Revista GEOCIÊNCIAS

unesp

ISSN: 1980-900X (online)

APLICACIÓN DE LA MINERIA DE DATOS A LA EVALUACION DE LA VULNERABILIDAD DE ACUIFEROS EN LAS CUENCAS CUYAGUATEJE Y COSTERA SUR DE PINAR DEL RÍO, CUBA

DATA MINING APPLICATION TO AQUIFERS VULNERABILITY ASSESSMENT IN CUYAGUATEJE AND COSTERA SUR BASINS OF PINAR DEL RÍO, CUBA.

Rosa María VALCARCE, Ernesto Solís MORALES

Universidad Tecnológica de La Habana "José Antonio Echeverría" (CUJAE). 3H3M+XJ6, La Habana, Cuba.
E-mails: rvalcarce1959@gmail.com; solismoralesernesto@gmail.com

RESUMO - As águas subterrâneas representam mais de 97% do total de água doce disponível na Terra e, obviamente, precisam de proteção. É necessário prevenir a deterioração da qualidade das águas subterrâneas e, neste sentido, os mapas de vulnerabilidade intrínseca à contaminação dos aquíferos fazem parte de um sistema de alerta precoce. Métodos de sobreposição de índices ponderados são comumente usados para mapear a vulnerabilidade de aquíferos, mas apresentam um conjunto de desvantagens que indicam a necessidade de aplicação de métodos alternativos que permitam obter resultados mais precisos. O objetivo desta pesquisa foi avaliar a vulnerabilidade à contaminação das águas subterrâneas nas bacias Cuyaguateje e Costera Sur de Pinar del Río, Cuba, usando técnicas de mineração de dados, análise de agrupamento e de árvore de decisão, e comparar os resultados com os obtidos pela aplicação do RISK método, que é um método de superposição de índices ponderados. As variáveis selecionadas para aplicação dessas técnicas de classificação foram: declividade topográfica do terreno, litologia do aquífero, índice de atenuação do solo aos contaminantes e densidade de dolinas por km^2. A análise de agrupamento não supervisionada e supervisionada foi capaz de discriminar melhor as áreas com diferentes graus de vulnerabilidade, demonstrando seu maior poder resolutivo.
Palavras-chave: Vulnerabilidade intrínseca de aquíferos. Águas subterráneas. Análise de agrupamento. Árvore de decisão.

ABSTRACT - Groundwater accounts for more than 97% of the total freshwater availability on Earth and obviously needs protection. It is necessary to prevent quality groundwater deterioration and aquifer intrinsic vulnerability maps are part of an early warning system. Weighted index overlay methods are commonly used to map aquifer vulnerability. However, these methods have disadvantages that make it convenient to apply alternative methods to obtain more accurate results. The research objective was evaluate groundwater contamination vulnerability in Cuyaguateje and Costera Sur basins of Pinar del Río, Cuba, using data mining techniques (cluster analysis and decision tree), and compare the results with those obtained by the RISK method, which is a weighted index superposition method. The variables used were: land topographic slope, aquifer lithology, soil attenuation index to pollutants and sinkholes density per square kilometer. The unsupervised an supervised grouping analysis was able to better discriminate the areas with different degrees of vulnerability, demonstrating greater resolving power.
Keywords: Aquifer intrinsic vulnerability. Groundwater. Cluster analysis. Decision tree.

INTRODUCCIÓN

La minería de datos se aplica con la finalidad de extraer patrones, describir tendencias, predecir comportamientos, y de esta forma extraer información útil y conocimiento de grandes cantidades de datos. Emplea algoritmos que se dividen en exploración, estadístico, clasificación y aprendizaje automático. Estos algoritmos se escogen en dependencia de la tarea a realizar teniendo en cuenta distintos criterios, como por ejemplo, el área del conocimiento donde se aplica y la complejidad de los datos (Medina & Gómez, 2014). La minería de datos ha sido utilizada en casi todas las esferas de la vida; en el comercio y sector financiero, en la medicina, en los deportes, en las ciencias sociales y en la tecnología. Su uso se ha expandido al campo de la Geociencias donde se reportan excelentes aplicaciones en la exploración y explotación de yacimientos de petróleo y gas, y de yacimientos minerales sólidos, en la cartografía geológica, y en años recientes en la evaluación de la vulnerabilidad a la contaminación de las aguas subterráneas.

La Asamblea General de las Naciones Unidas reconoce el recurso Agua como uno de los desafíos globales más importantes. Identifica que el agua es un elemento fundamental para lograr el desarrollo sostenible al ser un recurso imprescindible para alcanzar el desarrollo socioeconómico, la producción de alimentos, la conservación de los ecosistemas y la supervivencia humana. Destaca

que el cambio climático está influyendo negativamente en la cantidad y calidad del agua disponible a nivel mundial, poniendo en peligro el logro del Objetivo de Desarrollo Sostenible número 6 de la Agenda 2030 de las Naciones Unidas, que plantea como meta el acceso al agua limpia y el saneamiento para todos en los próximos diez años (UNESCO, 2020).

Las reservas de agua subterránea abastecen al 80% de la población mundial, representan una reserva a largo plazo para hacer frente a emergencias, sequías y cambios climáticos. Sin embargo, aun cuando son más resilientes a la contaminación que las aguas superficiales, los núcleos poblacionales, la industria, la minería y las explotaciones agropecuarias, son fuentes potenciales de su contaminación. Entre los contaminantes más comunes de las aguas subterráneas están los patógenos originados por la materia orgánica y los nutrientes producidos por aguas residuales, la agricultura y la ganadería; los residuos sólidos y sustancias tóxicas procedentes de actividades industriales y mineras, así como de la agricultura intensiva; y las sales disueltas como consecuencia de la intrusión salina.

Con el objetivo de proteger la calidad de las aguas subterráneas es necesario conocer la vulnerabilidad de los acuíferos. En términos generales se entiende por vulnerabilidad de un acuífero el grado de protección natural que ofrece el medio geológico a la contaminación. La vulnerabilidad natural de un acuífero, o su vulnerabilidad intrínseca, depende de sus características geológicas, de sus parámetros hidrogeológicos (conductividad hidráulica y transmisividad), de la presencia, espesor y características hidrogeológicas del suelo que lo cubre y de la zona no saturada. En los últimos años se han desarrollado varias metodologías para elaborar mapas de vulnerabilidad de un acuífero a la contaminación, y entre estas metodologías se destacan aquellas que se fundamentan en la superposición de índices ponderados (Vargas, 2010). Estos métodos se basan en la combinación de diferentes parámetros que se dividen en rangos, a los que se asigna determinada puntuación y un factor de ponderación para cuantificar su influencia en la vulnerabilidad del acuífero.

Posiblemente sea DRASTIC el método de superposición de índices ponderados más popular y más empleado para elaborar mapas de vulnerabilidad de acuíferos. Fue desarrollado desde la década del 80 del pasado siglo por investigadores de la Agencia de Proyección del Medio Ambiente de Estados Unidos (Aller et al., 1987). Este método 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. Desde entonces otros métodos han sido desarrollados, como el método AVI (Stempvoort et al., 1993), EPIK (Dörfliger et al., 1999), RISK (Dörfliger et al. 2004), COP (Vías et al., 2006), GOD (Foster et al., 2007), entre otros

Recientemente, Tziritis et al. (2020) desarrolla el método RIVA, un nuevo método de superposición de índices para evaluar la vulnerabilidad intrínseca a la contaminación del agua subterránea, que incluye cuatro parámetros principales: recarga del acuífero (R), condiciones de infiltración (I), protección que ofrece la zona vadosa (V) y conductividad hidráulica del acuífero (A).

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 también dependen de la experiencia de los investigadores y de criterios muchas veces subjetivos (Valcarce et al., 2021).

La aplicación de las técnicas de minería de datos en la cartografía de la vulnerabilidad de los acuíferos representa una estrategia alternativa para eliminar las desventajas propias de los métodos paramétricos ponderados, y permite obtener resultados más objetivos, e incluso, en muchas ocasiones, con una mayor capacidad de discriminación espacial.

En sus investigaciones, Deepesh et al. (2018) analizan la importancia de los mapas de vulne-

rabilidad para evaluar la contaminación potencial de las aguas subterráneas a la contaminación, las ventajas y desventajas de los diferentes métodos aplicados comúnmente en la elaboración de los mismos, y la tendencia actual de aplicar técnicas de minería de datos para lograr mayor precisión en la elaboración de estos mapas.

Yoo et al. (2016) evaluó la sensibilidad a la contaminación del agua subterránea en un acuífero de Korea empleando diferentes algoritmos de minería de datos, entre ellos redes neuronales artificiales y árbol de decisión. Sus resultados mostraron que los modelos propuestos estiman con mayor precisión la vulnerabilidad del agua subterránea a los contaminantes, que los modelos elaborados anteriormente empleando métodos de superposición de índices.

Las técnicas de minería de datos aplicando el algoritmo K- medias para evaluar la vulnerabilidad a la contaminación de acuíferos ubicados en la Provincia Alborz, en Irán, también fueron aplicadas por Javadi et al. (2017). Estos investigadores compararon los resultados obtenidos con esta técnica de agrupamiento y con el método DRASTIC, concluyendo 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.

Nadiri et al. (2018) destacan el poder resolutivo de las técnicas de aprendizaje no supervisado y supervisado para evaluar la vulnerabilidad intrínseca a la contaminación de un acuífero en Irán. Demuestran que la correlación estadística entre la concentración de nitratos en el agua subterránea y las clases de vulnerabilidad obtenidas por las técnicas de minería de datos es mayor que con las clases de vulnerabilidad definidas por el método DRASTIC.

Fue estudiada la vulnerabilidad a la salinización de acuíferos costeros provocada por la intrusión salina en la provincia de Mazandaran, al norte de Irán aplicando como técnicas de minería de datos el modelo aditivo generalizado, el modelo lineal generalizado y máquinas de soporte vectorial (Motevalli et al., 2019). 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.

Valcarce et al. (2021) evaluaron la vulnerabilidad a la contaminación de las aguas subterráneas de la cuenca kárstica Almendares–Vento en la provincia La Habana, Cuba, empleando la técnica de clasificación no supervisada análisis de agrupamiento. Las variables empleadas 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. Los resultados obtenidos demostraron mayor poder resolutivo que los métodos de superposición de índices ponderados aplicados anteriormente sobre esta cuenca, al obtenerse un mapa con mayor discriminación espacial de las zonas con diferentes grados de vulnerabilidad.

El objetivo de la presente investigación es profundizar en la evaluación de la vulnerabilidad intrínseca a la contaminación de las aguas subterráneas en las cuencas Cuyaguateje (PI) y Costera Sur de Pinar del Río (PII), en Cuba. La vulnerabilidad de estas cuencas fue evaluada según el método RISK (Valcarce & Solis, 2021) y en la presente investigación se re-evalúan empleando los algoritmos K – media y Árbol de Decisión como técnicas de clasificación no supervisada y supervisada respectivamente.

MATERIALES Y MÉTODOS

Las cuencas hidrogeológicas Cuyaguateje (PI) y Costera Sur (PII) pertenecen a la provincia de Pinar del Río, situada al sur de la zona más occidental de Cuba. Se extienden desde la península de Guanahacabibes hasta el oeste de la provincia de Artemisa.

Al norte limitan con el litoral de la referida península y la cordillera de Guaniguanico, mientras que hacia el sur limitan con las aguas del mar Caribe (Figura 1). Estas cuencas hidrogeológicas constituyen las principales fuentes de abasto de agua subterránea a la provincia de Pinar del Río. Los horizontes acuíferos de mayor interés están presentes en las calizas fosilíferas muy karstificadas de la formación Vedado y, en menor grado, de la formación Jaimanitas, así como en los depósitos terrígeno-carbonatados de la formación Paso Real.

La columna estratigráfica regional se caracteriza por sedimentos que abarcan desde el Jurásico hasta el Cuaternario reciente, y las formaciones que conforman este territorio están formadas en su mayoría por rocas carbonatadas donde existe un marcado desarrollo kárstico, sobre todo en las

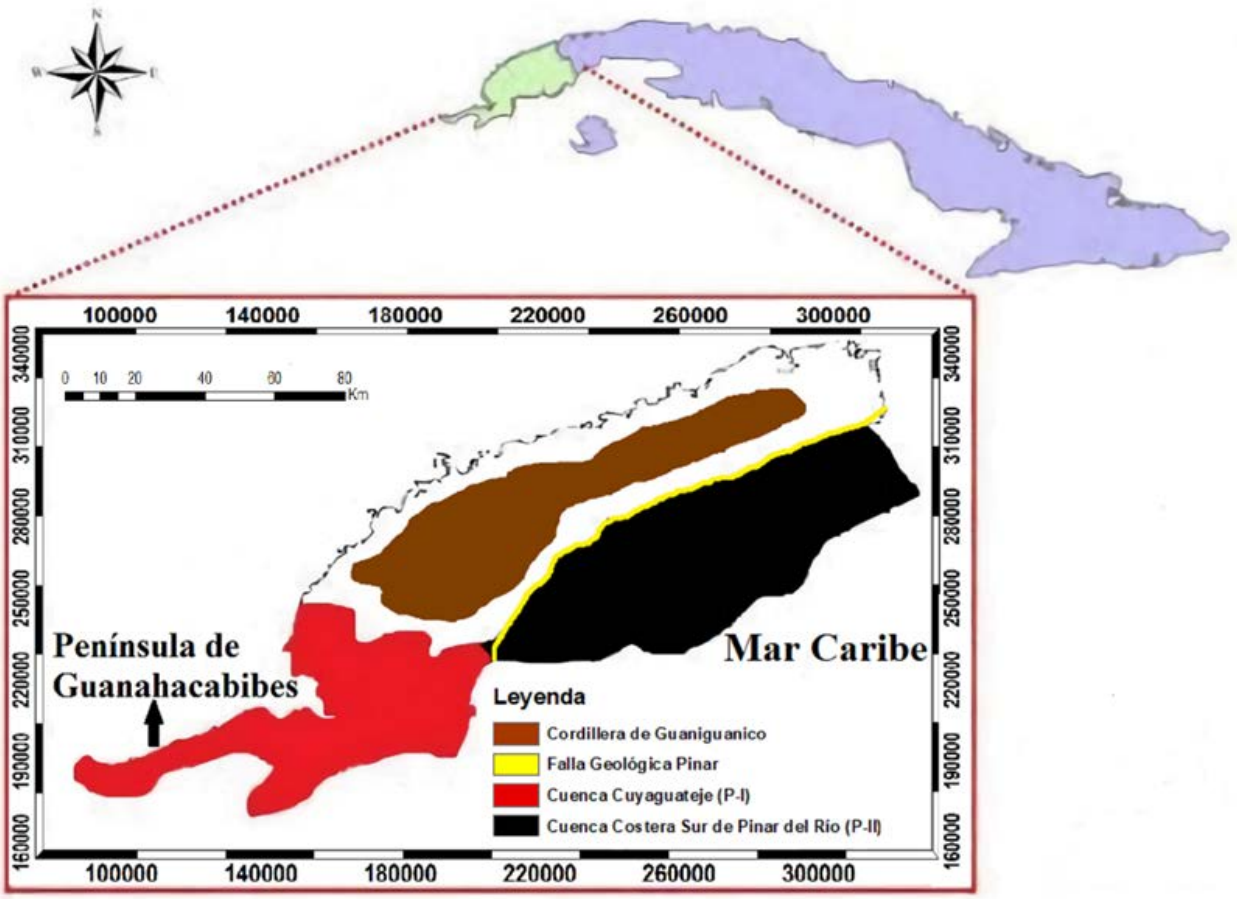

Figura 1 - Área de estudio (Valcarce & Solis, 2021)

formaciones Vedado, Paso Real y Jaimanitas.

La provincia de Pinar del Río está marcada por las subregiones geomorfológicas Llanura de Guanahacabibes y Llanura Sur de Pinar del Río, las cuales poseen un relieve relativamente llano con pocas alturas. En esta zona se ubican importantes asentamientos poblacionales siendo el más importante la ciudad de Pinar del Río, capital de la provincia del mismo nombre. También existe una importante actividad agrícola que requiere de estos recursos hidráulicos para su desarrollo

Valcarce & Solís (2021) evaluaron la vulnerabilidad natural a la contaminación de las aguas subterráneas en las cuencas PI y PII 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).

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 parámetro 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 parámetros fueron 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 para cada criterio y finalmente fue calculado el índice de vulnerabilidad *RISK* según la ecuación:

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

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. Por estas razones se decide aplicar los algoritmos K-medias y Árbol de Decisión para reevaluar la vulnerabilidad a la contaminación del agua subterránea en estas cuencas.

El algoritmo K-medias 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. Este algoritmo permite 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. 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 (Hamdan & Emad, 2017).

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; Imron et al., 2020).

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 algoritmo Árbol de Decisión tiene como objetivo desarrollar un gráfico de clasificación supervisada y predecir la clase objetivo en función del conjunto de datos de entrenamiento de entrada. Estés algoritmos puede aprender las relaciones entre las variables de entrada y las salidas correspondientes, y representar cada relación mediante reglas específicas. El árbol de decisión se visualiza similar a un diagrama de flujo, donde cada hoja del árbol está representada por los atributos utilizados, debajo de estas se encuentran los nodos que representan el valor del atributo y cada rama representa un resultado de la prueba (Jeihouni et al., 2019).

El software utilizado fue WEKA, software libre desarrollado por la Universidad de Waikato, de ahí su nombre: *Waikato Environment for Knowledge Analysis* (Martínez, 2018).

Para la evaluación de la vulnerabilidad a la contaminación de las aguas subterráneas mediante métodos estadísticos es necesario seleccionar los atributos a utilizar, es decir, se debe realizar una selección de las variables de predicción de la vulnerabilidad, las que no deben ser redundantes, o sea, no deben estar correlacionadas. Posterior-mente se deben analizar las propiedades del conjunto de datos seleccionados, observar las tendencias, los valores atípicos y la cantidad de clases de vulnerabilidad posibles.

En esta investigación se confeccionó una base de datos compuesta por los atributos siguientes: pendiente topográfica (PendTop), litología (Lit), índice de atenuación del suelo (IA), densidad de dolinas por Km^2 (DD). Estos atributos se digi-

talizaron cada 100 metros para un total de 575 808 puntos en el área de estudio. A continuación se justifica brevemente la selección de estos atributos.

A mayor pendiente topográfica (PendTop) predomina en el área la escorrentía superficial y por lo tanto, hay menor infiltración de contaminantes hacia el acuífero. Este atributo fue extraído del Modelo Digital de Elevaciones (MDE) a partir de la función *Slope* en la herramienta *Spatial Analyst* del sistema de información geográfico QGIS

La litología del acuífero (Lit) se empleó porque la composición mineralógica y grado de fracturación de las rocas, influyen en la velocidad de transferencia de un contaminante en el acuífero. El algoritmo K-medias requiere que todas las variables sean cuantitativas y 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 esta-bleció: valor 1 a margas y rocas muy arcillosas, 2 a rocas con intercalaciones de arcillas, 3 a calizas y dolomías poco fracturadas, y 4 a rocas carbonatadas y dolomías muy fracturadas.

El índice de atenuación del suelo (IA) 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 dolinas por km^2 (DD) fue calculada empleando la herramienta *Kernel density estimation* del software QGIS. Un incremento de este parámetro provoca mayor probabilidad de que ocurra infiltración directa de los contaminantes al acuífero, es decir, mayor vulnerabilidad a la contaminación del agua subterránea.

La tabla 1 presenta las fuentes 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*.

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

Variable	Rango de variación	Fuente de datos
Pendiente Topográfica (PendTop)	0 – 130	Modelo digital de elevación 10X10 (GEOCUBA, 2010)
Litología (Lit)	1 –4	Mapa Geológico de la República de Cuba a escala 1:100 000 (IGP, 2016)
Índice de atenuación del suelo (IA)	0 - 70	Mapa de Suelos a escala 1:25 000 Instituto de Suelos (1990).
Densidad de dolinas por km^2 (DD)	0 – 5	Mapa Geológico de la República de Cuba a escala 1:100 000 (IGP, 2016)

RESULTADOS Y DISCUSIÓN

A continuación se analizan y comparan los resultados obtenidos al evaluar la vulnerabilidad a la contaminación de las aguas subterráneas en las cuencas PI y PII empleando el método RISK, y empleando las técnicas de clasificación estadística no supervisada (K-medias) y supervisada (Árbol de Decisión).

La figura 2 muestra el mapa de vulnerabilidad a la contaminación de las aguas subterráneas obtenido por el método RISK en las cuencas estudiadas según Valcarce & Solis (2021). Se aprecia que fueron definidas zonas de vulnerabilidad muy alta, alta, moderada y baja, con predominio de zonas de moderada y alta vulnerabilidad.

Las zonas clasificadas con muy alta vulnerabilidad se localizan en la Península de Guanahacabibes y constituyen el 13,7% de la región de estudio.

Las zonas clasificadas de alta vulnerabilidad representan un 29,9 % del area total y en ella se encuentran las principales formaciones acuíferas de la cuenca Costera Sur de Pinar del Río, desarrolladas en las formaciones geológicas Paso Real y Güines. Las zonas clasificadas de vulnerabilidad moderada se sitúan mayormente en la parte sur y central del área de estudio y alcanzan el 55,1 % de la región. Por último, las zonas de baja vulnerabilidad ocupan solo el 1,3 % del área.

Para aplicar las técnicas de minería de datos las variables a emplear no deben ser redundantes. La tabla 2 muestra que se cumple esta premisa al mostrar que no existe correlación entre los atributos empleados.

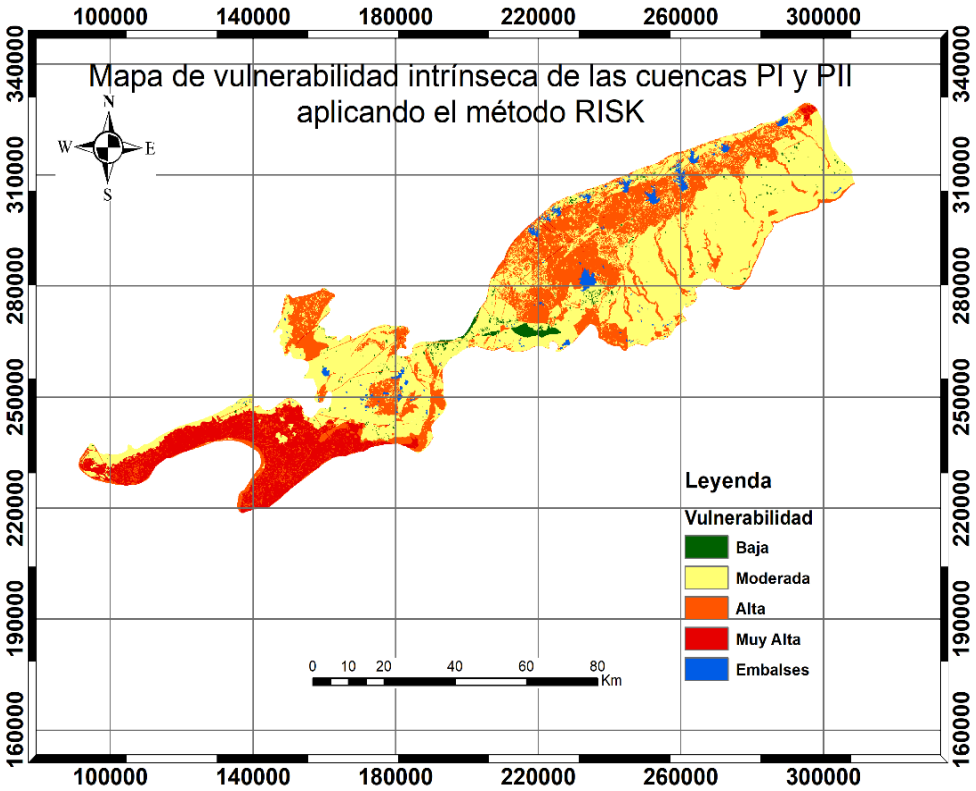

Figura 2 - Mapa de vulnerabilidad a la contaminación de las aguas subterráneas obtenido por el método RISK en las cuencas Cuyaguateje y Costera Sur de Pinar del Río, Escala 1:100 000 (Valcarce & Solis, 2021).

Tabla 1 - Matriz de correlación lineal entre los atributos empleados.

Atributo	PendTop	Lit	IA	DD
PendTop	1			
Lit	0,052	1		
IA	0,112	- 0,063	1	
DD	0,185	0,110	0,053	1

Con el algoritmo K- medias se obtuvieron modelos de clasificación con dos, tres, cuatro, cinco y seis grupos. Para seleccionar el número óptimo de grupos se aplicó el Método del Codo, que consiste en calcular la suma del error cuadrático dentro de cada clúster para cada modelo de clasificación (Moya, 2016). Este parámetro se calcula como la suma de las distancias al cuadrado de cada punto al centroide al cual pertenece. El número óptimo de conglomerados se define cuando incluir más clústeres no produce disminución significativa en la suma del error cuadrático dentro de los grupos. Según este criterio el modelo de agrupación con cinco clústeres resultó el óptimo.

La tabla 3 presenta los resultados obtenidos al aplicar el algoritmo K-medias para el modelo de clasificación con cinco grupos. El grado de vulnerabilidad asociado a cada clúster fue asignado interpretando los valores de los centroides de cada grupo.

Tabla 3 - Centroides y grado de vulnerabilidad de cada clúster.

		CENTROIDES				
CLUSTER	**TOTAL DE INSTANCIAS**	**PendTop (%)**	**Lit**	**IA**	**DD**	**GRADO DE VULNERABILIDAD**
1	46 582	31,6	1,4	62,7	0,39	Muy Baja
2	42 119	32,0	3,2	64,1	0,32	Baja
3	247 650	21,3	1,2	12,8	0,16	Moderada
4	67 099	49,6	2,8	15,5	1,81	Alta
5	172 358	22,0	3,6	10,5	0,16	Muy Alta

Los clústeres 1 y 2 presentan muy elevados valores del índice de atenuación del suelo, lo que imprime baja vulnerabilidad a estos grupos. Atendiendo a la variable litología se clasifica el clúster 1 de muy baja vulnerabilidad y el clúster 2 de baja vulnerabilidad.

Los clústeres 3, 4 y 5 presentan valores de índice de atenuación del suelo muy bajo, por lo que deben asociarse a estos grupos vulnerabilidad entre muy alta y moderada.

El clúster 4 presenta altos valores de densidad de dolinas lo que significa alta capacidad de infiltración de los contaminantes. Además, presenta mayor pendiente topográfica que el grupo 3, y su litología está compuesta fundamentalmente por rocas arcillosas y poco fracturadas. Por todo lo anterior se asocia a este grupo una alta vulnerabilidad.

La variable litología en el clúster 5 está constituida predominantemente por rocas muy fracturadas, lo que unido a los muy bajos valores del índice de atenuación del suelo y de la pendiente topográfica permite inferir que este grupo presenta la mayor vulnerabilidad del conjunto, por lo que clasifica como de muy alta vulnerabilidad.

Finalmente el análisis de los centroides del clúster 3 permite concluir que representa moderada vulnerabilidad dentro del conjunto.

Este modelo de agrupamiento fue depurado aplicando un método de clasificación supervisada, el algoritmo Árbol de Decisión. La tabla 4 presenta la denominada matriz de confusión, que muestra que el 99,93% de las instancias no fueron movidas a otro clúster, y solo el 0.07% fueron las instancias modificadas.

Se destaca que la mayor depuración se produce en el clúster de alta vulnerabilidad, donde se reclasifican 206 instancias, al pasar 13 a muy baja, 7 a baja, 69 a moderada y 97 a muy alta vulnerabilidad.

Este algoritmo reveló que la variable más informativa para la clasificación fue el índice de atenuación del suelo seguida por la litología del acuífero, la densidad de dolinas y la pendiente topográfica.

Tabla 4 - Matriz de confusión.

	CLASES DE VULNERABILIDAD				
	MUY BAJA	**BAJA**	**MODERADA**	**ALTA**	**MUY ALTA**
MUY BAJA	46 573	0	0	9	0
BAJA	0	42 099	0	20	0
MODERADA	0	0	247 575	75	0
ALTA	13	7	69	66 913	97
MUY ALTA	0	0	0	102	172 256

En la figura 3 se presenta el mapa de vulnerabilidad obtenido con la aplicación de las técnicas de minería de datos.

Al comparar los mapas obtenidos aplicando RISK y técnicas de minería de datos, se aprecia que estas últimas logran mayor poder resolutivo al discriminar mejor las clases de vulnerabilidad presentes.

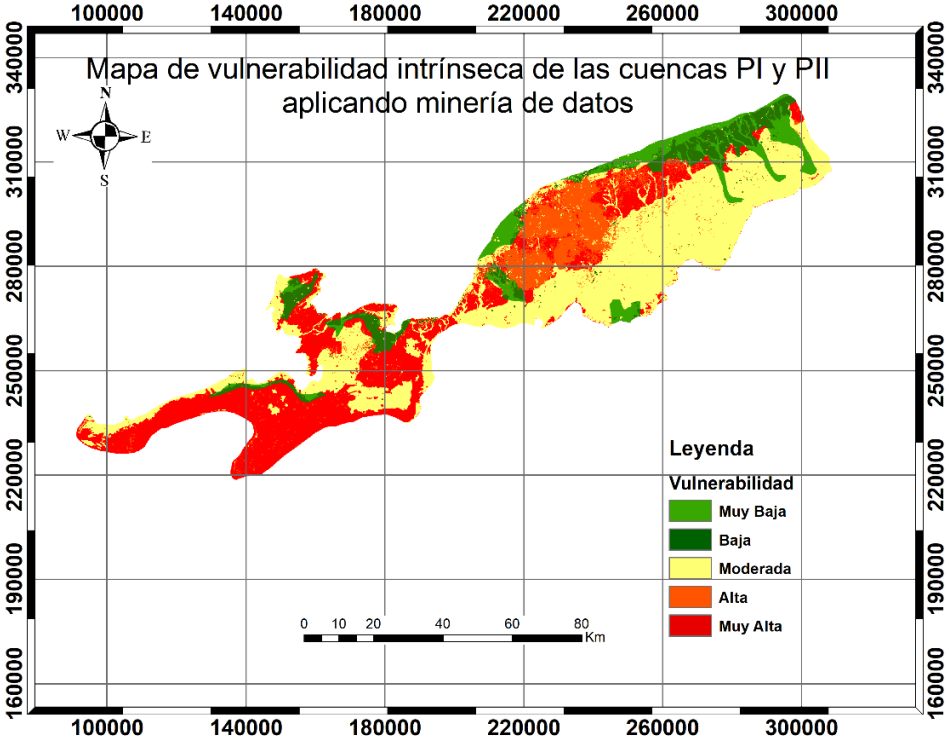

Figura 3 - Mapa de vulnerabilidad intrínseca a la contaminación de las aguas subterráneas de las cuencas PI y PII obtenido con la aplicación de los algoritmos k-medias y Árbol de Decisión, Escala 1:100 000.

Por ambos métodos se ratifica la elevada susceptibilidad a la contaminación de las aguas subterráneas en estas cuencas, teniendo en cuenta que las zonas clasificadas de alta y muy alta vulnerabilidad alcanzan el 43,7% y el 41,6% por el método RISK y minería de datos respectivamente, aunque este último discrimina mejor las zonas de mayor vulnerabilidad. A estas clases de vulnerabilidad pertenecen las principales formaciones acuíferas de la cuenca Costera Sur de Pinar del Río, desarrolladas en las formaciones geológicas Paso Real y Güines, e importantes formaciones acuíferas asociadas a la formación Vedado en la Península de Guanahacabibes. En estas regiones el suelo presenta poco desarrollo y se describen fundamentalmente rocas carbonatas fracturadas y karstificadas.

Predominan las zonas clasificadas de vulnerabilidad moderada, las que cubren el 55 % del área por el método RISK, y el 43% según las técnicas de clasificación aplicadas, y se localizan predominantemente en la zona central y sur de la Cuenca PII. En esta área predominan formaciones arcillosas, poco desarrollo kárstico, escaso espesor de suelo y relieve llano. Se describen la mayoría de los dépositos biogénicos, palustres, formaciones geológicas como Camacho, Loma Candela, Capdevila, Sigüanea y Güane.

Las técnicas de minería de datos, a diferencia del método RISK, logran definir zonas de baja y muy baja vulnerabilidad, las que se localizan hacia la zona norte de ambas cuencas, cubriendo el 15,3% del área total. La figura 4 permite comparar el % del área que ocupan las clases de vulnerabilidad obtenidas por el método RISK y por las técnicas de minería de datos. Puede apreciarse que el empleo de las técnicas de clasificación no supervisada y supervisada permitió un mayor poder resolutivo para cartografiar la vulnerabilidad natural de las cuencas, al brindar un modelo con cinco clases de susceptibilidad a la degradación del agua subterránea.

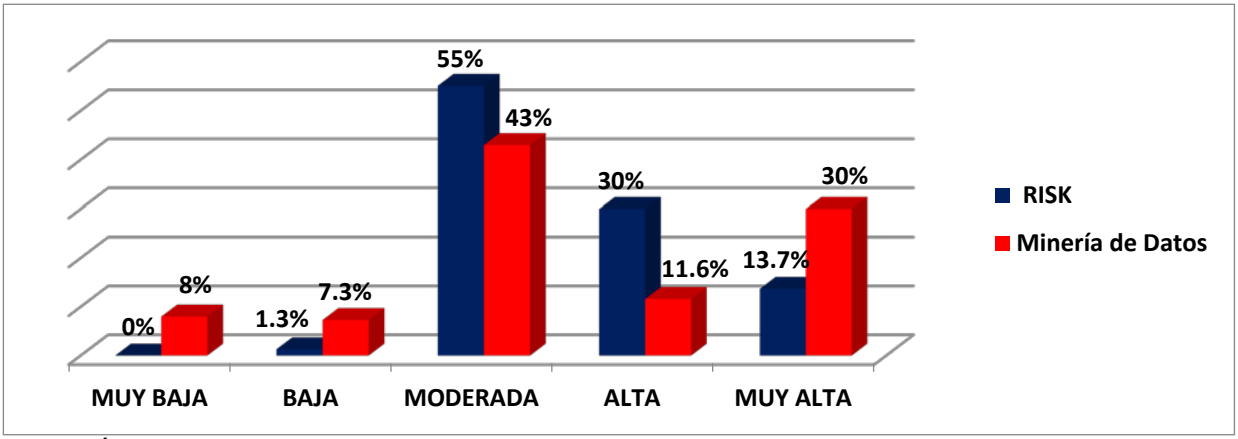

Figura 4 - Área que ocupan las clases de vulnerabilidad intrínseca a la contaminación del agua subterránea en las cuencas PI y PII aplicando el método RISK y técnicas de minería de datos.

Desde el punto de vista económico y ambiental estas investigaciones poseen enorme valor, teniendo en cuenta que la disponibilidad y calidad del agua representa uno de los principales desafíos para cualquier país. En particular, las aguas subterráneas de las cuencas Cuyaguateje y Costera Sur de Pinar del Río tienen gran importancia en el abasto de agua a la población y en el desarrollo agrícola e industrial de la provincia. Orientar políticas correctas en el ordenamiento territorial de estas cuencas para garantizar la protección de sus recursos hídricos subterráneos resulta imprescindible, y para ello, evaluar la vulnerabilidad natural de los acuíferos representa el primer paso. El costo de estas investigaciones no es elevado si se tiene en cuenta 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

La aplicación de las técnicas de minería de datos permitió obtener un modelo de clasificación con cinco clústeres que logra discriminar zonas de muy alta, alta, moderada, baja y muy baja vulnerabilidad natural a la contaminación de las aguas subterráneas en las cuencas Cuyaguateje y Costera Sur de Pinar del Río, Cuba.

Este resultado manifiesta mayor poder resolutivo que el obtenido por el método RISK, con el que fueron identificadas cuatro clases de

vulnerabilidad. Se demuestra que las técnicas de clasificación estadística no supervisada y supervisada no presentan las desventajas de los métodos de superposición de índices ponderados comúnmente usados para evaluar la vulnerabilidad de acuíferos, y permiten una cartografía más objetiva y precisa de la vulnerabilidad del acuífero a la contaminación.

La investigación desarrollada contribuye a la gestión integrada y sostenible de los recursos hídricos y promueve la aplicación de la ciencia y la tecnología para desarrollar sistemas de alerta temprana que contribuyan a la protección y reducción de la contaminación del agua. Por su alta eficiencia y poder resolutivo se recomienda el empleo de las técnicas de clasificación estadística multivariada para evaluar la sensibilidad a la contaminación de las aguas subterráneas.

REFERENCIAS

ALLER, L.; LEHR, J.; PETTY, R.; BENNETT, T. **DRASTIC: a standardized system to evaluate groundwater pollution potential using hydrogeologic setting**. Environmental Protection Agency (EPA). USA. 1987. Disp. en https://nepis.epa.gov/Exe/ZyPDF.cgi/20007KU4.PDF?Dockey=20007KU4.PDF Consultado el 1/11/2021.

DEEPESH, M.; MADAN, K.; JHA, V.P.; SINGH, C. **Assessment and mapping of groundwater vulnerability to pollution: Current status and challenges**. Earth, 2018. Disp. en: https://doi:10.1016/j.earscirev.2018.08.009 Consultado el 4/11/2021.

DÖRFLIGER, N.; JAUFFRET, D.; LOUBIER, S. **Cartographie de la vulnérabilité des aquifères karstiques en Franche-Comté**. BRGM/RP-53576-FR., 140 p., 2004.

DÖRFLIGER, N.; JEANNIN, P.; ZWAHLEN, F. Water vulnerability assessment in karst environments: a new method of defining protection areas using a multi-attribute approach and GIS tools (EPIK method). **Environmental Geology**, v. 39, n. 2, p. 165-176, 1999.

FOSTER, S.; HIRATA, R.; GOMES, D.; D ELIA, M.; PARÍS, M. **Groundwater quality protection**, 2nd printing. Washington, D.C. The World Bank. ISBN 0-8213-4951-1. 2007.

GRUPO EMPRESARIAL GEOCUBA. **Modelo Digital de Elevaciones de la República de Cuba**, GEOCUBA, La Habana, 2010.

HAMDAN, H. & EMAD, L. K-means clustering algorithm applications in data mining and pattern recognition. **International Journal of Science and Research**, v. 6, n. 8, p. 1577-1584. 2017.

HERNÁNDEZ-ORALLO, J.; RAMIREZ, M.J.; FERRI, C. **Introducción a la Minería de datos**. Pearson Educación, 680 p. 2004.

IMRON, M.; HASANAH, U.; HUMAIDI, B. Analysis of Data Mining Using K-Means Clustering Algorithm for Product Grouping. **International Journal of Informatics and Information System**, v. 3, n. 1, p. 12 - 22, 2020.

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. Ministerio de la Agricultura, La Habana, 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**, v. 549, p. 27-37, 2017.

JEIHOUNI, M.; TOOMANIAN A.; MANSOURIAN, A. Decision Tree-Based Data Mining and Rule Induction for Identifying High Quality Groundwater Zones to Water Supply Management: a Novel Hybrid Use of Data Mining and GIS. **Water Resources Management.** v. 34, n. 9, p. 139-154, 2019.

MARTÍNEZ, A**. Aplicación de técnicas de minería de datos con software Weka**. Universidad de Salamanca, 2018. Disp. en: http://dx.doi.org/10.1007/s10115-003-0128-3 Consultado el 29/09/2019.

MEDINA, R, F. & GÓMEZ S.C. Funcionalidades de la minería de datos. **Revista Ingeniería y Región**. v. 10, n. 12. p. 31 - 40, 2014.

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**. Spatial Modeling in GIS and R for Earth and Environmental Sciences, 2019. Disp. en: https://doi.org/10.1016.B978-0-12-815226-3.00025-9 Consultado el 4/10/2021.

MOYA, R. **Selección del número óptimo de clusters**. 2016. Disp. en https://jrrob.com/seleccion-del-numero-optimo-clusters. Consultado el 4 de octubre 2021.

NADIRI, A.; GHAREKHANI, M.; KHATIBI, R. Mapping aquifer vulnerability indices using artificial intelligence - running multiple frameworks (AIMF) with supervised and unsupervised learning. **Water Resources Management**, v. 32, p. 3023 - 3040, 2018.

STEMPVOORT, D.V.; EWERT, L. WASSENAAR, L. Aquifer Vulnerability Index: A GIS – compatible Method for groundwater vulnerability mapping. **Canadian Water Resources Journal**. v. 18, n. 1., p. 25 - 37, 1993.

TZIRITIS, E.; PISINARAS, V.; PANAGOPOULOS, A.; ARAMPATZIS, G**. RIVA: a new proposed method for assesing intrinsic groundwater vulnerability. Environmental Science and Pollution Research. 2020**. Disp. en https://doi.org/10.1007/s11356-020-10872-3. Consultado el 15/11/2021.

UNESCO. **Informe Mundial de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos 2020**: Agua y Cambio Climático, París, UNESCO, 2020. Disp. en: https://es.unesco.org/themes/water-security/wwap/wwdr/2020 Consultado el 15/11/2021.

VALCARCE, R. & SOLIS, E. Evaluación de la vulnerabilidad intrínseca a la contaminación de las aguas subterráneas en las cuencas Cuyaguateje y Costera Sur de Pinar del Río, Cuba. **Geociências**, v. 40, n. 3, p. 751 - 761, 2021.

VALCARCE, R.M., SUÁREZ, O.; RODRIGUEZ, W.; VEGA, M. Aplicación de la minería de datos a la evaluación de la vulnerabilidad de acuíferos. **Revista Cubana de Ciencias Informáticas**, v. 15, n. 2, p. 21 - 43, 2021.

VARGAS, M.S. **Propuesta metodológica para la evaluación de la vulnerabilidad intrínseca de los acuíferos a la contaminación**. Ministerio de Ambiente, Bogotá, 45 p., 2010.

VÍAS, J.M.; ANDREO, B.; PERLES, M. J.; CARRASCO, F.; VADILLO, I., JIMÉNEZ, P. Proposed method for groundwater vulnerability mapping in carbonate (karstic) aquifers: the COP method. Application in two pilot sites in Southern Spain". **Hydrogeology Journal**, v. 14, n. 2, p. 912-925, 2006.

YOO, K.; SUDHEER, K.; JAE. J.; KYUNGJOO, O.; JOONHONG, P. Decision tree – based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity. **Journal of Cleaner Production**, v. 122, p. 277 – 286, 2016.

Submetido em 3 de fevereiro de 2022
Aceito para publicação em 2 de junho de 2022