A new approach to pollution vulnerability assessment in aquifers using K-means analysis

Marisela Uzcategui-Salazar[a,b*], Javier Lillo Ramos[c,d]

[a] International Doctoral School, University of Rey Juan Carlos, 29833 Móstoles, Madrid, Spain.
[b] Geological Engineering School. TERRA research team. University of Los Andes. 5101 Mérida, Venezuela.
[c] Global Earth Change and Enviromental Geology Research Group, Department of Biology, Geology, Physics and Inorganic Chemistry. University of Rey Juan Carlos, 29833 Móstoles, Madrid, Spain.
[d] IMDEA Water Institute, Av. Punto Com, 2, 28805 Alcalá de Henares, Madrid, Spain

*Corresponding author.

E-mail: mariselauzcateguis@gmail.com
University of Rey Juan Carlos, Departmental II, office 256. Tulipán Street, s/n, 29833 Móstoles, Madrid, Spain
ORCID: 0000-0002-2894-5925

Abstract

The most used methods to evaluate the vulnerability to contamination of aquifers are based on overlay index maps, such us DRASTIC, GOD and AVI. These methods assign weighting and rating values to hydrogeological characteristics, introducing subjectivity in the evaluation. In this research, a new methodology is proposed to eliminate some of that subjectivity. The methodology evaluates the vulnerability to contamination of a detrital aquifer using K-means cluster analysis with a new set of parameters. The set is composed of some parameters extracted from these methods, as well as other new ones that have a significant influence on the movement of contaminants. Application of the Principal Components Analysis (PCA) technique before using K-means cluster allowed the selection of the most relevant parameters. In order to validate the methodology, this was applied to a detrital aquifer located at central Spain (the so-called "Aluviales Jarama-Tajuña" aquifer) with a significant agricultural development. To compare the traditional methods of vulnerability assessment with the K-means cluster, nitrate concentration was used as a pollution indicator. Thus, 23 groundwater quality samples were used to correlate (Spearman´s correlation coefficient) the vulnerability values with nitrate concentration to validate the most suitable method. The results showed that GOD and AVI were not appropriate methods to evaluate the vulnerability of the aquifer, because they have negative or very low correlation with nitrate concentration (-0.5 and 0.01 respectively). This is due to the use of very few variables that do not represent relevant features for the vulnerability assessment. Alternatively, DRASTIC and K-means cluster analysis obtained higher Spearman´s correlation coefficients (0.34 and 0.48 respectively). The relevant features selected by PCA analysis to use in the K-means low dimensional analysis were depth of groundwater (D), net recharge (R), and land use (L). The new proposed method grouped data in three clusters that represent low vulnerability (35.9 % of the study area), moderate (41.4%) and high vulnerability (22.7%). K-means increases the Spearman´s correlation by 14 % with respect to the most approximate conventional method (DRASTIC). Therefore, the results obtained confirm the advantage of joint application of PCA and K-Means analysis, which represents a novel approach for the assessment of groundwater vulnerability in detrital aquifers.

Key words: aquifer vulnerability assessment, groundwater quality, overlay index maps, K-means cluster.

1. Introduction

Aquifers represent the most important source of water in arid and semi-arid zones. Water in aquifers has a natural protection against evapotranspiration losses and inputs of anthropogenic agents from human land uses. However, the growing demand for water due to increasing industrial and agricultural activities puts aquifers at high risk of contamination. Rational management and prevention are the most appropriate strategies for groundwater protection (Saatsaz et al. 2013; Kadkhodaie et al. 2019). Vulnerability assessment is one of the most widely used tools to prevent aquifer pollution, since it allows the identification of the areas most susceptible to contamination taking into account their own hydrogeological characteristics (Babiker et al. 2005). Thus, intrinsic groundwater vulnerability depends on the natural conditions of aquifer, i.e. those hydrological and geological characteristics that affect and control the movement of groundwater (Aller et al. 1987).

There are different methods to assess the intrinsic vulnerability: simulation methods, statistical models and overlay index methods (Huan et al. 2012). The overlay index methods are widely used because of their simple approach. In this research, three well-known methods are considered: the GOD index (Foster 1987), the AVI index (Stempvoort et al. 1993) and the DRASTIC index (Aller et al. 1987). The former is the most common and established method (Rupert 2001; Panagopoulos et al. 2006; Huan et al. 2012; Kazakis and Voudouris 2015; Jafari and Nikoo 2016; Yang et al. 2017; Barzegar et al. 2020).

All of these overlay index methods are somewhat subjective, as they assign numerical weighting and rating values to the properties according to their importance and hydrogeological features of aquifer. However, they do not take into account influence of regional and local conditions (e.g. land uses among others) that can affect weighting and rating values which is a major disadvantage (Javadi et al. 2011; Hao et al. 2017). To improve the vulnerability assessment, researchers have modified the original methods by changing the weighting and rating values through statistical methods or by adding/ removing variables (Rupert 2001; Panagopoulos et al. 2006; Javadi et al. 2011; Mendoza 2012; Huan et al. 2012; Hao et al. 2017; Kadkhodaie et al. 2019).

Data mining techniques are being used in groundwater studies related to prediction of water quality, definition of hydrogeological models, aquifer assessment, and transport of contaminants (Pathak and Hiratsuka 2011; Conti and Gibert 2014; Yoo et al. 2016; Stumpp et al. 2016; Marín Celestino et al. 2018; Ouedraogo et al. 2019; Tahmasebi et al. 2020). The capability of data mining techniques to process hidden and big datasets allows to identify patterns that can be used to predict hydrogeological behavior of aquifers, which in turn improves the design of groundwater protection programs (Conti and Gibert 2014; Tahmasebi et al. 2020).

A useful data mining technique is the K-means clustering, an unsupervised pattern recognition method (Javadi and Hashemy 2016; Javadi et al. 2017) that allows information to be classified into different groups or clusters. It is an iterative algorithm that assigns individual points to a cluster such that the sum of the squared Euclidean distance between the data points and the centroid of the cluster is at the minimum (Dabbura 2020). One of the difficulties of the K-means method is to define the number of clusters, as it must be established at the beginning of the iterative process. Charrad et al., (2014) propose to estimate the optimal number of clusters through the calculation of various indices. Some of the variables or features help to identify clusters while others add noise, making clustering more difficult (Dash and Koot 2009). For this reason, is necessary to identify the relevant features in order to select the variables that have the greatest influence on the process (Song et al. 2010). The identification of critical variables generates a better understanding of the aquifer system and its interaction with causal indicators of potential impacts (Malmir et al. 2021).

Aquifers in areas where human activities are carried out that may generate pollution are, in principle, aquifers susceptible to contamination by the presence of potential pollutants. This is the case of aquifers located in areas of high urban, agricultural, livestock or industrial development, where there is a production of wastes or residues with pollutants that can be easily transported to the aquifer. In these areas, high concentrations of pollutants can be observed in chemical, physical or bacteriological analyses of groundwater, allowing the evolution of these concentrations to be monitored and, in some cases, the origin of the pollution to be identified.

The detrital Jarama-Tajuña aquifer, located in central Spain, is an important source of water supply for the agricultural and industrial activities in the region of Madrid. This aquifer was selected for this work because the use of agricultural products and wastewater for irrigation has significantly increased the concentration of nitrate (NO_3^-) in the groundwater (Arauzo et al. 2008; Mostaza-Colado et al. 2018; Mostaza 2019). Agricultural activities combined with the excessive application of fertilizers are a potential source of nitrate contamination of groundwater. Thus, nitrate is considered an indicator of groundwater quality (Kazakis and Voudouris 2015).

The aim of this research is to develop a new methodology for the assessment of the pollution vulnerability of a detrital aquifer in a simple approach using few variables. This is intended to facilitate data collection which, in some cases, could make the use of classical methodologies unfeasible. The analysis of these variables will be performed using clustering algorithms to eliminate the subjectivity associated with assigning weighting and rating values, allowing groups to be defined based on similarities between the data. The variables are draw from classical methodologies (DRASTIC, GOD and AVI) and from modified methodologies that have worked well in the cases studied (Rupert 2001; Babiker et al. 2005; Panagopoulos et al. 2006; Denny et al. 2006; Javadi et al. 2011; Mendoza 2012; Huan et al. 2012; Jafari and Nikoo 2016; Hao et al. 2017; Jang et al. 2017; Barzegar et al. 2020; Aslam et al. 2020) . However, the subjective loading of the used methods means the results that are not adjusted to the local reality of an aquifer. Therefore, it is important to incorporate those variables that control the mobility of pollutants and that are not considered in the methodologies mentioned above. On the other hand, the variables incorporated may be excessive and redundant with respect each other, which generates a bias in the results. Thus, the selection of the set of variables initially considered can be refined by applying principal component analysis (PCA). This allows the number of variables to be reduced to a minimum number of new variables and the old variables to be used as representatively as possible to eliminate redundancy and retain relevant information (Song et al. 2010). The K-means clustering method is used to make a more adjusted analysis of the dataset and to demonstrate how this data mining technique can be a very useful tool for the evaluation of the most relevant variables that make the aquifer vulnerable to contamination and thus relate them to the current state of the groundwater quality. The proposed methodology based on the joint application of PCA and K-means analysis provides a novel approach for the assessment of the vulnerability of aquifers to contamination.

2. Study area

The “Aluviales Jarama-Tajuña” aquifer is located in the southeast of Madrid (Spain) (Fig. 1). The area (133 km^2) is situated approximately between 3º38 and 3º25 W and between 40º7 and 43º21 N. The Jarama River flows north to south along the aquifer area, being the river and the aquifer hydraulically connected (Mostaza-Colado et al. 2018). The climate is Mediterranean temperate-continental, close to semi-arid conditions during summer. The average annual rainfall is 350 mm, estimated by Thiessen polygons method from three weather stations (“Center: Finca Experimental”, “Arganda” and “San Martín de la Vega”, Fig. 1) for the 2008-2018 period (data from the Spanish Agroclimatic Information System for Irrigation,Sistema de Información Agroclimática y de Regadíos -SIAR-, 2019).

The “Aluviales Jarama-Tajuña” is a shallow unconfined aquifer formed by Quaternary alluvial deposits of the Jarama River (Carreño Conde et al. 2014), consisting mainly of gravels and sands interbedded with layers of clay and silt layers (Bardají et al. 1990). The basement of the aquifer and its sidewalls are formed by Tertiary sedimentary units, which consist mainly of gypsum with intercalated beds of carbonate rocks and mudstones (Fig. 2) (Calvo et al. 1989; Carreño Conde et al. 2014). The aquifer has an average thickness of 10.97m and is characterized by a storage coefficient and a transmissivity of 0.07 and 700 m^2/d, respectively (Bardají et al. 1990). The water level in the aquifer varies from 2 m to 26 m depth. The highest values are found in the central-eastern part of the study area, mainly as a result of groundwater extraction in wells.

The study area has an important agricultural development, with artificial irrigation being one of the main sources of water for crops. The continued use of agricultural products in the area (fertilizers, pesticides, etc.) is significantly increasing the risk of contamination. Periodic monitoring of groundwater quality is annually carried out by the Hydrographic Confederation of El Tajo (CHET), showing that the concentration of nitrate in some wells exceeds the acceptable level defined at 50 mg/L (Arauzo et al. 2008; BOE 1996; Mostaza 2019).

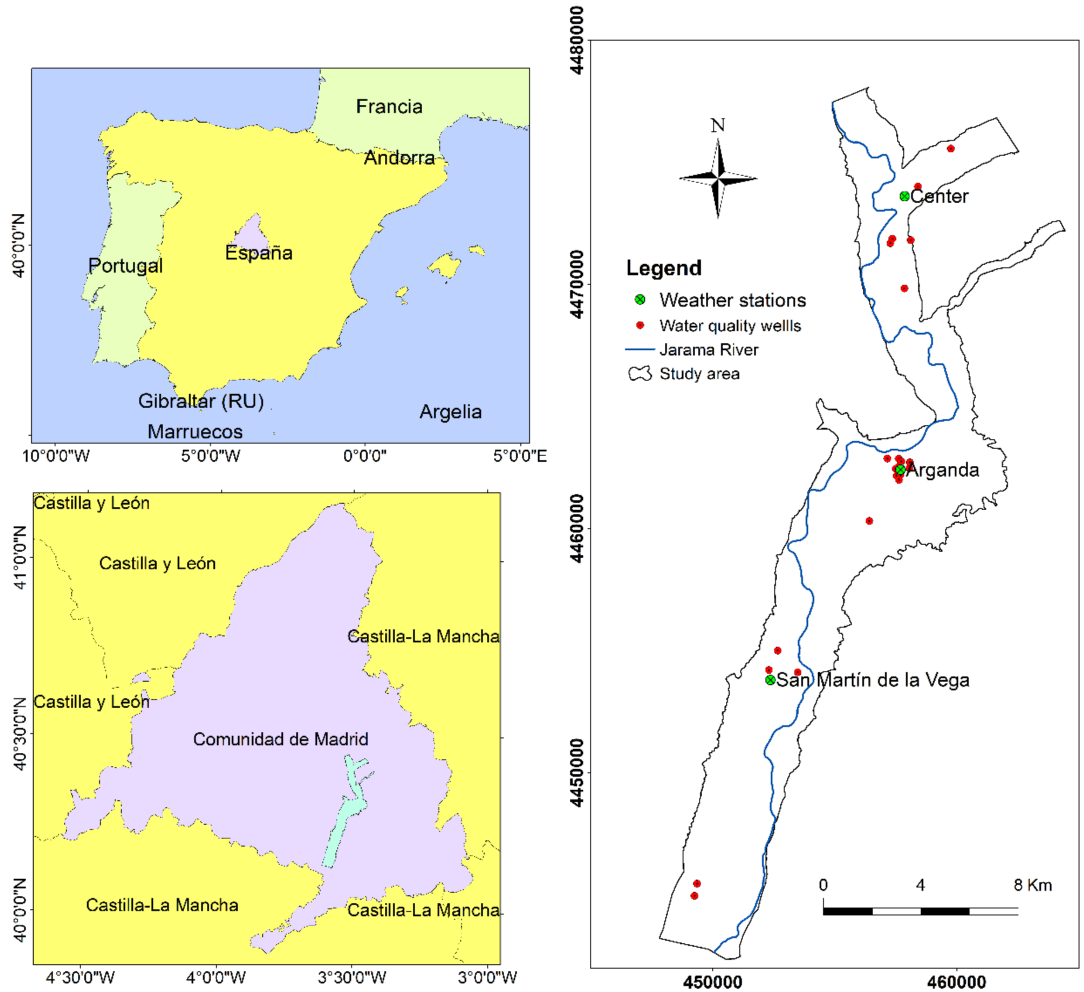

Fig. 1 Location map of the "Alluviales Jarama-Tajuña" aquifer showing the locations of weather stations and water quality monitoring wells (coordinates in UTM Zone 30N).

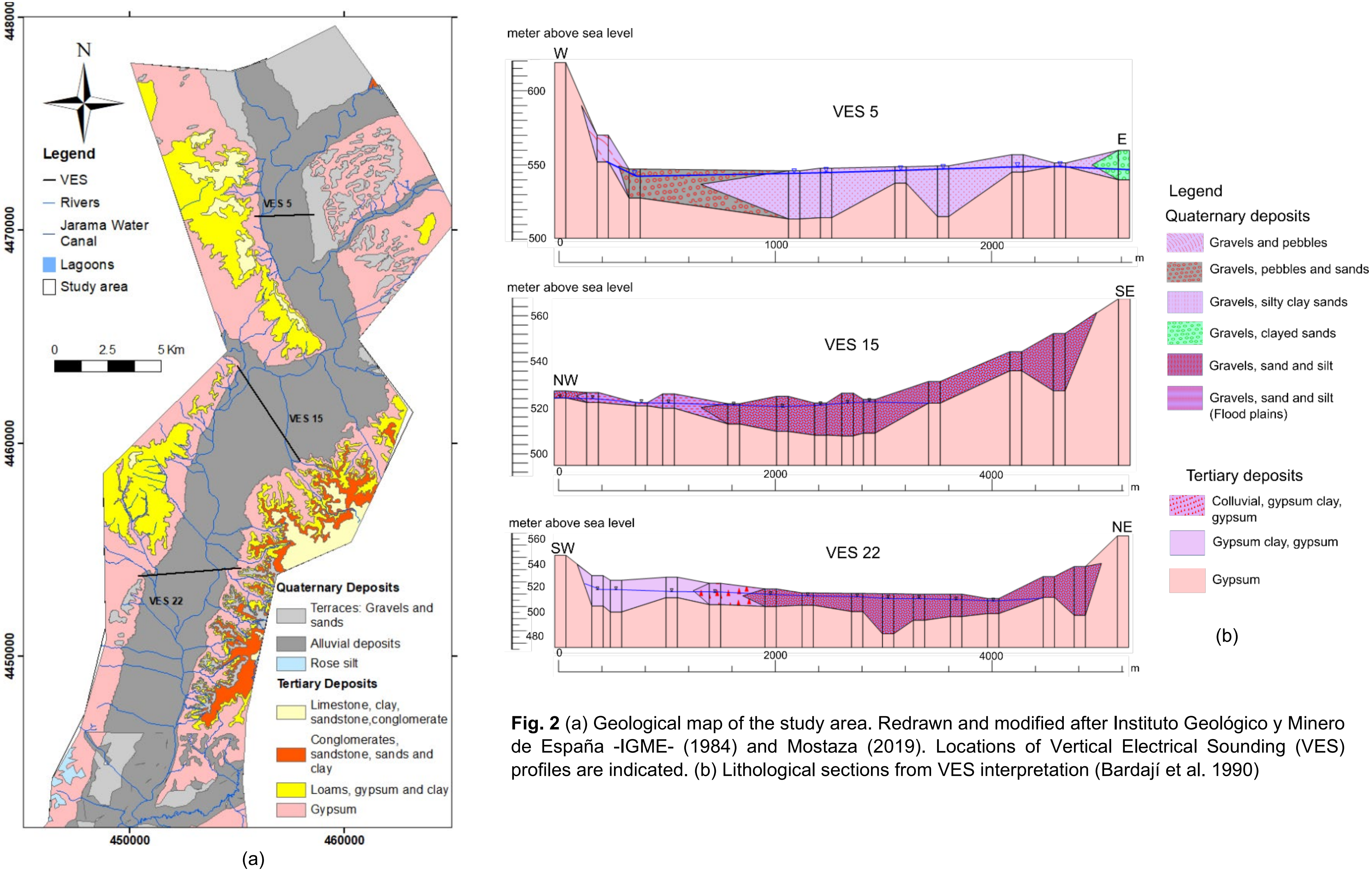

Fig. 2 (a) Geological map of the study area. Redrawn and modified after Instituto Geológico y Minero de España -IGME- (1984) and Mostaza (2019). Locations of Vertical Electrical Sounding (VES) profiles are indicated. (b) Lithological sections from VES interpretation (Bardají et al. 1990)

3. Materials and methods

3.1 Data set collection

The data considered in this work is displayed in Table 1.The data were storage as a geographic database in ArcGIS v10.2.1. The whole study area (133 Km^2) was divided in 5842 pixels with a cell size 150mx150m, in order to obtain a big data set to evaluate the different variables in each point.

The assessment was performed in the following six stages:

- Intrinsic vulnerability mapping using overlay index methods: DRASTIC, GOD and AVI
- Intrinsic vulnerability analysis using cluster analysis (K-means algorithm) with a high dimensionally data set.
- Selection of hydrogeological features by Principal Analysis Component (PCA), to reduce the dimension of the cluster analysis.
- Intrinsic vulnerability mapping using cluster analysis (K-means algorithm) with a low dimensionally data set.
- Validation of the vulnerability map. Comparison of the effectiveness of each method by statistical correlation between a quality water indicator (nitrate concentration) and the vulnerability value.

A max-min normalization method (Salazar and Del Castillo 2018) of the vulnerability index values obtained from each applied method was performed to standardize the ranges of values (0-1) (Equation 1).

$$Normalized\ Vulnerability\ index = \frac{(Vx - Vmin)}{(Vmax - Vmin)} \qquad \text{Equation 1}$$

Where *Vx* is the vulnerability index value evaluated in the *x* point, and *Vmin*, *Vmax* are the obtained minimum and maximum vulnerability index values of the range, respectively.

From the normalized vulnerability index values were defined four vulnerability classes: Low (Vulnerability index ≤ 0.25), Moderate (0.25 <Vulnerability index ≤ 0.50), high (0.50 <Vulnerability index ≤ 0.75) and very high (Vulnerability index > 0.75).

Table 1 Features and sources of the data considered in this research

Type	Description	Source	Scale	Time periods	Number of data	Application in the method
Geological data	Geological map	(IGME 1984)	1:50.000	-	Sheets: 559, 560, 582, 583, 605	Lithological sections
	VES (Vertical Electrical Sounding)	(Bardají et al. 1990)			22 Vertical Electric Soundings	
Hydrogeological data	Water table	(Mostaza 2019)	-	2012, 2013, 2015, 2016 and 2017	58 monitoring wells	Depth of water table, unsaturated thickness
	Pumping test	(Bardají et al. 1990)		-	11 Pumping test	Transmissivity and Hydraulic conductivity of the aquifer
	Empiric hydraulic conductivity	Smith and Weathcraft 1993; Domenico and Schwartz 1998; Sanders 1998; Coduto 1999; Fetter 2001				Hydraulic conductivity of unsaturated zone
Topography, land use, soil data	Digital model elevation	(IGN 2019)	1:25.000	-	Sheets: 559, 560, 582, 583, 605	Topography, Slope
	Land use map	(IGN 2018)	1:100.000			Land use
	Soil map	(IGN 2008)	1.3.000.000			Soil type
Climate data	Rainfall and Temperature	(SIAR 2019)	-	2008-2018	Three weather station “Center: Finca experimental”, “Arganda” and “San Martín de la Vega”	Natural Recharge
Agricultural data	Agricultural demand units (UDA)	(CHT 2015a)	-	2015-2021		Artificial Recharge
Water quality data	(CHT 2019; Mostaza 2019)	(CHT 2019; Mostaza 2019)		2015-2017	23 monitoring wells	Nitrate concentration

3.2 Intrinsic vulnerability assessment by overlay index methods

The DRASTIC, GOD and AVI methods were used to assess the vulnerability to aquifer contamination by overlay index maps.

3.2.1 Vulnerability analysis by DRASTIC method

The *DRASTIC* method assumes that contaminants are introduced from the surface and that they have the same mobility as water (Aller et al. 1987).

The method uses seven parameters, called “factors”: Depth to the water table (*D*), net recharge (*R*), aquifer media (*A*), soil type (*S*), topography (*T*), impact of the vadose zone (*I*) and hydraulic conductivity (*C*). Depending on the importance of each factor considered for assessment, this method assigns a weighting coefficient (W) from 1 to 5. In adittion, each factor is assigned a rating value (R) from 1 to 10 , depending on its expression. Thus, the vulnerability is calculated by the following equation (Aller et al. 1987):

$$DRASTIC\ Index\ (DI) = D_R D_w + R_R R_w + A_R A_w + S_R S_w + T_R T_w + I_R I_w + C_R C_w \qquad \text{Equation 2}$$

Where $D_R, R_R, A_R, S_R, T_R, I_R, C_R$ are the rating values and $D_w, R_w, A_w, S_w, T_w, I_w, C_w$ are the weighting coefficients (Table 2). Higher values of DRASTIC index (*DI*) represent higher vulnerability than lower values. In this work, the rating values were selected according to specific information of the study area (Table 2).

Depth to water table (*D*) was determined by interpolating depth values from 58 wells recorded in five years (2012, 2013, 2015, 2016, 2017) (Mostaza 2019). The kriging method was used to interpolate the values with an exponential variogram.

Net recharge (R) was calculated as the sum of natural and artificial recharge (Fig. S1). Natural recharge was obtained from a water balance for the three Thiessen polygons defined in the study area (See supplemenrary information, Fig. S2) using the following equation (Custodio and Llamas 2002) (for a closed hydrogeological basin):

$$R = P - ETR - ESC \qquad \text{Equation 3}$$

Where R is natural recharge, P is monthly precipitation, ETR the real evapotranspiration and ESC is the surface runoff (See Supplementary information for details).

Artificial recharge was calculated from irrigation return flows in the agricultural demand units (UDA) (See Supplementary information ,Table S2, Fig. S3). Information on land use (IGN 2018) and agricultural demand (CHT 2015b) was necessary to determine the irrigation zones in the study area. Artificial recharge was estimated by intersecting irrigation zones and agricultural demand units.

The lithology of the aquifer (A) as well as the impact of vadose zone (I) were obtained by integrating the information from 22 vertical electrical soundings (VES) (Bardají et al. 1990) (See Supplementary information, Fig. S4) and the geological map of the area (IGME 1984) (Fig. 2a). Lithological information from three drill cores allowed to calibrate the lithology in the VES (See Supplementary information, Fig. S5), thus obtaining 22 lithological sections by correlation (Fig. 2b) shows three representative of them). This information was used for a complete lithological interpretation of the aquifer and the vadose zone. Subsequently, numerical values were assigned to the lithological units according to their permeability properties. Different rating values were considered for each type of lithology as shown in Table 2. The aquifer media (A) and impact of the vadose zone (I) maps were produced by kriging interpolation with an exponential variogram.

The soil type (S) factor was obtained using the soil map and soil texture (Monturiol and Alcalá del Olmo 1990; IGN 2008; United States Department of Agriculture (USDA) 2017). Soil type is a descriptive variable, and numerical values were assigned according to the DRASTIC method (Table 2).

The slope topography (T) was calculated from the digital elevation model (IGN 2019), by using the slope tool in ArcGIS.

The hydraulic conductivity (*C*) was determined by dividing the transmissivity values by the aquifer thickness values. The transmissivity values were obtained from data from 11 aquifer tests previously carried out by Bardají et al. (1990). The data were interpolated for the whole study area by the Inverse Distance Weighted (IDW) method, with a distance of 500 m and a minimum number of points equal to one. The aquifer thickness data were obtained from the lithological sections (Fig. 2b) and depth data (CHT 2019; Mostaza 2019). The aquifer thickness represents the saturated material from the groundwater level to the basement of the aquifer (gypsum).The derivative map was made by interpolation of the data using the kriging method (exponential variogram).

All the parameters of the thematic maps were reclassified by defining classes and assigning rating values from 1 to 10, as shown in Table 2.

Equation 2 was used to obtain the DRASTIC vulnerability index (DI) for the study area. The vulnerability map was produced using the raster calculation tool in ArcGIS. Finally, a normalization of the DI values made it possible to define the vulnerability classes.

Table 2. Weighting and rating values of the DRASTIC parameters in the study area (Adapted from Aller et al., 1987)

Drastic Parameters	Range	Rating values (R)	Weighting Values (w)
Water level **D** (m)	1.5- 4.6	9	5
	4.6 – 9.1	7	
	9.1 -15.2	5	
	15.2 – 22.9	3	
	22.9 – 30.5	2	
Net Recharge **R** (mm)	0 -50	1	4
	50 - 103	3	
	103 - 178	6	
	178 - 254	8	
	>254	9	
Aquifer media **A**	Sand and Gravel 4 - 9	Colluvial, gypsum clay 5	3
		Gravel, sand, sandy clay 6	
		Gravel and silty or clayey sand 7	
		Gravel, sand and silt 8	
Soil type **S**	Loam	5	5
	Silty loam	4	
Topography **T** (%)	0 – 2	10	3
	2 – 6	9	
	6 – 12	5	
	12 – 18	3	
	>18	1	
Impact of vadose zone **I**	Gavel, sand and silt	8	4
	Gravel and silty or clayey sand	7	
	Gravel, sand, sandy clay	6	
	Gypsum clay, gypsum, gravel, sand and clay	5	
Hydraulic conductivity **C** (m/d)	0.04 – 4.08	1	2
	4.08 – 12.22	2	
	12.22 – 28.55	4	
	28.55 – 40.75	6	
	40.75 – 81.49	8	
	>81.49	10	

3.2.2 GOD method

The *GOD* method is based on three parameters or “factors” to assess the vulnerability of aquifer: the groundwater occurrence (*G*), the overall lithology of aquifer (*O*) and the water table Depth (*D*) (Foster 1987).

The vulnerability index is calculated by the equation 4, where each factor has a rating value from 0 to 1 (Foster 1987) (Table 3):

$$Vulnerability\ Index = G * O * D \qquad \text{Equation 4}$$

The vulnerability is considered zero when the GOD index is less than 0.1. An index of 0.1 to 0.3 represents low vulnerability. An index of 0.3 to 0.5 represents a moderate vulnerability, and an index of 0.5 to 0.7 refers high vulnerability, and above 0.7 is related to very high vulnerability (Foster and Hirata 1991).

The groundwater occurrence parameter (*G*) defines the type of aquifer. This parameter has been obtained from the lithological sections and the depth to groundwater data, as well as from other hydraulic data such as the storage coefficient (obtained from the pumping test carried out by Bardají et al. (1990)).

The overall lithology of aquifer (*O*) factor is equivalent to the impact of vadose zone factor in *DRASTIC*, but the ratings assigned to each lithology type in *GOD* are different. Similarly, the water table depth (*D*) was obtained from the previous map in DRASTIC method, but new rating values were considered (Table 3).

The vulnerability index (*GOD*) was calculated for the entire study area using raster calculator tools in ArcGIS. The vulnerability index values were normalized (equation 1) and then classified to define the classes in the vulnerability map.

Table 3. GOD ranges and rating values for three parameters in the study area (Based on Foster, 1987)

GOD Parameters	Range	Rating values
Groundwater occurrence **G**	Unconfined aquifer	1
Overall lithology of aquifer **O**	Alluvial silt, clay, marl, fine limestone	0.5
	Alluvial sand and gravels	0.6
	Wind sand, sandstone	0.7
	Colluvial gravel	0.8
Depth of water **D** (m)	20 - 50	0.6
	10 - 20	0.7
	5 - 10	0.8
	2 - 5	0.9

3.2.3 Aquifer vulnerability index. AVI

AVI is a simplified method to assess the aquifer vulnerability by considering a single parameter, the hydraulic resistance (*C*). This parameter is an estimate of the travel time of contaminants through the unsaturated zone (vertical direction from the ground surface to the groundwater level), measured in years (Stempvoort et al. 1993). To apply the methodology, it is necessary to know the thickness of the unsaturated zone and its hydraulic conductivity.

The hydraulic resistance is calculated using the following equation (Stempvoort et al. 1993):

$$C = \sum_i \frac{di}{Kvi} \qquad \text{Equation 5}$$

Where *i* is the number of layers, *di* is the thickness of each unsaturated layer and *Kvi* is the vertical hydraulic conductivity of each unsaturated layer.

There is an inverse relationship between hydraulic resistance and pollution vulnerability class, as hydraulic resistance controls the travel time of contaminants in the unsaturated zone.

The unsaturated thickness parameter (*d*) was obtained from the VES lithological sections located at the study area (Bardají et al., 1990) (Fig. 2b).

The vertical hydraulic conductivity (*Kv*) of the unsaturated zone was estimated from the geological map and the lithological sections (Fig. 2), assigning empirical values from several authors (Smith and Weathcraft 1993; Domenico and Schwartz 1998; Sanders 1998; Coduto 1999; Fetter 2001). The empirical values obtained correspond to horizontal values of hydraulic conductivity (*Kh*). For this reason, it was necessary to consider the effects of compaction and consolidation that reduce the soil void ratio in the unsaturated zone. For the vertical hydraulic conductivity (*Kv*), a ratio of $Kh/Kv = 10$ was assumed due to the lack of information on grain size, which is commonly used for alluvial aquifers (Neilson-Welch and Allen 2007).

The hydraulic resistance map was obtained using equation 5. Normalization and classification were applied to the obtained vulnerability index values to define the vulnerability index classes. It is important to note that the classification of the AVI vulnerability index map is inverse to that in other methods. In this case, high normalized ranges represent low vulnerability and low normalized ranges correspond to high vulnerability.

3.3 Intrinsic vulnerability assessment by K-means cluster analysis

Clustering analysis allows grouping objects according to their similarities (Rahmani et al. 2019; Javadi et al. 2020). The similarity between two objects is the distance between them (Euclidean distance is commonly considered) (Rahmani et al. 2019;

Dabbura 2020). Unsupervised methods, as K-means clustering, do not use predefined classes to predict classification, which gives greater objectivity in the results. In addition, the independence of weighting and rating values in the evaluation of parameters is an advantage of using clustering. This assumes that the data of all parameters explain the vulnerability of the aquifer by themselves. This iterative process is achieved by the following procedure:

- Creation of "*n*" x "*d*" matrix dataset, where *n* is the number of data points in a *d*-dimensional feature space (in this case, all parameters chosen to assess the vulnerability).
- Selection of the number of clusters "K". The optimal number of clusters was determined using the R package NbClust, which provided 26 indices (Table S3). The best number of clusters was obtained using the majority rule.
- Each point was randomly assigned to the closest cluster The Euclidean distance is used to find the distance of each point to a temporaral cluster.

Recalculation of the temporal clusters with new centroids based to the nearest points located in them. This is achieved by minimizing the sum of squared errors of the distance "*A*" between each point to the centroid of each cluster, using the following equation (Dabbura 2020):

$$A = \ min \sum_{i=1}^{k} \sum_{x \in ki} ||x_k - mi||^2 \qquad \text{Equation 6}$$

Where x_k = *(x1, x2, x3,……..xn)* are the data belonging to the k_i cluster; and m_i is the centroid of the cluster k_i :

$$m_i = \frac{\sum_{k=1}^{Ni} x_k}{N_i} * x_k \in k_i \qquad \text{Equation 7}$$

Where N_i is the number of data objects in the cluster *i.*

The procedure ends when no points are reallocated from one cluster to another or when a predefined number of iterations is reached (Dabbura 2020).

The selection of the parameters to be used in K-means cluster analysis on a high dimensional dataset was carefully studied to consider non-redundant variables in order to avoid noise to create clusters. In addition, it was important to take into account parameters that influenced in facilitating the transport of pollutants (Rahmani et al. 2019) and selection of parameters associated with water resources systems should be based on indicators and their causal relationships (Malmir et al. 2021).

The parameters considered were extracted from different methods as DRASTIC (Aller et al. 1987), AVI (Stempvoort et al. 1993), GOD (Foster 1987) and others parameters by modified methods. The six parameters considered are described below.

Depth of water table (*D*), which considered the unsaturated thickness and the hydraulic head of the aquifer (Aller et al. 1987; Debernardi et al. 2008). Aquifer recharge (net recharge, *R*), which considered soil conditions, cover vegetation and land slope (See Supplementary information), (Aller et al. 1987; Kazakis and Voudouris 2015). Land use (*L*), which considered different activities developed in the area that have influence on the vulnerability to pollution, as well as, the irrigation network (Jarama irrigation water channel) (Arezooman et al. 2015; Kazakis and Voudouris 2015; Asadi et al. 2017; Hao et al. 2017). Land use is a qualitative parameter, for this reason, it was assigned numerical values from 1 to 5, according to tendency to contamination, i.e. the highest probability of contamination has a value of 5 and the lowest probability of contamination has a value of 1. (Table 4). Aquifer hydraulic conductivity (*C*), which considers aquifer media and permeability (Aller et al. 1987; Hao et al. 2017). Hydraulic conductivity of the unsaturated zone (*Kv*), which considered the vertical permeability and the impact of vadose zone (Aller et al. 1987; Foster 1987; Stempvoort et al. 1993).

Aquifer thickness (*Th*), which considered the dilution phenomena of the contaminant within the aquifer (Debernardi et al. 2008; Hao et al. 2017).

The data processing was carried out using RStudio v.4.0.5 software. Each parameter was normalized with the max-min scaling method, in order to reduce the bias caused by the predominance of very high ranges over lower ranges. The extract point value tool in ArcGIS v.10.2.1 was used to obtain the data of each variable for all the 5842 points.

Table 4 Quantitative land use values in the study area. Higher values represent a higher tendency to contamination and vice versa

Land use	Value
Urban areas	5
Industrial-commercial areas	5
Landfill	5
Irrigation crops	5
Non-irrigated arable land	3
Water courses	3
Non-vegetation	2
Forest and green areas	1

3.3.1 Feature selection by Principal Analysis Component (PCA)

Principal Component Analysis (PCA) was used to identify the relevant features from the original dataset, following the procedure proposed by Song et al., (2010).

The contribution of each eigenvector was calculated as the sum of whole absolute eigenvalues within the eigenvector. They were arranged in descending order, representing the hierarchy of the importance of each variable. The PCA reduces the dimension of dataset, explaining as much variance as possible. Thus, the dimension of the dataset could be reduced and the new low-dimensional data set was considered to select the relevant variables in the principal components. The application of PCA allowed a large number of correlated variables to be replaced by a smaller number of uncorrelated variables (eliminating redundancy), while retaining the most information from the original model (guaranteed with a high cumulative variance).

3.3.2 K-means cluster by low dimensional data set

K-means clustering analysis was applied to the new low-dimensional data set obtained from the PCA. Following the procedure described in the section 3.3, a new smaller data set was created with an "*n* "x "*d"* matrix, where *n* is the number of data points (5842) in a *d* low dimensional feature space.

3.4 Vulnerability map validation by using the nitrate concentration as an indicator of contamination.

The main pollutant in the "Aluviales Jarama-Tajuña" aquifer is nitrate, because of the intense agricultural activity (Arauzo et al. 2008; Mostaza-Colado et al. 2018). For this reason, nitrate concentration has been considered in this work as a reference indicator to validate the obtained vulnerability maps. Concentration data from 23 monitoring wells were classified into four categories as pollution indicator (low <12mg/L, moderate 12-25mg/L, high 25-50mg/L and very high>50mg/L). Values above 50mg/l were considered as very high nitrate concentration because they exceed the limit recommended by the Spanish Government (BOE 1996).

The ArcGis extraction tool allowed to obtain the corresponding the vulnerability index value for each nitrate concentration monitoring well.

Finally, a statistical analysis using Spearman's correlation coefficient was carried out to verify the degree of association between the vulnerability index and nitrate concentration. This analysis was performed to validate the vulnerability results of the different methods (Panagopoulos et al. 2006; Javadi et al. 2011; Yang et al. 2017; Barzegar et al. 2019).

4. Results and discussion

4.1 DRASTIC vulnerability analysis

The spatial distribution of the classes defined for each DRASTIC parameter is shown in Fig. 3.The maximum depth of groundwater (*D*) (29.7m), is found in the central zone of the study area and minimum values (around 6.3m) are located in the north and south sectors (Fig. 3a). Almost 70% of the study area has a water table depth of less than 9m, which determines that the most of the area is vulnerable to contamination due to the small thickness of the unsaturated zone.

The net recharge (*R*) varies from 0 to 984 mm per year. The maximum values correspond to irrigated zones (mostly located in the south), covering an area of 20%. In the central and northern zones, the recharge is a combination of rainfall and irrigation (more than 60% of the study area) (Fig. 3b). Despite the area with highest recharge is small, the recharge is higher than 254mm (maximum limit established by DRASTIC methodology) what favors contaminant infiltration from the surface. The natural recharge of the aquifer (from rainfall) is very low (See Supplementary material, Tables S1a, S1b and S1c), compared to the recharge by infiltration from irrigation returns (See Supplementary material, Table S2), which is in agreement with the results of previous works (Mostaza-Colado et al. 2018).

The aquifer media factor (*A*) of the study area is defined by the dominant presence of sands and gravels (rating values 4-9). The highest permeability of the aquifer occurs in the central zone (35% of total area), which is considered the most vulnerable to contamination (Fig. 3c). Although the permeability of the aquifer is high, it shows little variation. Therefore, this parameter does not affect the distribution or variability of the DRASTIC vulnerability index.

The soil media factor (*S*) is defined by the occurrence of loamy and silty loamy textures, the latter being found in most of the study area (approximately 70%) (Fig. 3d). This type of soil texture helps to protect the vadose zone from the entry of contaminants.

More than 70% of the study area has a very low slope (*T*) (between 0-2%), only increases at the boundaries of the area and along the river banks (Fig. 3e). The gentle topography results in low surface runoff, which favors vulnerability to irrigation-related infiltration of pollutants.

The vadose zone (*I*) is defined by the occurrence of gravels, sands, clays, and silts (more than 80% of the study area). The most permeable materials are located in the central zone (Fig. 3f). Permeability contributes to the movement of pollutant movement from the surface to aquifer increasing the vulnerability there.

The aquifer hydraulic conductivity factor (*C*) ranges from 0 to 476 m/d. The hydraulic conductivity in more than 50% of the study area was higher than 81.49 m/d (which is the highest limit established by the DRASTIC methodology). The highest values are located in the south and in some areas in the north (Fig. 3g). These areas are susceptible to have high vulnerability, due to their high transmissivity and low unsaturated thickness. The lowest values of hydraulic conductivity are located in the central zone. Therefore, this area is less vulnerable to contamination due to its low hydraulic conductivity.

The DRASTIC Vulnerability Index (*DI*) ranged from 94 to 207. The distribution of the four defined vulnerability classes (low, moderate, high and very high vulnerability) is shown in the DRASTIC vulnerability map (Fig. 4). Almost 20% of the study area (mostly in the southern part) shows a very high vulnerability, influenced by the recharge related with high crop irrigation. A large part of the aquifer shows high vulnerability (53% of the study area). This is located along study area (north, central-west and edges at south zones), related to the high permeability of materials in these zones. Moderate and low vulnerability values are identified in the central-eastern part of the study area, where the water level is deeper and the hydraulic conductivity is lower. The DRASTIC vulnerability map shows that almost 70% of the study area has high and very high vulnerability. This result reveals that the “Aluviales Jarama-Tajuña” aquifer is highly vulnerable to contamination.

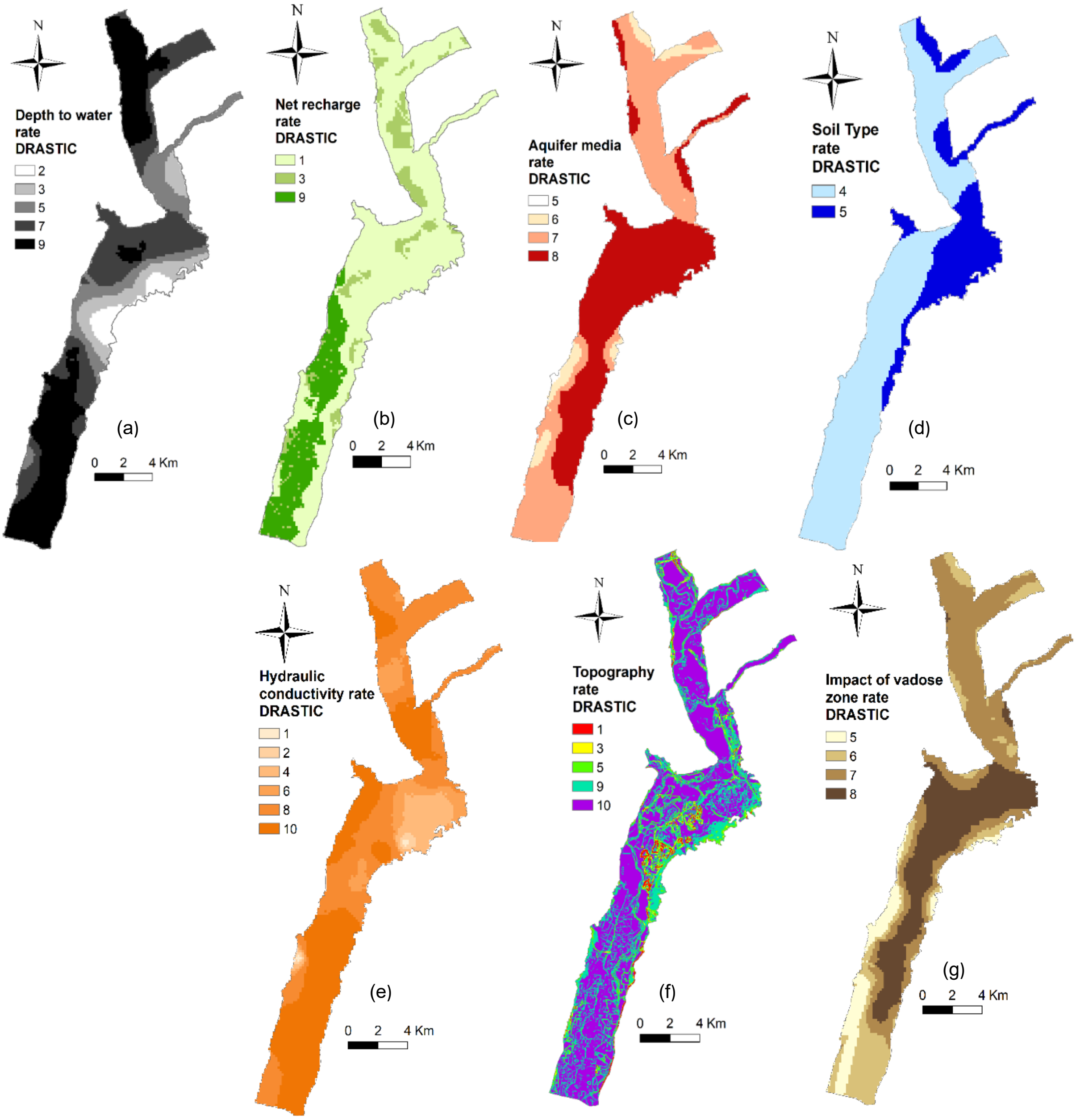

Fig. 3 DRASTIC method maps. (a) Depth to water table, (b) Net recharge , (c) Aquifer media , (d) Soil type , (e) Topography , (f) Impact of vadose zone , (g) Hydraulic conductivity

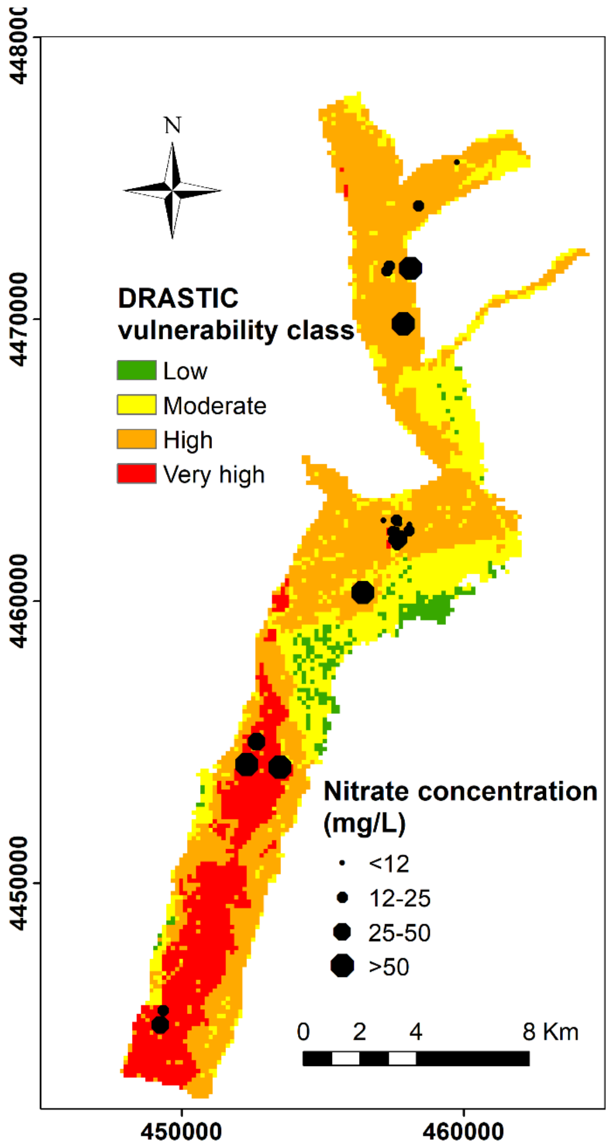

Fig. 4 Vulnerability index map of DRASTIC, showing nitrate concentration ranges at wells location

4.2 GOD vulnerability analysis

The spatial distribution of the classes defined for each GOD parameter is shown in Fig. 5a and Fig. 5b. The "Aluviales Jarama-Tajuña" aquifer is unconfined, according to lithological sections (Fig. 2b) and pumping test (Carreño Conde et al. 2014). Therefore, 100% of study area is unconfined aquifer and the map of groundwater occurrence (*G*) is defined by a single value equal to one (1) according to *GOD* method. The lithology of aquifer factor (*O*), equivalent to vadose zone factor in *DRASTIC*, varies from 0.5 to

0.7, as gravels, sands, clays and silts constitute 100% of study area. There is little variation of the permeability regarding the thickness of unsaturated material, which makes the area very vulnerable (Fig. 5a). As in the DRASTIC method, depth of groundwater factor (*D*) is low (depth is less than 10m in more than 70% of area), which is contributing to the high vulnerability (Fig. 5b).

The *GOD* vulnerability index ranged from 0.32 to 0.70. The map in Fig. 5c, shows the distribution of the normalized and classified GOD vulnerability index. Almost 60% of the study area has very high and high vulnerability (40% and 20%, respectively). This occurs in three well-defined zones located in the south, central and north parts of the study area. The very high to high vulnerability is due to the join effect of the high permeability of the materials and the low thickness of the unsaturated zone. 36% of the study area shows moderate vulnerability, mainly in the central zone where the relatively high depth of groundwater decreases the possibility that the pollutant reaching the aquifer. Only 4.6% of area displays low vulnerability, which occurs at the lateral edge of the aquifer at east of the central zone and in the southwestern part of study area. There, the materials consist mainly of clays and gypsum that reduce the infiltration.

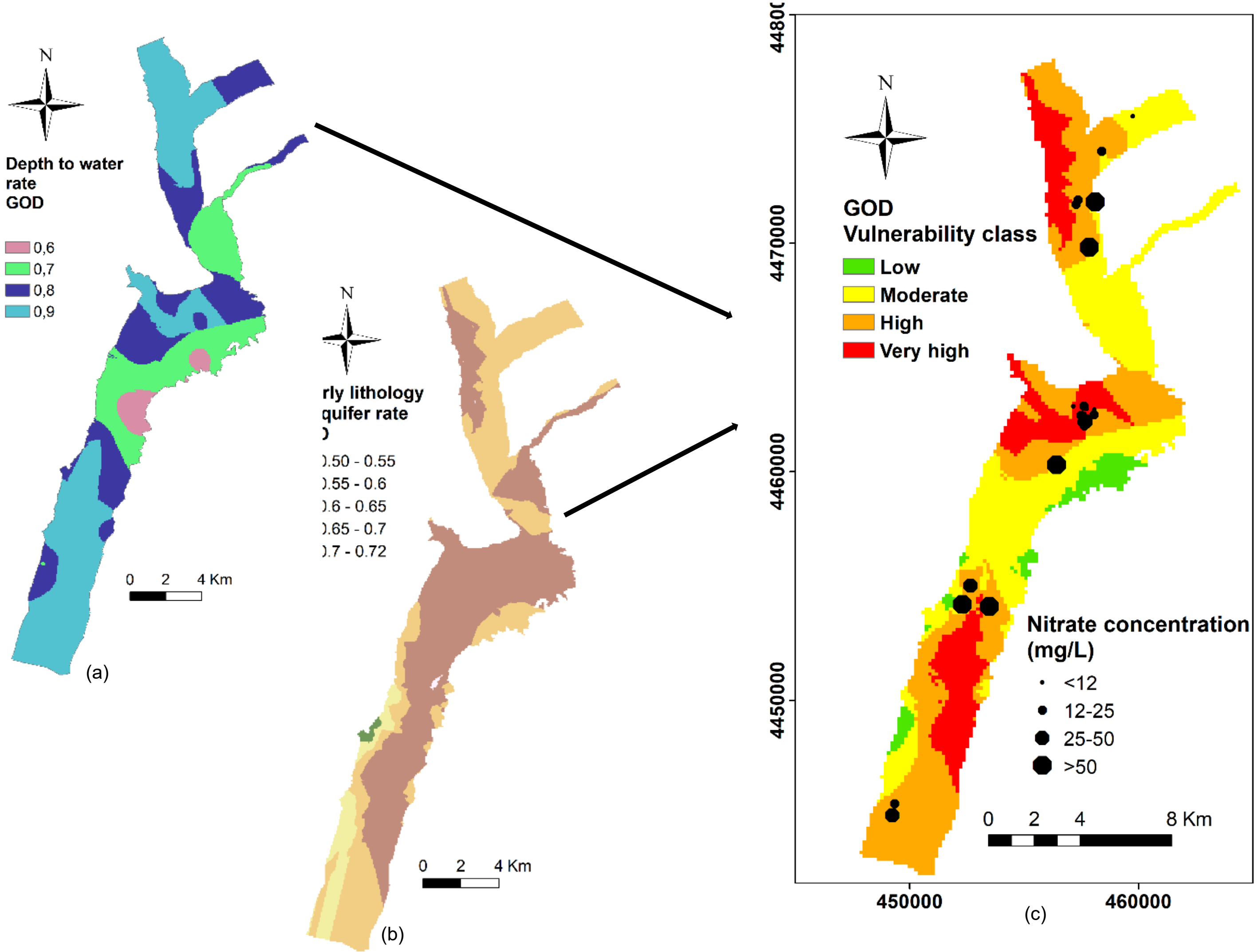

Fig. 5 GOD method maps. (a) Depth of water table (D), (b) lithology (O), (c) Vulnerability index map of GOD, showing nitrate concentration ranges at wells location

4.3 AVI vulnerability analysis

The hydraulic resistance values varied between 0 to 767 years (Fig. 6a). 68.2% of the study area shows very high vulnerability from north to south (only moderate to low vulnerability predominates in the southernmost part, Fig. 6b). The low values are related to the high hydraulic conductivity of the unsaturated zone, together with a low thickness (6m on average) of this layer.

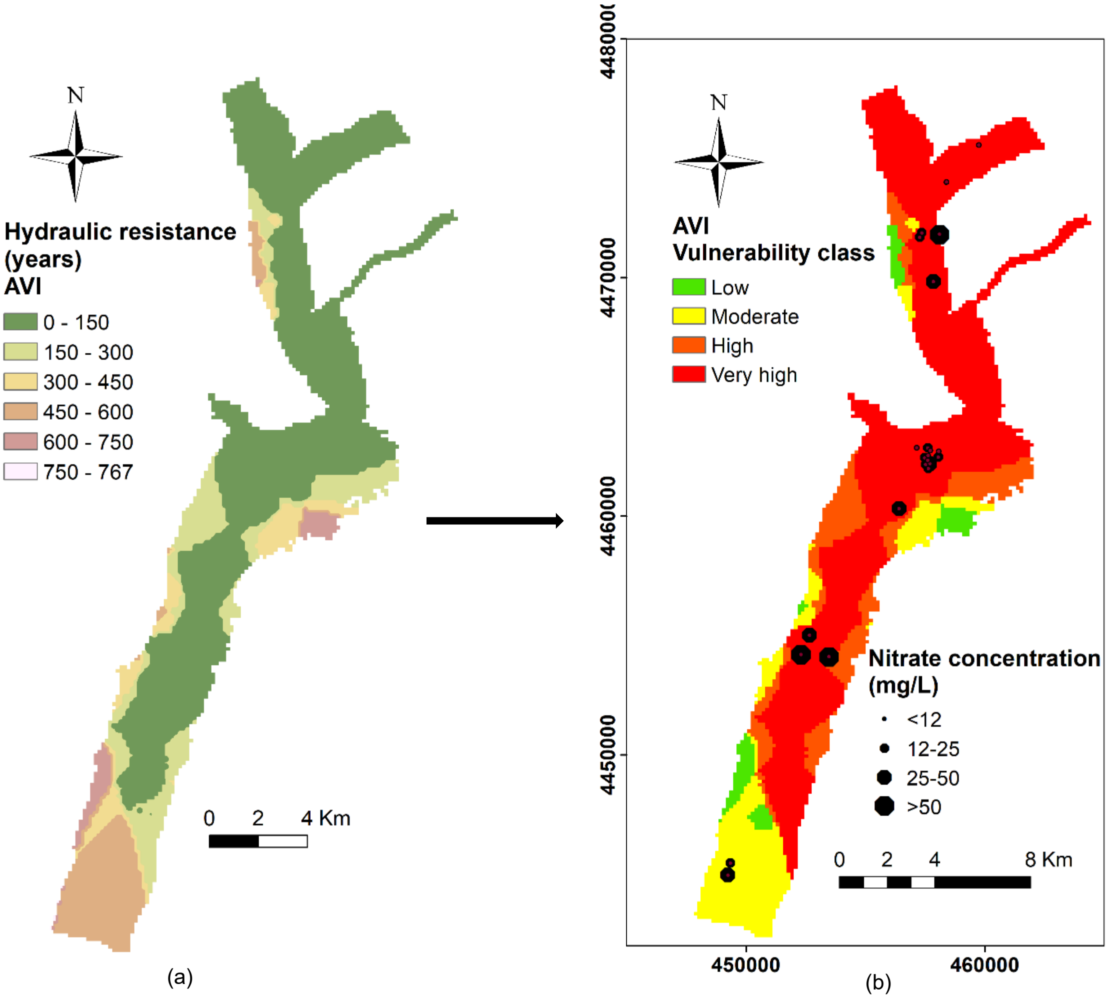

Fig. 6 AVI method maps. (a) Hydraulic resistance, (b) Vulnerability index map of AVI, showing nitrate concentration ranges at well location

4.4 K-means Cluster analysis

The parameters considered in the K-means cluster analysis are shown in Fig. 7.

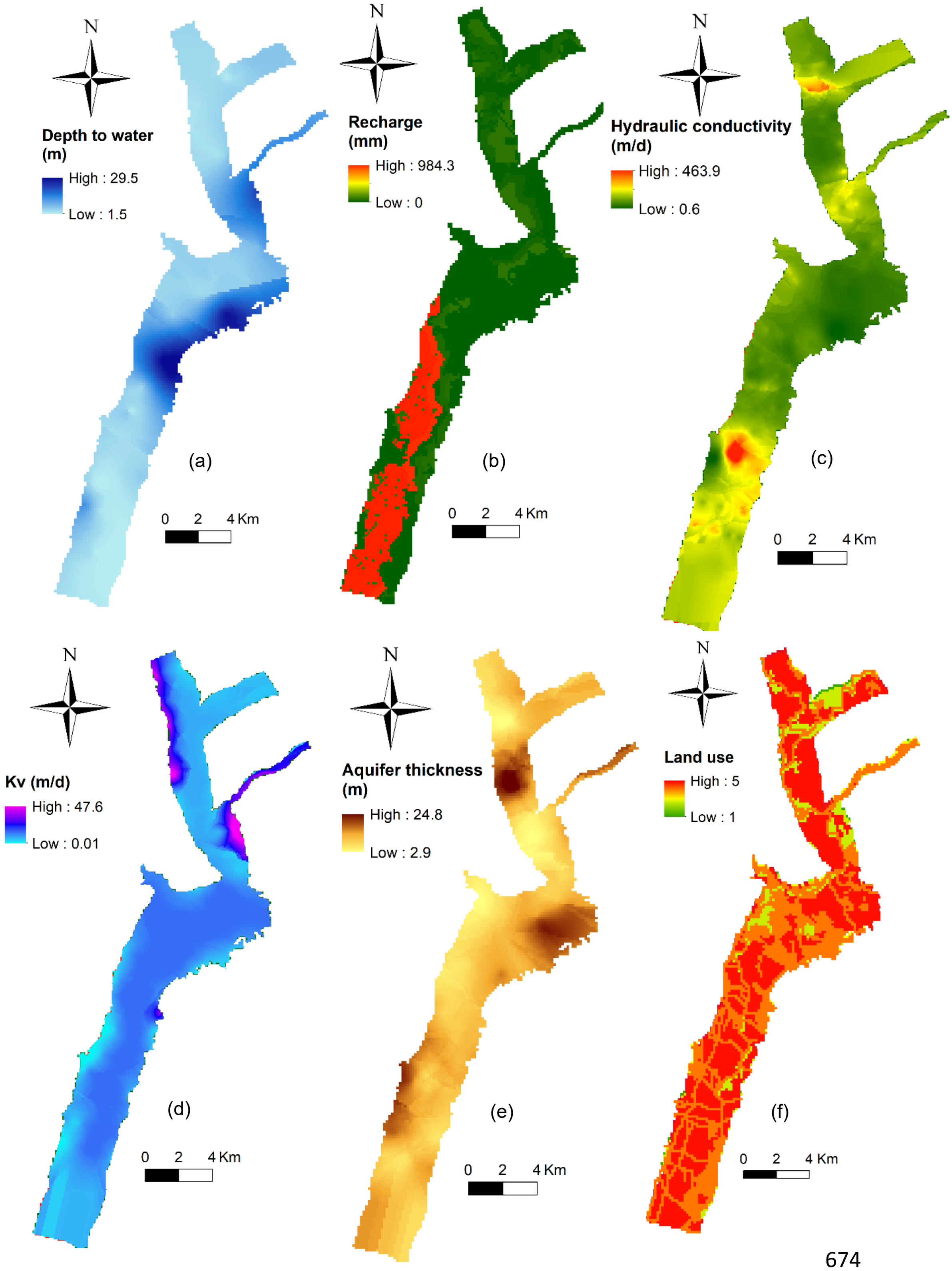

Fig. 7 K-means parameters maps. (a) Depth to water table, (b) Net recharge , (c) Hydraulic conductivity, (d) Vertical hydraulic conductivity on unsaturated zone, (e) Aquifer thickness, (f) Land use

The “*n x d”* data matrix was made of 5842 data points and six feature space (“*D*”, “*R*”, “*C*”, “*Kv*”, “*Th*” and “*L*”). After max-min normalization of database, the method resulted in an optimal number of three (3) clusters, proposed by 12 of 26 index.

The results of high dimensional K-means cluster analysis are summarized in Table 5.

Table 5 Variation of feature data in the three identified clusters (High dimensional dataset)

			D (m)	R (mm/year)	C (m/d)	Kv(m/d)	Th(m)	L	Vulnerability
Cluster	points	%	Mean	Mean	Mean	Mean	Mean	Mean	
1	2461	42.1	8.8	16.3	82.0	9.0	10.9	2.6	Low
2	2147	36.8	8.7	35.3	80.1	8.5	11.4	5.0	Moderate
3	1234	21.1	5.3	967.9	132.6	6.9	10.9	4.3	High

Cluster 3 includes the lowest values of depth of groundwater (*D*) and the highest values of recharge and hydraulic conductivity (*R,C*)*.* In addition, land use (*L*), has a high value. All these conditions contribute to define high vulnerability. On the other hand, Cluster 1 represents the opposite scenario of low vulnerability with the lowest values of recharge (*R*) and land use (*L*). Cluster 2 shows moderate vulnerability with higher recharge (*R*) than cluster 1, but lower than cluster 3. Although land use (*L*) in cluster 2 has the highest value, it was very similar to cluster 3, Thus recharge (*R*) and land use (*L*) together contribute to define moderate vulnerability in cluster 2. Note that vertical permeability on unsaturated zone (*Kv*) and aquifer thickness (*Th*) did not influence vulnerability ranking, as they were very similar in all clusters.

4.4.1 K-means cluster by low dimensional analysis

To perform the K-means clustering in a low dimensional dataset, three of the six features were selected using PCA analysis (Table 6). According to the procedure described in the Materials and Methods section, the selected features explain more than 86% of the variance. The relevant features were net recharge (*R*), Depth of water table (*D*) and land use (*L*), in this order of importance calculated by their contribution. These three features, selected by PCA for 5842 points across the study area, produced the low dimensional data set.

Table 6 Eigen vectors and Eigen values, varimax component matrix and eigenvectors contribution obtained from the PCA. Bold numbers in eigenvectors represent the maximum eigen values associated to each parameter.

Parameters	PC1	PC2	PC3
D	0.1766423	-0.113552	**0.9424516**
R	**-0.906886**	0.334078	0.2368487
C	-0.124457	0.072504	-0.197747
Kv	0.0499861	0.0002084	0.0742895
Th	0.0089448	-0.083407	0.0905335
L	-0.358171	**-0.929131**	-0.05356
Standard deviation	0.413	0.3031	0.2198
Proportion of Variance	0.4746	0.2555	0.1344
Cumulative Proportion	0.4746	0.7301	0.8645
Contribution	1.63	1.53	1.6

The K-means cluster analysis on the low dimensional data set resulted in three clusters as the optimal number of clusters, as was the case for the high-dimensional dataset. The results of the low dimensional K-means cluster analysis are summarized in Table 7.

Table 7 Variation of features data in the three identified clusters (Low dimensional dataset)

	points	%	D (m)	R (mm/year)	L	Vulnerability
Cluster	5842	100	Mean	Mean	Mean	
1	2461	42.1	8.8	16.3	2.6	Low
2	2147	36.8	8.7	35.3	5.0	Moderate
3	1234	21.1	5.3	967.9	4.3	High

The results of K-means cluster analysis on the low dimensional data set show the same behavior as the high dimensional data set. The clusters consist of the same number of points and represent the same vulnerability classes of vulnerability. Fig. 8, shows the clustering vulnerability map, where 35.9 % of the study area corresponds to low vulnerability, 41.4 % to moderate vulnerability and 22.7 % to high vulnerability.

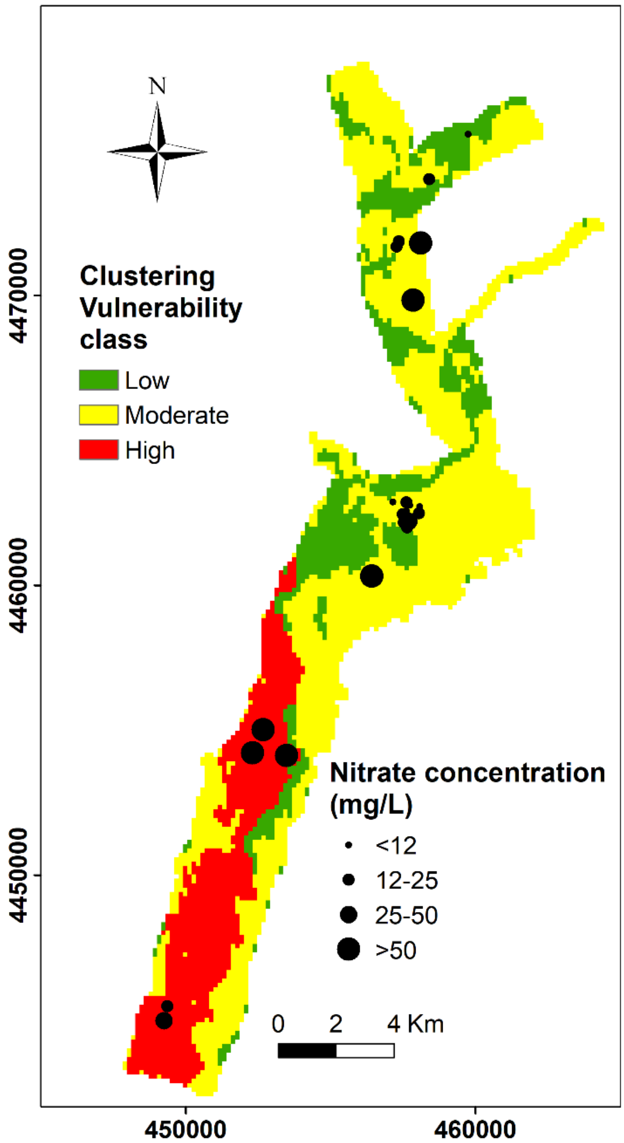

Fig. 8 Vulnerability map by K-means cluster analysis. Information on nitrate concentration (range and location) is included.

4.5 Vulnerability method validation

The nitrate observation points were classified by their concentration in four categories (Table 8). The classified points have been projected onto the vulnerability maps (Fig. 4, Fig. 5c, Fig. 6b and Fig. 8).

Table 8 Nitrate pollution indicator (four classes) based on nitrate concentration in water quality monitoring wells.

	Nitrate Concentration (mg/L)			
	<12	12 - 25	25 - 50	>50
Samples	5	10	3	5
Percentage (%)	21.7	43.5	13.0	21.7
Nitrate pollution indicator	low	moderate	High	very high

The graphical coincidences for high and low vulnerability and high and low nitrate pollution are noticeable in *DRASTIC* and K-means maps (Fig. 4 and Fig. 9). In contrast, *GOD* and *AVI* methods show less graphical agreement (Fig. 5c and Fig. 6b). Table 9 shows the Spearman´s correlation coefficient between nitrate concentration samples and each method used to assess the vulnerability to contamination.

Table 9 Spearman correlation coefficient between nitrate concentration and vulnerability p-value of the studied methods.

Method	Spearman rank correlation (rho)	p-value
DRASTIC	0.34	0.049*
GOD	-0.50	0.007**
AVI	0.01	0.48
K-means (Low dimensional data set	0.48	0.019*

*Spearman test p-value<0.05

The vulnerability indices *GOD* and *AVI* vulnerability did not yield a valid correlation with nitrate concentration values. Better correlations were obtained by the *DRASTIC* and the K-means methods. However, the cluster analysis showed a better correlation with nitrate concentration, with higher correlation coefficients compared to those for *DRASTIC* method. K-means cluster analysis resulted in 48% of Spearman´s correlation coefficients. The p-values confirms that the best methods (*DRASTIC*, K-means) were statistically significant.

Fig. 9, shows the percentage of area with very high, high, moderate and low vulnerability, depending on the applied assessment method, as well as the nitrate contamination range classes. The results obtained from the *AVI* method were completely different from the rest of the methods, as the *AVI* method assigned very high vulnerability to a large portion of the aquifer (more than 60% of the study area). This contrasting result is due to the fact that this assessment method only considers the travel time of the contaminant through the unsaturated zone. The low correlation of the *AVI* method with nitrate pollution (Table 9) shows that more characteristics need to be considered to obtain better or more adjusted vulnerability assessment. Thus, the *AVI* method is not suitable to be applied to an aquifer whose vulnerability is dominated by hydrological and hydrogeological features as net recharge, depth of water table and land use. The *GOD* method showed a negative correlation, meaning that the high values of nitrate concentration correspond with low vulnerability values and vice versa. This method does not take the aquifer recharge into account like the *AVI* method, which confirmed that recharge is a feature of paramount importance in the vulnerability assessment of the study area. In addition, the vulnerability assessed in the study area by the *GOD* method is strongly influenced by depth of water table over the other parameters considered in the methodology. The low correlation of *GOD* and nitrate concentration (Table 9) is due to the fact that the depth of groundwater in this case, is not sufficient to define vulnerability zones, suggesting that in detrital aquifers is necessary to consider others parameters. *DRASTIC* resulted in a lower proportion of very high vulnerability, similar to the percentage of samples with very high nitrate contamination (around 19%). On the other hand, *DRASTIC* showed different proportions of high, moderate and low vulnerability compared to the percentage of samples of nitrate concentration classes (Fig. 9). Despite this, the Spearman`s correlation coefficient between the vulnerability index of *DRASTIC* and the nitrate concentration was higher than *GOD* and *AVI* methods (34%, Table 8), indicating that some of the parameters considered on *DRASTIC* method had a major influence on improving the vulnerability assessment in the aquifer. The K-means method showed the highest Spearman´s correlation coefficient between vulnerability classes and nitrate concentration (48%). This showed that it is important to select non-redundant parameters and, in this case, the most influencing parameters were net recharge, depth of groundwater and land use, as obtained by PCA analysis. Considering nitrate as an indicator contamination (Table 8, Fig. 9), almost 22% of the samples corresponded to the very high pollution class, with the nitrate concentration exceeding the recommended limit (50mg/L). These samples are located on high vulnerability values areas in the cluster map. The high nitrate concentrations located in the areas with agricultural uses show that land use (*L*) is a very important variable in determining vulnerability. In addition, the low recharge by rainfall (fresh water) and the high recharge with water from irrigation returns (which are loaded with nitrogen fertilizers) have an significant influence on the high vulnerability of the aquifer, confirming the findings of Mostaza-Colado et al. (2018). Many water quality samples (43%) are indicative of moderate pollution (12-25 mg/L), the most numerous being located in the central zone of the aquifer coinciding with the moderate vulnerability zones in the K-means cluster map (where the water depth is higher and net recharge is lower). According to Mostaza-Colado et al. (2018); and Mostaza (2019), good agricultural practices influence the reduction of nitrate concentration in the central zone of the aquifer. However, this zone has a moderate vulnerability with moderate values of nitrate concentration (12-25 mg/L), which are mainly due to the aquifer conditions and land use and not to irrigation techniques. Therefore, good agricultural practices are not a significant factor for the vulnerability assessment of the studied aquifer and the low nitrate concentrations in this zone would be caused by the low recharge and the high depth of the water table, which makes it difficult for the nitrate to reach the aquifer.

K-means cluster analysis based on relevant features emerges as the best method to assess the vulnerability to pollution of a detrital aquifer, being more objective that the overlay index methods. The advantages of this refined K-means methodology are in line with Foster et al. (2013), who indicate that the best application of the pollution vulnerability assessment methodologies will be achieved when these methods incorporate (as simply and sensitively as possible) the main parameters controlling hydraulic accessibility and natural protection of the aquifer.

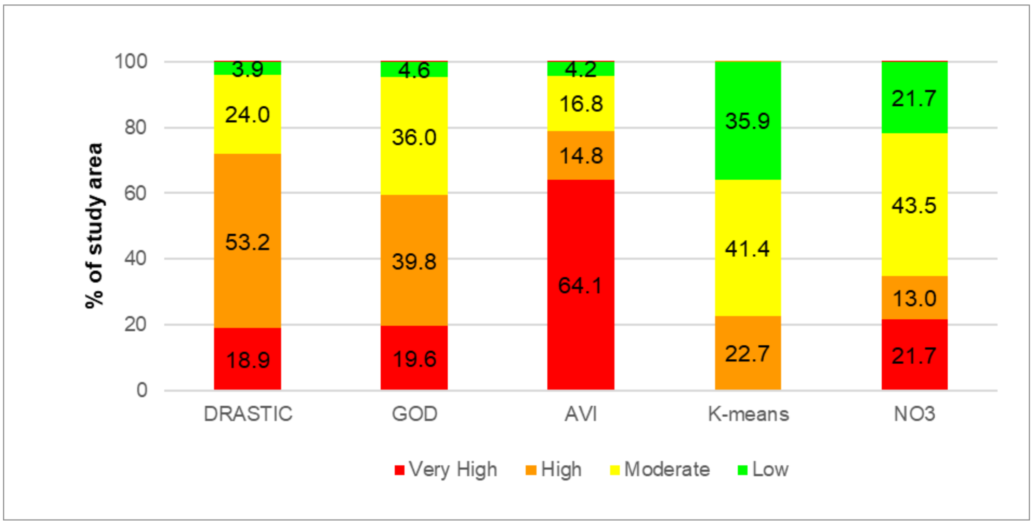

Fig. 9 Percentage of area with low, moderate, high and very high vulnerability related to each applied assessment method and percentage of samples in nitrate contamination range classes

5. Conclusions

Although vulnerability assessment maps have proven to be a useful tool to prevent and control the process of groundwater contamination, the selection of the most appropriate method is paramount. In this work, vulnerability of the "Aluviales Jarama-Tajuña" aquifer in Spain has been assessed by overlay index maps methods (*DRASTIC, GOD, AVI*) and K-means clustering analysis. The vulnerability maps obtained by each method were compared with the concentration of nitrate in groundwater samples as an indicator of contamination, in order to validate the most appropiate method to use. The results showed that is important to take into account the relevant features in a specific area, as the lack of appropriate parameters could lead to inappropriate results. Furthermore, methods with a short number of parameters should be used with caution in studies of detrital aquifers, as the few parameters considered may not be relevant or sufficient to obtain a good vulnerability assessment. This is the case for *GOD* and *AVI* methods, which did not take into account relevant features such as net recharge and land use in the aquifer under study. *DRASTIC* gave better results, as it considers some of these features as well as other parameters that control the vulnerability of the aquifer. The *DRASTIC* results improved significantly the correlation with nitrate concentration (34%). However, not all parameters used in *DRASTIC* were relevant for the assessment. This was demonstrated by the K-means analysis, which considered a new set of parameters extracted from index methods. Six parameters were identified (Depth of groundwater (*D*), recharge of the aquifer (*R*), land use (*L*), hydraulic conductivity of the aquifer (*C*), hydraulic conductivity of unsaturated zone (*Kv*), aquifer thickness (*Th*)). The PCA analysis was applied to that set, obtaining the key hydrogeological parameters that affect the vulnerability of the detrital aquifer. The parameters identified as relevant after PCA analysis were depth of water table (*D*), net recharge (*R*), and land use (*L*). The new proposed method grouped data in three clusters that represent low vulnerability (42.1% of the study area), moderate (36.8%) and high vulnerability (21.1%). Nitrate concentration has been used as indicator of contamination to validate the results obtained by the methods used in the study. The application of K-means cluster yielded the best correlation (48%) between vulnerability values and nitrate concentration, increasing significantly that obtained from the other methods. The study shows that cluster analysis methods can be applied to significantly eliminate the subjectivity of the traditional vulnerability assessment methods, as they do not associate rating or weighting coefficients. Also, the few variables selected facilitate data collection and guarantee optimal results, as they represent key factors for the aquifer studied. Thus, the use K-means cluster analysis confirmed the advantage of applying data mining techniques in the assessment of groundwater vulnerability in detrital aquifers.

Acknowledgments

Authors wish to acknowledge Francisco Carreño (Universidad Rey Juan Carlos) and David Mostaza (IMIDRA) for their data contribution and collaboration in this study.

Declaration of Competing Interest

This paper is considered as a scientific research production for the promotion of the knowledge from the University Rey Juan Carlos, Madrid- Spain.

There is no conflict of interest with the objectives of government institutions or water policies, and the research follows scientific ethics and scientific integrity principles.

This research has not been supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution Statement

All authors contributed to the study conception and design***. Marisela Uzcategui-Salazar***: Material preparation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing- Original draft preparation, Visualization. ***Javier Lillo Ramos****:* Investigation, Validation, Supervision, Visualization, Writing- Reviewing and Editing.

All authors reviewed the results and approved the final version of the manuscript.

References

Aller L, Benett T, Lehr JH, et al (1987) DRASTIC: A Standardized System for Evaluating Ground Water Pollution Potential Using Hydrogeologic Settings. Robert S. Kerr Environmental Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Ada, Oklahoma

Arauzo M, Martinez-Bastidas JJ, Valladolid M (2008) Contaminación por nitrógeno en el sistema “río-acuífero aluvial” de la cuenca del Jarama (Comunidad de Madrid, España) ¿Origen agrícola o urbano? Limnetica 27(2):195–210

Arezooman omidi, KHashei S, Javadi S, Hashemi S (2015) Groundwater Vulnerability Assessment by the use of Drastic-Nw Modified model (Case study: Kuchesfehan-Astane plain). Iranian Journal of Irrigation and Drainage 9:75–82

Asadi P, Ataie-Ashtiani B, Beheshti A (2017) Vulnerability assessment of urban groundwater resources to nitrate: the case study of Mashhad, Iran. Environmental Earth Sciences 76:. https://doi.org/10.1007/s12665-016-6357-z

Babiker IS, Mohamed MAA, Hiyama T, Kato K (2005) A GIS-based DRASTIC model for assessing aquifer vulnerability in Kakamigahara Heights, Gifu Prefecture, central Japan. Science of The Total Environment 345:127–140. https://doi.org/10.1016/j.scitotenv.2004.11.005

Bardají I, Cabra P, Calvo JP, et al (1990) Estudio para la ordenación de la actividad extractiva del tramo bajo del Río Jarama

Barzegar R, Asghari Moghaddam A, Adamowski J, Nazemi AH (2019) Delimitation of groundwater zones under contamination risk using a bagged ensemble of optimized DRASTIC frameworks. Environ Sci Pollut Res 26:8325–8339. https://doi.org/10.1007/s11356-019-04252-9

Barzegar R, Asghari Moghaddam A, Norallahi S, et al (2020) Modification of the DRASTIC Framework for Mapping Groundwater Vulnerability Zones. Groundwater 58:441–452. https://doi.org/10.1111/gwat.12919

BOE (1996) Real Decreto 261/1996, de 16 de febrero, sobre protección de las aguas contra la contaminación producida por los nitratos procedentes de fuentes agrarias. https://www.boe.es/eli/es/rd/1996/02/16/261. Accessed 21 May 2020

Calvo JP, Ordonez S, Sanz E (1989) Sedimentologia de los complejos lacustres miocenos de la Cuenca de Madrid. ACTA GEOLOGICA HISPANICA 24:281–298

Carreño Conde FC, García Martínez SG, Lillo Ramos JL, et al (2014) Building a 3D geomodel for water resources management: case study in the Regional Park of the lower courses of Manzanares and Jarama Rivers (Madrid, Spain). Environ Earth Sci 71:61–66. https://doi.org/10.1007/s12665-013-2694-3

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust package : Manual

CHT (2015a) Plan Hidrológico de la parte española de la Demarcación Hidrológica del Tajo. Anejo 3 Usos y demandas de agua. In: Confederación Hidrográfica del Tajo. http://www.chtajo.es/LaCuenca/Planes/PlanHidrologico/Planif_2015-2021/Paginas/Plan_2015-2021.aspx. Accessed 2 Nov 2021

CHT (2019) Red de control de las aguas subterráneas en la Cuenca Hidrográfica del Tajo. In: Confederación Hidrográfica del Tajo. http://www.chtajo.es/LaCuenca/CalidadAgua/AguasSubterraneas/Paginas/default.aspx. Accessed 20 Jan 2020

CHT (2015b) Plan Hidrológico de la parte española de la Demarcación Hidrológica del Tajo. Memoria parte española de la demarcación hidrográfica del Tajo Plan hidrológico de cuenca

Coduto D (1999) Geotechnical Engineering: Principles and Practices. Prentice-Hall, Englewood Cliffs, NJ

Conti D, Gibert K (2014) Discovering Comprehensible Hydrogeological Profiles in the Margarita Island's Aquifers including Post-processing in a data mining process. San Diego, California USA, pp 1434–1441

Custodio E, Llamas M (2002) Hidrología subterránea. Tomo I, Omega. Barcelona, España

Dabbura I (2020) K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. In: Medium. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. Accessed 26 May 2021

Dash M, Koot PW (2009) Feature Selection for Clustering. In: LIU L, ÖZSU MT (eds) Encyclopedia of Database Systems. Springer US, Boston, MA, pp 1119–1125

Debernardi L, De Luca DA, Lasagna M (2008) Correlation between nitrate concentration in groundwater and parameters affecting aquifer intrinsic vulnerability. Environ Geol 55:539–558. https://doi.org/10.1007/s00254-007-1006-1

Domenico P, Schwartz F (1998) Physical and Chemical Hydrogeology, 2nd Edition. Wiley

Fetter C (2001) Applied Hydrogeology, 4th Edition. Prentice-Hall

Foster S, Hirata R (1991) Determinación del riesgo de la contaminación de aguas subterráneas. Una metodología basada en datos existentes. Centro Panamericano de Ingeniería Sanitaria y Ciencias del Ambiente (CEPIS), Lima, Perú

Foster S, Hirata R, Andreo B (2013) The aquifer pollution vulnerability concept: aid or impediment in promoting groundwater protection? Hydrogeol J 21:1389–1392. https://doi.org/10.1007/s10040-013-1019-7

Foster SSD (1987) Fundamental Concepts in Aquifer Vulnerability, Pollution Risk and Protection Strategy: International Conference, 1987, Noordwijk Aan Zee, the Netherlands Vulnerability of Soil and Groundwater to Pollutants The Hague, Netherlands Organization for Applied Scientific Research P69-86. Netherlands Organization for Applied Scientific Research

Hao J, Zhang Y, Jia Y, et al (2017) Assessing groundwater vulnerability and its inconsistency with groundwater quality, based on a modified DRASTIC model: a case study in Chaoyang District of Beijing City. Arabian Journal of Geosciences 10:. https://doi.org/10.1007/s12517-017-2885-4

Huan H, Wang J, Teng Y (2012) Assessment and validation of groundwater vulnerability to nitrate based on a modified DRASTIC model: A case study in Jilin City of northeast China. Science of The Total Environment 440:14–23. https://doi.org/10.1016/j.scitotenv.2012.08.037

IGME (1984) MapasIGME - Portal de cartografía del IGME: MAGNA 50 - Mapa Geológico de España a escala 1:50.000 (2ª Serie). Sheets: 559, 560, 582, 583 and 605. In: Instituto Geológico y Minero de España .Cartografía del IGME. http://info.igme.es/cartografiadigital/geologica/Magna50.aspx. Accessed 1 Oct 2019

IGN (2019) Digital elevation model (25 meters). España – MDT25/MDT25- LIDAR. In: Centro de Descargas del Centro Nacional de Información Geográfica (CNIG). http://centrodedescargas.cnig.es. Accessed 1 Nov 2019

IGN (2018) CORINE Land Cover . Scale 1:100.000. In: Instituto Geográfico Nacional.Centro de Descargas del Centro Nacional de Información Geográfica (CNIG). http://centrodedescargas.cnig.es

IGN (2008) Soil Map of Spain. In: Instituto Geográfico Nacional . Centro de Descargas del Centro Nacional de Información Geográfica (CNIG)

Jafari SM, Nikoo MR (2016) Groundwater risk assessment based on optimization framework using DRASTIC method. Arab J Geosci 9:742. https://doi.org/10.1007/s12517-016-2756-4

Javadi S, Hashemy Shahdany SM, Neshat A, Chambel A (2020) Multi-parameter risk mapping of Qazvin aquifer by classic and fuzzy clustering techniques. Geocarto International 0:1–23. https://doi.org/10.1080/10106049.2020.1778099

Javadi S, Hashemy SM (2016) Evaluation of Groundwater Vulnerability Using Data Mining Technique in Hashtgerd Plain. Journal of the Earth and Space Physics 42:35–41. https://doi.org/10.22059/jesphys.2016.57743

Javadi S, Hashemy SM, Mohammadi K, et al (2017) Classification of aquifer vulnerability using K-means cluster analysis. Journal of Hydrology 549:27–37. https://doi.org/10.1016/j.jhydrol.2017.03.060

Javadi S, Kavehkar N, Mohammadi K, et al (2011) Calibrating DRASTIC using field measurements, sensitivity analysis and statistical methods to assess groundwater vulnerability. Water International 36:719–732. https://doi.org/10.1080/02508060.2011.610921

Kadkhodaie F, Asghari Moghaddam A, Barzegar R, et al (2019) Optimizing the DRASTIC vulnerability approach to overcome the subjectivity: a case study from Shabestar plain, Iran. Arab J Geosci 12:527. https://doi.org/10.1007/s12517-019-4647-y

Kazakis N, Voudouris KS (2015) Groundwater vulnerability and pollution risk assessment of porous aquifers to nitrate: Modifying the DRASTIC method using quantitative parameters. Journal of Hydrology 525:13–25. https://doi.org/10.1016/j.jhydrol.2015.03.035

Malmir M, Javadi S, Moridi A, et al (2021) A new combined framework for sustainable development using the DPSIR approach and numerical modeling. Geoscience Frontiers 12:101169. https://doi.org/10.1016/j.gsf.2021.101169

Marín Celestino ME, Martínez Cruz DA, Otazo Sanchez EM, et al (2018) Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, Principal Component Analysis and Spatial Analysis: A Case Study. Water 10:. https://doi.org/doi.org/10.3390/w10040437

Mendoza SN (2012) Calibración del método DRASTIC en acuíferos utilizando nitratos como indicador. 11

Monturiol F, Alcalá del Olmo L (1990) Mapa de las Asociaciones de Suelos de la Comunidad de Madrid. Consejo superior de investigaciones científicas (CSIC). Consejería de Agricultura y Cooperación, Madrid, España

Mostaza D (2019) Estudio de la relación entre las aguas superficiales y subterráneas de la Masa de Agua Subterránea (MAS) 030.007 "Aluviales: Jarama-Tajuña”. Ph.D. Thesis, Universidad Rey Juan Carlos

Mostaza-Colado D, Carreño-Conde F, Rasines-Ladero R, Iepure S (2018) Hydrogeochemical characterization of a shallow alluvial aquifer: 1 baseline for groundwater quality assessment and resource management. Science of The Total Environment 639:1110–1125. https://doi.org/10.1016/j.scitotenv.2018.05.236

Neilson-Welch L, Allen D (2007) GROUNDWATER AND HYDROGEOLOGICAL CONDITIONS IN THE OKANAGAN BASIN, BRITISH COLUMBIA A STATE-OF-THE-BASIN REPORT. Okanagan Basin Water Board For Objective 1 of the Phase 2 Groundwater Supply and Demand Project, Simon Fraser University

Ouedraogo I, Defourny P, Vanclooster M (2019) Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeol J 27:1081–1098. https://doi.org/10.1007/s10040-018-1900-5

Panagopoulos GP, Antonakos AK, Lambrakis NJ (2006) Optimization of the DRASTIC method for groundwater vulnerability assessment via the use of simple statistical methods and GIS. Hydrogeol J 14:894–911. https://doi.org/10.1007/s10040-005-0008-x

Pathak DR, Hiratsuka A (2011) An integrated GIS based fuzzy pattern recognition model to compute groundwater vulnerability index for decision making. Journal of Hydro-environment Research 5:63–77. https://doi.org/10.1016/j.jher.2009.10.015

Rahmani B, Javadi S, Shahdany SMH (2019) Evaluation of aquifer vulnerability using PCA technique and various clustering methods. Geocarto International 36:2117–2140. https://doi.org/10.1080/10106049.2019.1690057

Rupert MG (2001) Calibration of the DRASTIC Ground Water Vulnerability Mapping Method. Ground Water 39:625–630. https://doi.org/10.1111/j.1745-6584.2001.tb02350.x

Saatsaz M, Eslamian S, Javadi S (2013) Development of a coupled flow and solute transport modelling for Astaneh-Kouchesfahan groundwater resources, North of Iran. International Journal of Water 7:80. https://doi.org/10.1504/IJW.2013.051980

Salazar C, Del Castillo S (2018) FUNDAMENTOS BÁSICOS DE ESTADÍSTICA, 1st edn.

Sanders L (1998) A Manual of Field Hydrogeology. Prentice-Hall

SIAR (2019) Consulta de datos meteorológicos. In: Sistema de Información Agroclimática y de Regadíos . Ministerio de agricultura y pesca, alimentación y medio ambiente. http://www.siar.es

Smith L, Weathcraft J (1993) Groundwater Flow. Maidment, D.R. (Ed.)- Handbook of Hydrology. McGraw Hill.

Song F, Guo Z, Mei D (2010) Feature selection using principal component analysis. 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization. https://doi.org/DOI 10.1109/ICSEM.2010.14

Stempvoort DV, Ewert L, Wassenaar L (1993) AQUIFER VULNERABILITY INDEX: A GIS - COMPATIBLE METHOD FOR GROUNDWATER VULNERABILITY MAPPING. Canadian Water Resources Journal 18:25–37

Stumpp C, Żurek AJ, Wachniew P, et al (2016) A decision tree tool supporting the assessment of groundwater vulnerability. Environ Earth Sci 75:1057. https://doi.org/10.1007/s12665-016-5859-z

Tahmasebi P, Kamrava S, Bai T, Sahimi M (2020) Machine learning in geo- and environmental sciences: From small to large scale. Advances in Water Resources 142:103619. https://doi.org/10.1016/j.advwatres.2020.103619

United States Department of Agriculture (USDA) (2017) Soil Survey Manual (SSM) | NRCS Soils

Yang J, Tang Z, Jiao T, Malik Muhammad A (2017) Combining AHP and genetic algorithms approaches to modify DRASTIC model to assess groundwater vulnerability: a case study from Jianghan Plain, China. Environ Earth Sci 76:426. https://doi.org/10.1007/s12665-017-6759-6

Yoo K, Shukla SK, Ahn JJ, et al (2016) Decision tree-based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity. Journal of Cleaner Production 122:277–286. https://doi.org/10.1016/j.jclepro.2016.01.075