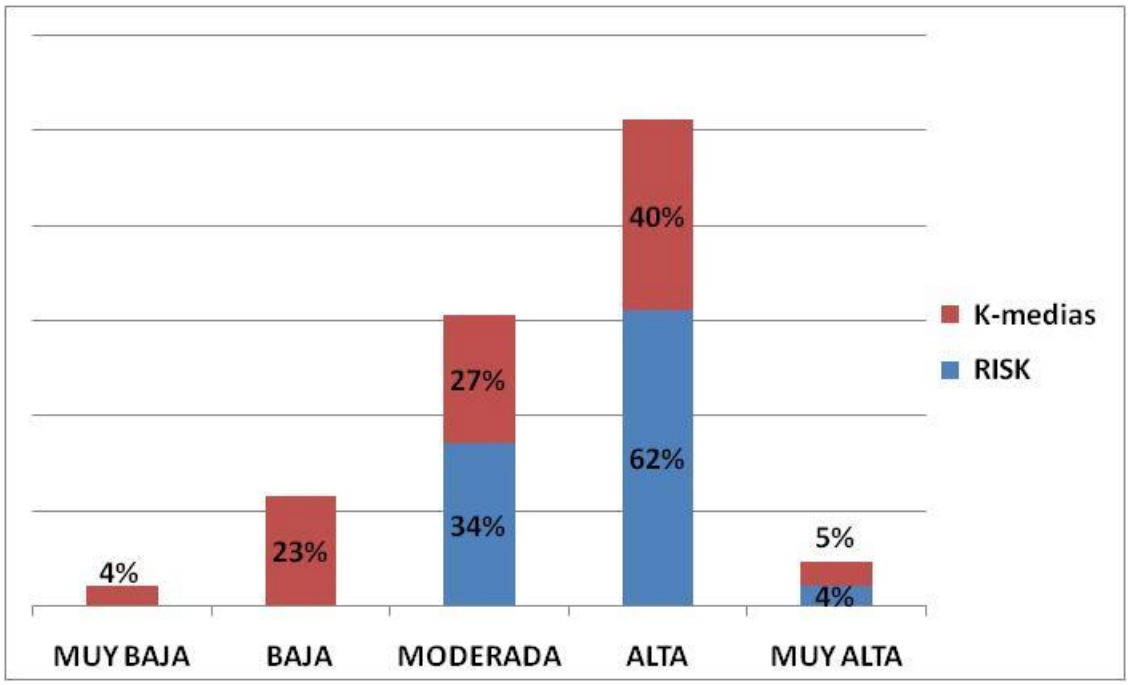

Fig. 6 – Comparación entre clases de vulnerabilidad intrínseca a la contaminación del agua subterránea en la cuenca Almendares–Vento aplicando el método RISK y el algoritmo K-medias.

Desde el punto de vista económico y ambiental el aporte de investigaciones de este tipo es invaluable, teniendo en cuenta que la disponibilidad y calidad del agua representa uno de los principales desafíos para cualquier país y Cuba no es la excepción. De manera particular, la cuenca Almendares-Vento presenta importancia de primer orden en el desarrollo económico y social de Cuba, y su contaminación por la acción irresponsable del hombre podría ser irreversible, o requerir enormes recursos y tiempo para que las acciones de remediación fueran efectivas. Orientar políticas correctas en el ordenamiento territorial de esta cuenca y garantizar la protección de sus recursos hídricos subterráneos resulta imprescindible, y para ello, evaluar la vulnerabilidad natural del acuífero representa el primer paso. El costo de estas investigaciones no es elevado, considerando que la información necesaria para acometer con éxito estas tareas se encuentra disponible en archivos y bases de datos de empresas vinculadas a la actividad

geológica del país, y existen los recursos humanos y la capacidad técnica y profesional para su desarrollo.

Conclusiones

Fue obtenido un modelo de clasificación con cinco clusters asociados a diferente grado de sensibilidad a la contaminación de las aguas subterráneas que logra discriminar zonas de muy alta, alta, moderada, baja y muy baja vulnerabilidad en la cuenca kárstica Almendares–Vento. Este resultado manifiesta mayor poder resolutivo que el método RISK, con el que solo fueron identificadas tres clases de vulnerabilidad. Se constata que esta técnica de clasificación estadística no supervisada no presenta las desventajas de los métodos de superposición de índices ponderados comúnmente usados para evaluar la vulnerabilidad de acuíferos, y permite una cartografía más objetiva y precisa de la vulnerabilidad del acuífero a la contaminación

La investigación desarrollada contribuye a los estudios de protección de los acuíferos en Cuba. Por su alta eficiencia y poder resolutivo se recomienda el uso del algoritmo K-medias para evaluar otras importantes cuencas del territorio nacional.

Referencias

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 http://www.gacetaoficial.cu/

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

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: https://doi.org/10.1063/1.4991238

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.

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: 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=2ahUKEwj-176iprLqAhVHUt8KHUsLBpQQkeECKAB6BAgLECo&biw=1999&bih=979

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.

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

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

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.

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

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.

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

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: http://dx.doi.org/10.1007/s10115-003-0128-3

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: https://doi.org/10.1016.B978-0-12-815226-3.00025-9

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.

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.

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.

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.

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.

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.

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.

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.

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.

Conflicto de interés

Los autores del artículo **Aplicación de la minería de datos a la evaluación de la vulnerabilidad de acuíferos** declaran que no existen conflictos de intereses y autorizan la distribución y uso del artículo.

Contribuciones de los autores

1. Conceptualización: Rosa María Valcarce Ortega
2. Curación de datos: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda.
3. Análisis formal: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño.
4. Adquisición de fondos: Rosa María Valcarce Ortega
5. Investigación: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño.
6. Metodología: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño.
7. Administración del proyecto: Rosa María Valcarce Ortega,
8. Recursos: Rosa María Valcarce Ortega,
9. Software: Willy Rodríguez Miranda, Oscar Suárez González.
10. Supervisión: Rosa María Valcarce Ortega,
11. Validación: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño
12. Visualización: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño
13. Redacción – borrador original: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño

14. Redacción – revisión y edición: Rosa María Valcarce Ortega, Oscar Suárez González, Willy Rodríguez Miranda, Marina Vega Carreño

Financiación

Departamento de Geociencias, Facultad de Ingeniería Civil, Universidad Tecnológica de La Habana “José Antonio Echeverría”, CUJAE.