Stoch Environ Res Risk Assess (2013) 27:435–447
DOI 10.1007/s00477-012-0676-8

ORIGINAL PAPER

Groundwater quality assessment using data clustering based on hybrid Bayesian networks

Pedro A. Aguilera · Antonio Fernández · Rosa F. Ropero · Luís Molina

Received: 9 May 2012 / Accepted: 4 December 2012 / Published online: 21 December 2012

Abstract Bayesian networks (BNs) have become a standard in the field of Artificial Intelligence as a means of dealing with uncertainty and risk modelling. In recent years, there has been particular interest in the simultaneous use of continuous and discrete domains, obviating the need for discretization, using so-called hybrid BNs. In these hybrid environments, Mixtures of Truncated Exponentials (MTEs) provide a suitable solution for working without any restriction. The objective of this study is the assessment of groundwater quality through the design and application of a probabilistic clustering, based on hybrid Bayesian networks with MTEs. Firstly, the results obtained allows the differentiation of three groups of sampling points, indicating three different classes of groundwater quality. Secondly, the probability that a sampling point belongs to each cluster allows the uncertainty in the clusters to be assessed, as well as the risks associated in terms of water quality management. The methodology developed could be applied to other fields in environmental sciences.

P. A. Aguilera · R. F. Ropero
Informatics and Environment Laboratory, Department of Plant Biology and Ecology, University of Almería, Almería, Spain
e-mail: aguilera@ual.es

R. F. Ropero
e-mail: rfr723@alboran.ual.es

A. Fernández (✉)
Department of Statistics and Applied Mathematics, University of Almería, Almería, Spain
e-mail: afalvarez@ual.es

L. Molina
Department of Hydrogeology and Analytic Chemistry, University of Almería, Almería, Spain
e-mail: lmolina@ual.es

Keywords Hybrid Bayesian networks · Mixtures of truncated exponentials · Probabilistic data clustering · Groundwater quality

1 Introduction

Groundwater quality is very important for sustaining both natural ecosystems and human activities (Lischeid 2009; Papaioannou et al. 2010; García-Díaz 2011). With the aim of assessing groundwater quality, multivariate procedures, such as cluster analysis, have been applied to physico-chemical information obtained from monitoring programmes (Liu et al. 2011; Evin and Favre 2012; Ghorban 2012; Vousoughi et al. 2012; Wang and Jin 2012). Cluster analysis (Anderberg 1973; Jain et al. 1999) is a statistical technique that groups observations (sampling points) into clusters. Thus, sampling points with similar water quality can be grouped to optimize monitoring programmes (Atlas et al. 2011; Lu et al. 2011). However, when using these groups as part of a decision-making process, the uncertainty involved by including an observation into a group can not be quantified. In this context, managers have an increasing interest in the development of new operational tools to assess uncertainty and risk, which can facilitate the decision-making process (Refsgaard et al. 2007).

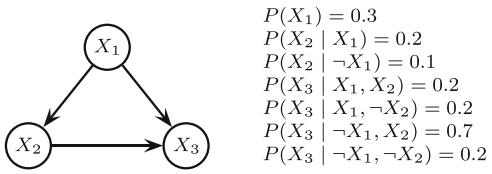
Bayesian networks (BNs) (Pearl 1988; Jensen and Nielsen 2007) are considered to be one of the most powerful tools for representing complex systems in which the relationships between variables are subject to uncertainty. Their main purpose is to provide a framework for efficient reasoning about the system they represent, in terms of updating information about unobserved variables, given that some new information is incorporated to the system (Jensen et al. 1990; Shenoy and Shafer 1990). Variables in

BNs are modelled by means of probability distributions; therefore risk and uncertainty can be estimated more accurately than in other models (Uusitalo 2007; Liao et al. 2010; Liu et al. 2012). BNs graphical interpretation allows stakeholders to easily understand the relationships between variables and refine the learned model manually adding or removing arcs (even variables) from the graph to better represent reality (Voinov and Bousquet 2010). Most data available in environmental sciences are continuous or hybrid (discrete and continuous), and even though BNs can manage them, the limitations are too restrictive in many cases (Nyberg et al. 2006). The most widely-used solution in environmental modelling is to discretise the variables, accepting a loss of information (Bromley et al. 2005; Uusitalo 2007). To date, several new solutions to this problem have been proposed, such as the conditional Gaussian (CG) model (Lauritzen 1992; Lauritzen and Jensen 2001), the mixture of truncated exponentials model (MTE) (Moral et al. 2001), the mixtures of polynomials model (MoP) (Shenoy and West 2011) and the mixtures of truncated basis functions (MoTBFs) model (Langseth et al. 2012).

Aguilera et al. (2011) reviewed the application of BNs in environmental modelling. Hybrid BNs have scarcely been applied in environmental modelling. There are few papers published concerning BNs in groundwaters and none of them use a solution based on hybrid BNs, but discretisation is applied. They are related to management and decision-making (Molina et al. (2009a, 2011); Carmona et al. 2011; Henriksen and Barlebo 2008; Henriksen et al. 2007; Santa Olalla et al. 2007; Santa Olalla et al. 2005), participative modelling (Martínez-Santos et al. 2010; Zorrilla et al. 2010) and prediction (Molina et al. 2009b).

BNs have been developed to resolve a wide variety of problems in the field of artificial intelligence (Larrañaga and Moral 2011). One of these is the so-called data clustering problem (Anderberg 1973; Jain et al. 1999), which is very useful in tasks such us pattern recognition or Mach Learn. Data clustering is understood to be a partition of a data set into groups in such a way that the individuals in one group are similar to each other but as different as possible from the individuals in other groups. BNs are valid tools for solving probabilistic clustering problems which, in contrast to traditional clustering, allow an individual to belong to more than one cluster depending on a probability distribution.

The aim of this article is to develop a probabilistic clustering model based on hybrid BNs that can be applied in the assessment of groundwater quality. To do this, inference is applied to a probability distribution of a data set. The probability distributions of the BN are modelled using MTEs, which means that there is no restriction on the model's structure, i.e., any combination of discrete and/or continuous nodes with discrete and/or continuous parents is allowed. In addition, continuous and discrete data can be used simultaneously without the need for any discretization.



(a) Qualitative component. **(b)** Quantitative component.

Fig. 1 An example of a Bayesian network with three variables

The article is organized as follows: Sect. 2 introduces the basic concepts about hybrid BNs and how they can be used to solve a probabilistic data clustering problem. Section 3 is dedicated to the application of the clustering model to management of groundwater quality. Lastly, Sect. 4 presents the most important conclusions drawn from the study.

2 Probabilistic clustering based on hybrid Bayesian networks

2.1 Bayesian networks

A Bayesian network (Jensen et al. 1990; Shenoy and Shafer 1990) is a statistical multivariate model for a set of variables $\mathbf{X}$[1], which is defined in terms of two components:

- A qualitative component, defined by means of a directed acyclic graph (DAG), in which each vertex represents one of the variables in the model, so that the presence of an arc linking two variables indicates the existence of statistical dependence between them. For example, the graph depicted in Fig. 1(a) could be the qualitative component of a BN for variables X_1, X_2 and X_3.
- A quantitative component, specified using a conditional distribution $p(x_i|\ \mathrm{pa}(x_i))$ for each variable X_i, $i = 1, \ldots, n$ given its parents in the graph, denoted as $\mathrm{pa}(X_i)$. Figure 1b shows an example of the conditional distributions $p(x_1)$, $p(x_2|x_1)$ and $p(x_3|x_1,x_2)$ for the DAG in Fig. 1a.

The success of BNs stems from the fact that the DAG structure gives us information about which variables are relevant or irrelevant for some other variable of interest,

[1] Uppercase letters denote random variables and boldfaced uppercase letters denote random vectors, e.g. $\mathbf{X} = \{X_1, \ldots, X_n\}$. The domain of $\mathbf{X}$ is denote as $\Omega_{\mathbf{X}}$ By lowercase letters x (or $\mathbf{x}$) we denote some element of Ω_X (or $\Omega_{\mathbf{X}}$).

taking into account the d-separation concept (Jensen and Nielsen 2007). This allows us to simplify, to a significant extent, the joint probability distribution (JPD) of the variables necessary to specify the model. In other words, BNs provide a compact representation of the JPD over all the variables, defined as the product of the conditional distributions attached to each node, so that

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}(x_i)). \quad (1)$$

For instance, the JPD associated to the network in Fig. 1, $p(x_1, x_2, x_3)$, is simplified as the product $p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2)$.

There are two approaches to training a BN: automatic and manual (or a mixture of the two). The first approach involves using algorithms which, starting with a set of training data, calculate the optimum structure for these data (Spirtes et al. 1993; Cooper and Herskovits 1992). From here, the corresponding probability distributions are calculated. In contrast, using manual approximation, expert opinion is included as part of the process to indicate which variables are related and how strongly. This second option is often used when no training data are available or where some data are missing.

A BN can carry out an efficient reasoning for a given scenario under conditions of uncertainty. This is what is known as probability propagation or probabilistic inference. Hence, the objective is to obtain information about a set of variables of interest (unobserved variables) given known values of other variables (observed or evidenced variables). If we denote the set of evidence as $\mathbf{E}$, and its values as $\mathbf{e}$, then we can calculate the posterior probability distribution, $p(x_i|\mathbf{e})$, for each variable of interest $X_i \notin \mathbf{E}$.

2.2 Hybrid Bayesian networks based on the MTE model

BNs were originally proposed for handling discrete variables and nowadays, a broad and consolidated theory can be found in the literature (see for instance Jensen and Nielsen (2007)). However, in real problems, it is very common to find continuous and discrete domains simultaneously in so-called hybrid BNs.

In a hybrid framework, the simplest and the most common solution is to discretise the continuous data and treat them as if they were discrete. Thus, existing methods for discrete variables can be easily applied. However, discretisation of variables can lead to a loss in precision and this is why other approaches have received so much attention over the last few years.

So far, several approaches have been devised to represent probability distributions in hybrid BNs. In order of their appearance they are: the CG model (Lauritzen 1992; Lauritzen and Jensen 2001), the mixtures of truncated exponentials (MTEs) model (Moral et al. 2001), the mixtures of polynomials (MOPs) model (Shenoy and West 2011) and the mixtures of truncated basis functions (MoTBFs) model (Langseth et al. 2012).

Although the CG model is used extensively by researchers and works well in many cases, it puts some restrictions on the network. It is only useful in situations where the joint distribution of the continuous variables, for each configuration of the discrete ones, follows a multivariate Gaussian. Moreover, CG models are not valid in frameworks where a discrete variable has continuous parents.

Discretisation is equivalent to approximating a density by a mixture of uniforms, meaning that each interval is approximated by a constant function. Thus, the accuracy of the final model could be increased if, instead of constants, other functions with better fitting properties were used. A good choice are exponential functions since they are closed under restriction, marginalisation and combination. This is the idea behind the so-called MTE model (Moral et al. 2001), explained next.

During the probability inference process, when the posterior distributions of the variables are obtained given some evidence, the intermediate probability functions are not necessarily density functions. Therefore, a general function called MTE potential needs to be defined as follows:

Definition 1 (MTE potential) Let $\mathbf{X}$ be a mixed n-dimensional random vector of variables. Let $\mathbf{Z} = (Z_1, \ldots, Z_d)^{\mathsf{T}}$ and $\mathbf{Y} = (Y_1, \ldots, Y_c)^{\mathsf{T}}$ be the discrete and continuous parts of $\mathbf{X}$, respectively, with $c + d = n$. We say that a function $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ is a mixture of truncated exponentials potential (MTE potential) if one of the following conditions holds:

(i) $\mathbf{Z} = \emptyset$ and f can be written as

$$f(\mathbf{x}) = f(\mathbf{y}) = a_0 + \sum_{i=1}^{m} a_i \exp\{\mathbf{b}_i^{\mathsf{T}} \mathbf{y}\} \quad (2)$$

for all $\mathbf{y} \in \Omega_{\mathbf{Y}}$, where $a_i \in \mathbb{R}$ and $\mathbf{b}_i \in \mathbb{R}^c$, $i = 1, \ldots, m$.

(ii) $\mathbf{Z} = \emptyset$ and there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Y}}$ into hypercubes such that f is defined as

$$f(\mathbf{x}) = f(\mathbf{y}) = f_i(\mathbf{y}) \quad \text{if } \mathbf{y} \in D_i,$$

where each f_i, $i = 1, \ldots, k$ can be written in the form of Eq. 2.

(iii) $\mathbf{Z} \neq \emptyset$ and for each fixed value $\mathbf{z} \in \Omega_{\mathbf{Z}}$, $f_{\mathbf{z}}(\mathbf{y}) = f(\mathbf{z}, \mathbf{y})$ can be defined as in (ii).

For example, the function f defined as

$$f(y_1,y_2)=\begin{cases}2+\mathrm{e}^{3y_1+y_2}+\mathrm{e}^{y_1+y_2} & \text{if } 0\leq;y_1\leq 1, 0\leq;y_2\leq;2,\\ 1+\mathrm{e}^{y_1+y_2} & \text{if } 0\leq;y_1\leq 1, 2\leq y_2\leq;3,\\ \frac{1}{4}+\mathrm{e}^{2y_1+y_2} & \text{if } 1\leq;y_1\leq;2, 0\leq;y_2\leq;2,\\ \frac{1}{2}+5\mathrm{e}^{y_1+2y_2} & \text{if } 1\leq;y_1\leq;2, 2\leq y_2\leq;3.\end{cases}$$

is an MTE potential since all of its parts are MTE potentials.

Thus, in this hybrid framework an MTE potential f is an MTE density if

$$\sum_{\mathbf{z}\in\Omega_{\mathbf{Z}}}\int_{\Omega_{\mathbf{Y}}} f(\mathbf{z},\mathbf{y})d\mathbf{y}=1.$$

A conditional MTE density can be specified by dividing the domain of the conditioning variables and specifying an MTE density for the conditioned variable for each configuration of splits of the conditioning variables.

Consider the following example. Let X and Y be two continuous variables. A possible conditional MTE density for Y given X is:

$$f(y|x)=\begin{cases}1.26-1.15\mathrm{e}^{0.006y} & \text{if } 0.4\leq x\leq;5,\ 0\leq y\leq;13,\\ 1.18-1.16\mathrm{e}^{0.0002y} & \text{if } 0.4\leq x\leq;5,\ 13\leq y\leq;43,\\ 0.07-0.03\mathrm{e}^{-0.4y}+0.0001\mathrm{e}^{0.0004y} & \text{if } 5\leq x\leq;19,\ 0\leq y\leq;5,\\ -0.99+1.03\mathrm{e}^{0.001y} & \text{if } 5\leq x\leq;19,\ 5\leq y\leq;43.\end{cases}$$

Since MTEs are defined into hypercubes, they admit a tree-structured representation in a natural way. Moral et al. (2001) proposed a data structure to represent MTE potentials, the so-called *mixed probability trees* or mixed trees for short which are specially appropriate for this kind of conditional densities.

In a similar way to the discretisation process, the more intervals used to divide the domain of the continuous variables, the better the MTE model accuracy, but also the more complex it becomes. Furthermore, in the case of MTEs, using more exponential terms within each interval substantially improves the fit to the real model, but again more complexity is assumed.

The MTE model has been the main focus of research for several years by the Laboratory of Probabilistic Graphical Models group[2] and it forms the basis of the clustering presented in Sect. 2.3. For more details about learning and inference tasks in these models, see Moral et al. (2001, 2002, 2003), Rumí et al. (2006), Rumí and Salmerón (2007), Romero et al. (2006), Cobb and Shenoy (2006), Cobb et al. (2007), Morales et al. (2007), Fernández et al. (2010), Langseth et al. (2009), Langseth et al. (2010), Aguilera et al. (2010) and Fernández et al. (2012).

The last two approaches, dealing with hybrid BNs (MOPs and MoTBFs) are very recent. The idea behind the MOPs (Shenoy and West 2011) model is to replace the basis function of the MTE (exponential) by a polynomial, yielding several advantages. The MoTBFs (Langseth et al. 2012) imply a generalisation of the MTEs and MOPs in the sense that any function can be used as a basis to represent the potentials. We do not use any of these approaches since they are still the subject of research and so there is not yet any software available.

2.3 Bayesian networks for clustering

In the context of Mach Learn, there are two types of classification algorithms: *supervised* and *unsupervised*. Let $\mathcal{D}=\{\mathbf{d}_1,\ldots,\mathbf{d}_k\}$ be a set of instances where $\mathbf{d}_i=\{x_{i1},\ldots,x_{in},c_i\}$ are the values for the ith-individual with features $X_1,\ldots,X_n$ and target variable C. Supervised classification involves inferring a function, f, such that $f(x_1,\ldots,x_n)$ gives us information about the best class state c for an individual $x_1,\ldots,x_n$. On the other hand, if data about C are missing, we start from a collection of unlabelled data and the classification problem becomes unsupervised.

Clustering (Anderberg 1973; Jain et al. 1999), or unsupervised classification, is understood to be the partition of a data set into groups in such a way that individuals in one group are similar to each other but as different as possible from individuals in other groups. Different types of clustering algorithms can be found in the literature depending on the approach they follow. On one hand, there is a *hard* clustering, in which clusters are exclusive, i.e., an individual belongs to a cluster in a deterministic way. The second approach is *soft* clustering or probabilistic clustering, meaning that an individual can belong to more than one cluster depending on a probability distribution. BNs can solve a probabilistic clustering problem by performing inference on the model, as explained next.

[2] http://elvira.ual.es/programo

Since unsupervised classification (or clustering) is mainly based on supervised classification, let us first explain how to carry out supervised classification based on hybrid BNs.

A BN can be used for supervised classification if it contains a class variable C, and a set of feature variables $X_1, \ldots, X_n$ where an individual with observed features $x_1, \ldots, x_n$ will be classified as belonging to class c^* obtained as follows:

$$c^* = \arg\max_{c \in \Omega_C} f(c|x_1, \ldots, x_n), \quad (3)$$

where Ω_C denotes the set of possible values of C.

Note that $f(c|x_1, \ldots, x_n)$ is proportional to $f(c) \times f(x_1, \ldots, x_n|c)$, and therefore, solving the classification problem would require a distribution to be specified over the n feature variables for each value of the class. The associated computational cost can be very high. However, using the factorisation determined by the network, the cost is reduced. Although the ideal would be to build a network without restrictions on the structure, usually this is not possible due to the limited data available. Therefore, networks with fixed and simpler structures and specifically designed for classification are used.

The extreme case is the so-called *naïve Bayes* (NB) structure (Duda et al. 2001; Friedman et al. 1997). It consists of a BN with a single root node and a set of attributes having only one parent (the root node). The NB model structure is shown in Fig. 2.

Its name comes from the naïve assumption that the feature variables $X_1, \ldots, X_n$ are considered independent given C. This strong independence assumption is somehow compensated by the reduction in the number of parameters to be estimated from data, since in this case, it holds that

$$f(c|x_1, \ldots, x_n) \propto f(c) \prod_{i=1}^{n} f(x_i|c), \quad (4)$$

which means that, instead of one n-dimensional conditional distribution, n one-dimensional conditional distributions are estimated. Despite this extreme independence assumption, the results are amazing in many cases, and for this reason it has become the most widely used Bayesian classifier in the literature.

Unsupervised classification or data clustering is performed in a similar way as for supervised classification. The only difference stems from the fact that, since there is no information about the target variable C, certain

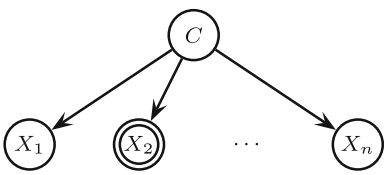


Fig. 2 Structure of a hybrid *naïve Bayes* classifier where X_2 is a continuous variables and the remaining ones are discrete

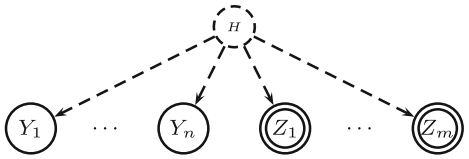


Fig. 3 Hybrid naïve Bayes classifier for probabilistic data clustering, where **Y** and **Z** are the set of discrete and continuous variables, respectively, and H the hidden variable useful for the clustering

considerations have to be taken into account when training the model. The key idea is to consider a hidden variable H as part of the dataset, that is, a variable whose values are missing in all the records. Thus, using an iterative process based on *data augmentation* (Tanner and Wong 1987), a model is built for a specific number of clusters. The iterative process includes two essential steps that are repeated until the probability of the model no longer improves:

1. For each ith-individual $x_{i1}, \ldots, x_{in}$ in the data set, a value is simulated for h_i from the posterior distribution $f(h|x_{i1}, \ldots, x_{in})$.
2. The probability distributions of the BN are re-learnt using the newly-generated database.

Figure 3 shows the model for carrying out a probabilistic clustering based on the hybrid naïve Bayes structure. Note that both continuous and discrete features are allowed in the model. The variable H must be discrete where its states represent the clusters obtained. The specific steps for building this model are detailed in Sect. 3.1.3.

3 Application in the assessment of groundwater quality

3.1 Methodology

3.1.1 Study area

The Campo de Dalías is located in the far southeastern end of Andalusia (Spain), covering around 330 km^2 (Fig. 4). It is bounded to the north by the Sierra de Gádor and to the south by the Mediterranean Sea. Its climate, together with technological innovations, have allowed the development of intensive agriculture in plastic-covered greenhouses. The cultivated area is approximately 20,000 hectares, which represents the largest cultivated area under greenhouse cover in Europe. Water for crop irrigation and for human consumption comes mostly from groundwater abstractions.

The study area can be differentiated into three hydrogeological units (Pulido-Bosch et al. 1991; Molina 1998): Balerma-Las Marinas, Balanegra and Aguadulce. The Balanegra unit occupies the western part, while the Aguadulce unit is to the east. Both these basically consist of carbonate deposits that form part of the Gádor nappe. The Balerma-

Fig. 4 Study area

Las Marinas unit is the largest and occupies the central-southern portion of the area. It is basically made up of Pliocene calcarenites that can exceed 100 m in thickness, though there are local Quaternary deposits as well.

The largest abstractions are made from the carbonate deposits of the Balanegra and Aguadulce units, given their calcium-magnesium bicarbonate water type. Accordingly, piezometric levels currently lie between −31 and −17 m a.s.l. In the Balerma-Marinas unit, the water facies is sodium-chloride and so abstractions are much lower. Since many wells have been abandoned as a result, the piezometric level over the entire unit is positive (10 and 40 m a.s.l). Although under a natural regime the hydraulic relationships between these three units would have been close, their subsequent exploitation means that they are now quite well individualized.

3.1.2 Monitoring and water analysis

A total of 125 wells (sampling points) were chosen, their distribution being representative of the three Campo de Dalías hydrogeological units. Water samples were taken according to the criteria given by the Enviromental Protection Agency (EPA 1991) and analysed for electrical conductivity, nitrate, Cu, Fe and pesticides. Conductivity was measured in situ using a WTW MultiLine P4 digital pH-conductivity meter. Nitrate was determined using ion chromatography, Cu and Fe, using atomic absorption spectroscopy, while pesticides were analysed using gas chromatography.

3.1.3 Data clustering methodology

This section describes the methodology for constructing a probabilistic clustering model based on a groundwater data set, and the strategy devised to find the optimal number of clusters.

Algorithm 1 (Gámez et al. 2006; Fernández et al. 2011) shows the steps for carrying out a probabilistic data clustering based on hybrid BNs using groundwater samples. Algorithms 2, 3 and 4 are subroutines of Algorithm 1 and they are shown in boldface. The algorithms were implemented in Elvira software (Elvira-Consortium 2002).

At the beginning of Algorithm 1, we have data only for the five physico-chemical variables but no information about the hidden variable H is available (i.e. the number of clusters and the associated probability distribution are still unknown). Therefore, the first task was to construct a preliminary model according to Algorithm 2, where the conditional MTE distributions for the variables are approximated by the marginal MTE distribution learnt directly from data (see Rumí et al. 2006 for more details). On the other hand, the initial number of clusters in H are fixed to two and their probabilities are equitatively initialised to 0.5. The results may depend on the initial model selected as this algorithm ensures finding a local maximum, but not the global one. Anyway, the clusters obtained are consistent with the expected results from an environmental point of view. An attempt to solve this issue would be to run the algorithm several times using different initializations, although this solution does not guarantee the global maximum either.

Once created, the initial model is refined using the *data augmentation* method (Tanner and Wong 1987) (see Algorithm 3). This method returns the most likely model with two clusters. In order to run this method, the missing data corresponding to the hidden variable H is initialised using zeros (step 4 in Algorithm 1). In a similar way, in step 6 we impute values for the hidden variable simulating them from the posterior distribution of H after propagating the values for the physico-chemical variables in the model, i.e., from $f(h|\mathbf{d}_i)$ (this imputation is needed for the following steps).

From this point, the idea is to create a new model by adding a cluster (see Algorithm 4). In this task the last state, h_n, of H is split to create a new one, h_{n+1}, and the new distributions generated are approximated from the current ones. The results may be slightly influenced by the choice of the state to be split, but we can not have this information a priori. The optimal solution would be to check all the states to find the optimal solution, but we did not consider this option because it adds complexity to the procedure that is disproportionate to the benefits in terms of accuracy.

After adding a cluster, the *data augmentation* method is run again to refine the new model iteratively using $n + 1$ clusters. Then, using the probability measure described below, we checked if the new model is an improvement over the earlier one. Assuming a data set of n independent and identically distributed observations for testing the model $\mathcal{T} = \{X^{(1)}, \ldots, X^{(n)}\}$, then the log-likelihood[3] of a model $\mathcal{M}$ according to this test set is defined as:

$$\mathcal{L}(\mathcal{M}|\mathcal{T}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log P_{\mathcal{M}}(X_j^{(i)}|\mathrm{pa}(X_j)^{(i)}), \tag{5}$$

where i and j index over the instances and nodes in the model respectively, $X_j^{(i)}$ is the value for the j-variable in the i-instance and $\mathrm{pa}(X_j)^{(i)}$ are the values for X_j's parents in the i-instance.

[3] For math convenience the logarithm of the likelihood is computed instead.

The process is repeated until the log-likelihood of the model for $n + 1$ clusters does not improve the earlier model containing n clusters, so that n, the optimal number of clusters, is finally determined.

Once the training stage has finished, the model in Fig. 5 is reported. It is then applied to perform the data clustering. Thus, an individual $(x_1, \ldots, x_n)$ will belong to the cluster c^* according to Eq. 3. In this way, an individual can belong to more than one cluster depending on the probability distribution. This feature of fuzzy problems is particularly interesting in the environmental sciences, in particular, in the assessment of groundwater quality.

Algorithm 1: Probabilistic clustering based on hybrid Bayesian networks for the groundwaters data set

Input: The data set containing 125 samples for the five physico-chemical variables $\mathbf{X} = \{\texttt{Conductivity}, \texttt{NO}_3, \texttt{Cu}, \texttt{Fe}, \texttt{Pesticides}\}$.
Output: A model M to carry out the clustering

Divide the data set into two parts randomly: *train* (80%) and *test* (20%).
$M_0 \leftarrow$ **LearnInitialModel** (*train*) (see Algorithm 2).
Lets denote the `cluster` variable as H (hidden).
Add a data column H with 0s to the *train* data set.
$M_0 \leftarrow$ **DataAugmentation** (M_0, *train*) (see Algorithm 3).
Add a hidden variable H to *test* by simulating values in M_0 from $f(h \mid \mathbf{d}_i)$.
$L_0 \leftarrow$ log-likelihood (M_0, *test*) (see Eq. 5).
repeat
  $L \leftarrow L_0$.
  $M_0 \leftarrow$ **AddCluster** (M_0) (see Algorithm 4).
  $M_0 \leftarrow$ **DataAugmentation** (M_0, *train*) (see Algorithm 3).
  Update H in *test* by simulating values in M_0 from $f(h \mid \mathbf{d}_i)$.
  $L_0 \leftarrow$ log-likelihood (M_0, *test*) (see Eq. 5).
  if $L_0 > L$ **then**
    $M \leftarrow M_0$.
until $L_0 < L$;
return M.

Algorithm 2: LearnInitialModel

Input: The *train* database with variables $\mathbf{X}$.
Output: A model M with variables $\mathbf{X}$ and a hidden one H.

foreach X_i *in* $\mathbf{X}$ **do**
  Learn an MTE potential $f(x_i)$ from *train* (see Rumí et al (2006)).
  $f(x_i \mid h) \leftarrow f(x_i)$.
$P(h_0) \leftarrow 0.5$.
$P(h_1) \leftarrow 0.5$.
Let M a naïve Bayes model with distributions $P(H)$ and $f(x_i \mid h), \forall X_i \in \mathbf{X}$.
return M.

3.2 Results

3.2.1 Data clustering

The results obtained from applying Algorithm 1 allow the sampling points to be grouped into three clusters (Fig. 6). Table 1 shows the log-likelihood values for the different number of clusters (from 2 to 6) according to Eq. 5. The entry in boldface indicates the optimal log-likelihood value which is reached with three clusters. Despite the fact that the algorithm stops when this measure does not improve with respect to the previous iteration, we forced the algorithm to run up to six clusters just to investigate the behaviour of the algorithm beyond. As shown, the log-likelihood decreases as the number of clusters increases, meaning that the inclusion of new clusters yields less accurate models in this case.

The average values of the physico-chemical variables for each of the clusters are presented in Table 2.

Algorithm 3: DataAugmentation

Input: A model M_0 with hidden variable H and a database D.
Output: The new model M after refining it.

Divide D into two datasets: *train* and *test*.
$L_0 \leftarrow$ log-likelihood (M_0, *test*) (see Eq. 5).
$M \leftarrow M_0$.
repeat
  $L \leftarrow L_0$
  Update H in *train* and *test* by simulating values in M_0 from $f(h \mid \mathbf{d}_i)$.
  $M_0 \leftarrow$ Learn a new model (*train*).
  $L_0 \leftarrow$ log-likelihood (M_0, *test*) (see Eq. 5).
  if $L_0 > L$ **then**
    $M \leftarrow M_0$.
until $L_0 < L$;
return M.

Algorithm 4: AddCluster

Input: A model M_0 with n states in the hidden variable H_0.
Output: A new model M with $n + 1$ states in the hidden variable H.

$M \leftarrow M_0$.
Let $h_1, \ldots, h_n$ be the states of the hidden variable H in M.
Add a new state, h_{n+1} to H.
Update the probability distribution of H by re-computing the probability of h_n and h_{n+1} as follows:
  $a \leftarrow p(h_n)/2$.
  $p(h_n) \leftarrow a$.
$p(h_{n+1}) \leftarrow a$.
foreach *feature* X_i *in* M **do**
  $f(x_i \mid h_{n+1}) \leftarrow f(x_i \mid h_n)$.
return M.

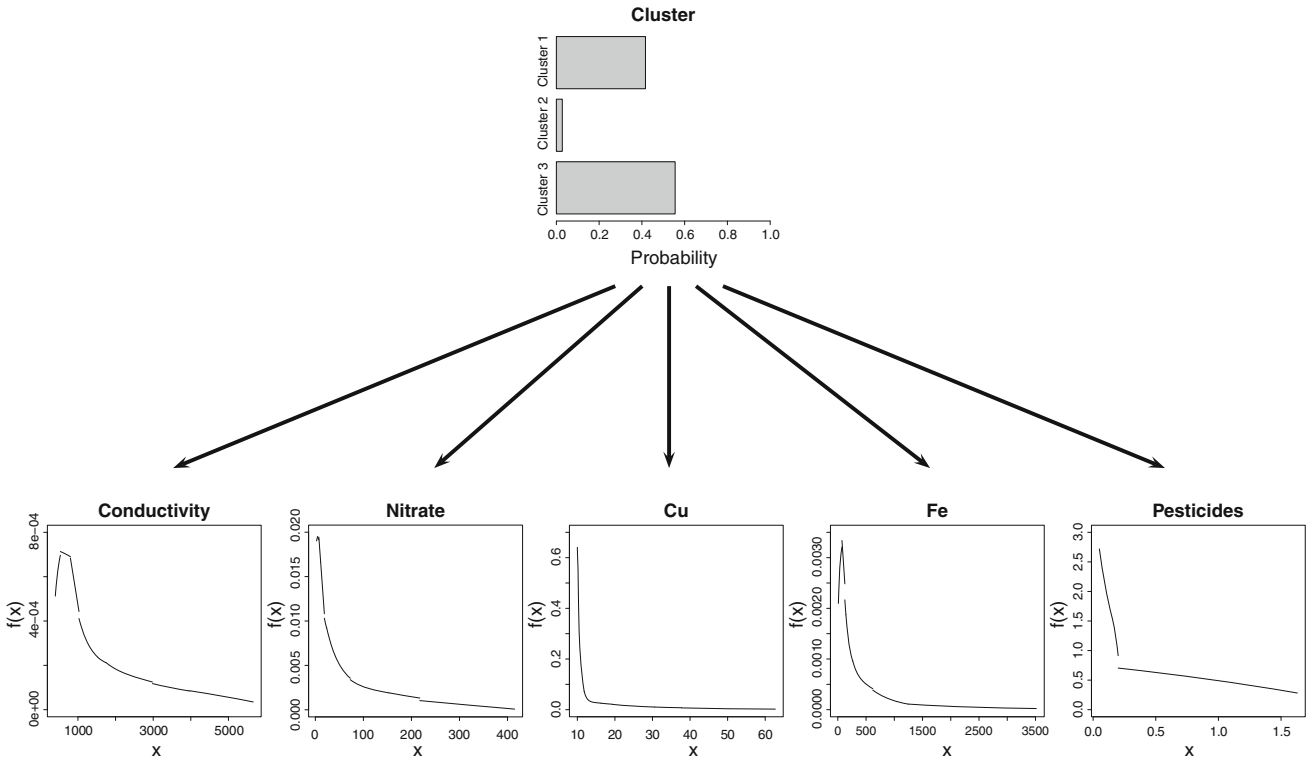


Fig. 5 Probabilistic data clustering model with the marginal MTE probability distributions for the variables. Note that the features are continuous, whilst the cluster variable is discrete. The marginal probability for the cluster variable represents the frequency (in terms of probability) of samples assigned to each cluster after running the algorithm (the marginal for the physico-chemical variables has similar interpretation). The number of states of the cluster variable corresponds to the optimum number of clusters found by Algorithm 1

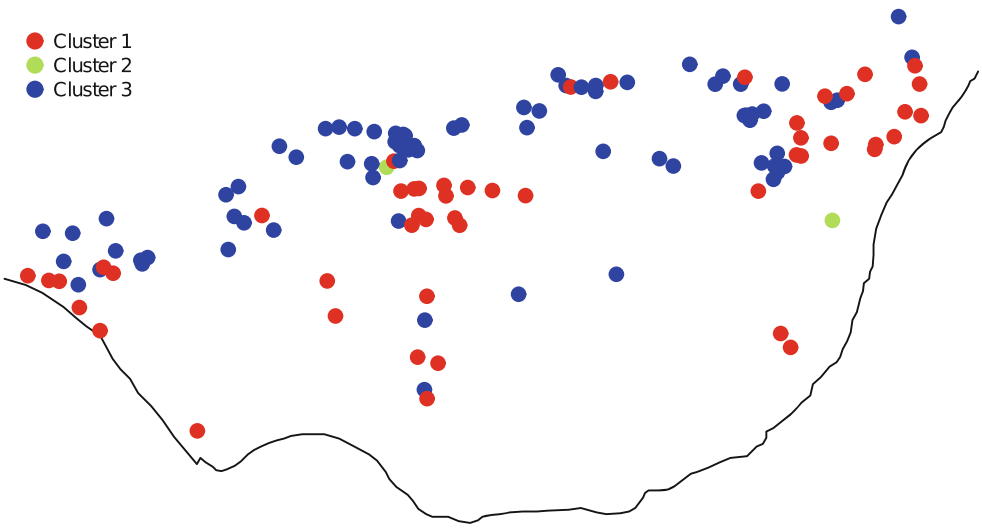


Fig. 6 Assignment of sampling points to its most probable cluster. (Color figure online)

Group 1 comprises 50 sampling points. The average probability of these points belonging to this group is 0.92, with a standard deviation of 0.13 (Table 3). The sampling points are situated over Pliocene calcarenites and Quaternary detritic deposits that form the uppermost part of the aquifer. The surface wells have a depth of between 20 and 150 m. Dissolution of deposits that overlie the sampling points mean that samples have a high conductivity. In turn, elevated nitrates, Cu and Fe are the result of the proximity of the phreatic level to the ground surface, which facilitates entry of these agricultural pollutants into the groundwater (Molina 1998). In this cluster, 18 % of the samples contain

Table 1 Accuracy in terms of log-likelihood for different clustering models depending on the number of clusters assigned. The entry in boldface indicates the optimal number of clusters

# Clusters	Log-likelihood
2	−888.6936
3	**−857.4265**
4	−861.5178
5	−864.0865
6	−873.5870

Table 3 Minimum, average and maximum probability, and standard deviation for the sampling points in clusters 1 and 3. The values for cluster 2 are not shown since, given only two sampling points, their statistical significance is meaningless

	# Sampling points	Minimum	Maximum	Average	SD
Cluster 1	50	0.55	0.99	0.92	0.13
Cluster 3	73	0.54	0.99	0.95	0.09

pesticides. The elevated contaminant concentrations are the reason that these waters are used neither for human consumption nor agricultural irrigation.

Group 2 consists of just two sampling points, with probabilities of belonging of 0.97 and 0.75. These boreholes are located in Plioquaternary deposits at depths of 30 and 80 m. The deeper sampling site gave a conductivity of 2,180 μmhos/cm, nitrate content of 124 mg/l, Fe of 182 μg/l, Cu of 16 μg/l, and a high pesticide content (1.63 μg/l). This sampling point is positioned between the calcarenites and the limestones in an abandoned borehole. The other, shallower sampling point gave a conductivity of 4,070 μmhos/cm, nitrate of 415 mg/l, Fe of 337 μg/l, Cu of 22 μg/l with presence of pesticides as well (0.28 μg/l). This sampling point is located over calcarenites at the eastern end of the study area and has been polluted due to the intensive agricultural activities in the vicinity; for this reason it has been abandoned for some time.

Group 3 is formed by 73 sampling points. The average probability of belonging to this group is 0.95, with standard deviation 0.09 (Table 3). This group is characterized by the fact that all the boreholes abstract water from the limestones and dolomites of the Balanegra and Aguadulce units that lie along the southern edge of the Sierra de Gádor. The depth of the boreholes along the edge of the Sierra is between 200 and 300 m. The aquifer gets steadily deeper as a result of a number of fractures, towards the centre of the area, reaching a depth of up to 900 m. The marked depth of these boreholes favours abstraction of better quality water, except at the eastern and western flanks where marine intrusion intervenes (Pérez-Parra et al. 2007). In this group only two sampling points indicate the presence of pesticides, for which reason neither of them are exploited.

3.2.2 *Uncertainty, risk and probabilistic clustering*

In addition to the data clustering, Algorithm 1 provides information about the probability that a certain sampling point belongs to a particular cluster (Fig. 7). This information allows the uncertainty and risk in the groundwater quality management to be assessed.

Thus, in group 1 (Fig. 8a), 36 of the sampling points have a probability greater than 0.95 of belonging to this group. Eight of the observations have a probability of belonging to the group of between 0.70 and 0.95, while six sampling points give a value lower than 0.70.

In group 3 (Fig. 8b), 57 sampling points have a probability greater than 0.95 of belonging to the group, 14 points have a probability of between 0.70 and 0.95, while two points have a probability lower than 0.70.

These data suggest that 72 and 78.1 %, respectively, of the sampling points in these two groups show a high degree of certainty of belonging to its group (probability $\geq$0.95), i.e., there is a lower risk that the water quality of a well belongs to a different group than the one assigned using the BN clustering. As a consequence, these sampling points could be used as reference observations during classification of groundwater quality or during groundwater monitoring programmes.

Moreover, for another series of sampling points (Table 4; Fig. 9), there is greater uncertainty (probability $<$0.70) that the points belong to the assigned groups. These points share physico-chemical and hydrogeological characteristics from both groups 1 and 3.

Table 2 Average values of the physico-chemical parameters measured at the sampling points grouped in clusters 1, 2 and 3. Conductivity is expressed in μmhos/cm, nitrate in mg/l and Cu, Fe and pesticides in μg/l

	# Sampling points	Conductivity	Nitrate	Cu	Fe	Pesticides
Cluster 1	50	2,833	106.53	22.65	447.17	0.087
Cluster 2	2	3,125	269.4	19.15	260	0.955
Cluster 3	73	927	12.25	10.11	148.82	0.024

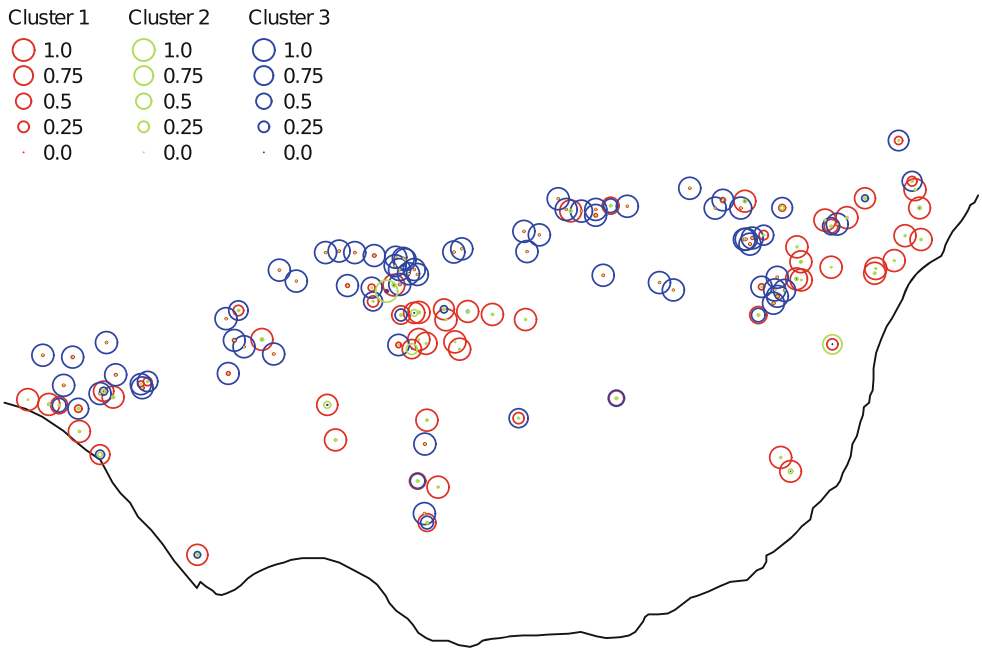
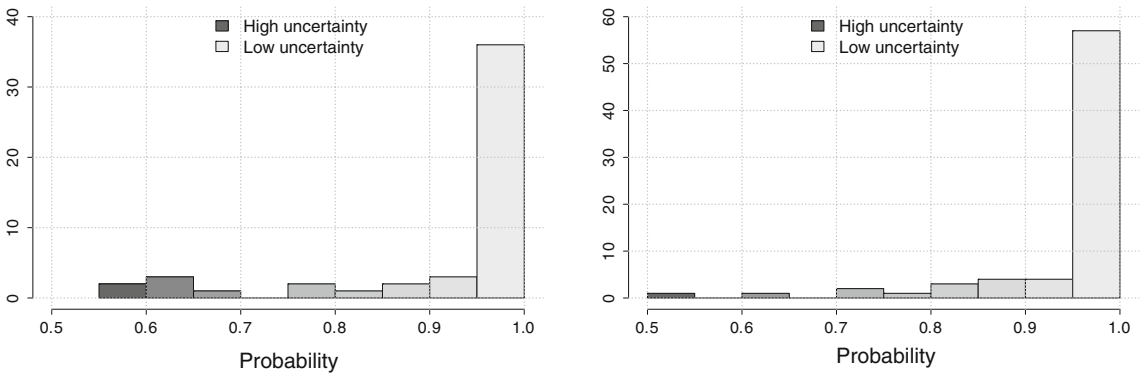


Fig. 7 Probability of a sampling point belonging to each of the clusters. For each sampling point, three concentric circles are shown on the map (using a different colour for each cluster), whose area is proportional to the probability of belonging to the corresponding cluster. (Color figure online)



(a) Cluster 1. **(b)** Cluster 3.

Fig. 8 Frequency histogram of the sampling points belonging to clusters 1 and 3 for various probability intervals. The higher the probability, the lower the uncertainty regarding assigning a sampling point to a particular cluster, and so the lower the risk of making a wrong decision about groundwater quality as part of the management process

Table 4 Sampling points with lower probabilities of belonging to groups 1 and 3

Sampling points	Cluster	Probability of belonging to cluster 1	Probability of belonging to cluster 3
65	1	0.60	0.40
394	1	0.69	0.31
839	1	0.60	0.40
1189	1	0.60	0.40
1201	1	0.65	0.45
1202	1	0.55	0.45
766	3	0.45	0.55
821	3	0.40	0.60

- Point 65 shares the high nitrate and iron contamination of group 1, and the low conductivity of group 3. It lies in the Pliocene calcarenites at a depth of 100 m.
- Point 394 shares the elevated concentrations of nitrate, iron and copper with group 1, and the low conductivity with group 3; it lies at a depth of 40 m in the Pliocene calcarenites.
- Point 839 has the high Fe of group 1 and the low conductivity of group 3, tapping limestones/dolomites at a depth of 600 m.
- Point 1189 shares the elevated Fe and Cu of group 1 and the low conductivity of group 3, it lies at 400 m depth in the Gádor limestones and dolomites.

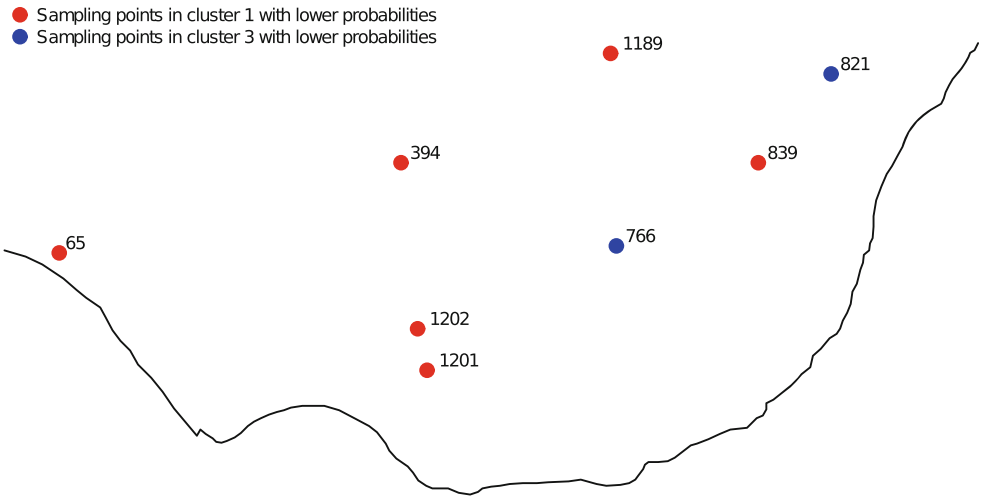


Fig. 9 Location of sampling points with lower probabilities of belonging to clusters 1 and 3. (Color figure online)

- Point 1201 shares the high Fe with group 1 and the low conductivity with group 3, again tapping the limestone/dolomite at a depth of 400 m.
- Point 1202 shares the Fe and Cu contamination with group 1 but has the low conductivity of group 3. It lies in the Sierra de Gádor limestones and dolomites at a depth of 900 m.
- Point 766 has high nitrates, Fe and Cu, as well as pesticides in common with group 1, with the low conductivity of group 3. It lies at a depth of 30 m in Quaternary deposits.
- Point 821 has elevated nitrates, Fe and Cu as in group 1, together with the low conductivity characteristic of group 3. It exploits the Pliocene calcarenites and lies at 200 m depth.

In traditional multivariate statistics, a cluster would normally allow the detection of sampling points with similar water qualities, i.e., groups of sampling points tapping groundwater with a homogeneous water quality. However, using these groups as part of a decision-making process does not take into account the uncertainty involved nor the risk of making the wrong decision in terms of water quality management.

In contrast, probabilistic clustering demonstrates that if all the sampling points assigned to each group are taken (as the traditional clustering methods do), we are committing the error of including sampling points that are not fully representative of a particular water quality.

The model developed here allows the uncertainty associated to be known and to determine, probabilistically, which sampling points should be chosen for groundwater monitoring programmes and conversely, which sampling points should be excluded on the basis of being less representative of a particular water quality.

4 Conclusions

This paper presents a novel technique for resolving the problem of probabilistic data clustering in the field of groundwater management. The probabilistic model based on MTEs allows simultaneous treatment of continuous and discrete variables without the need to discretise the data, thus increasing the precision of the modelling. Moreover, the grouping of sampling points using BNs allows optimization of the number of sampling points required for making an assessment of groundwater quality. It reduces the risk of wrong decisions being taken in the decision-making process by considering only those points that show higher probabilities of belonging to a particular group during the water quality monitoring programme. The technique of clustering presented in this article can be applied to any other field within the environmental sciences for risk assessment using probabilities, and thus contributes greater diversity to a field in which hybrid BNs were not previously applied.

Acknowledgments This work has been supported by the Spanish Ministry of Science and Innovation through project TIN2010-20900-C04-02, by the Regional Ministry of Economy, Innovation and Science (Junta de Andalucía) through project P08-RNM-03945 and by ERDF funds.

References

Aguilera PA, Fernández A, Reche F, Rumí R (2010) Hybrid Bayesian network classifiers: application to species distribution models. Environ Model Softw 25(12):1630–1639

Aguilera PA, Fernández A, Fernández R, Rumí R, Salmerón A (2011) Bayesian networks in environmental modelling. Environ Model Softw 26:1376–1388

Anderberg M (1973) Cluster analysis for applications. Academic Press, New York

Atlas L, Isik M, Kavurmaci M (2011) Determination of arsenic levels in the water resources of Aksaray Province, Turkey. J Environ Manag 92:2182–2192

Bromley J, Jackson NA, Clymer OJ, Giacomello AM, Jensen FV (2005) The use of Hugin^R to develop Bayesian networks as an aid to integrated water resource planning. Environ Model Softw 20:231–242

Carmona G, Varela-Ortega C, Bromley J (2011) The use of participatory object-oriented Bayesian networks and agro-economic models for groundwater management in Spain. Water Resourc Manag 25:1509–1524

Cobb BR, Shenoy PP (2006) Inference in hybrid Bayesian networks with mixtures of truncated exponentials. Int J Approx Reason 41:257–286

Cobb BR, Rumí R, Salmerón A (2007) Advances in probabilistic graphical models, chap Bayesian networks models with discrete and continuous variables. In: Studies in fuzziness and soft computing. Springer, Heidelberg, pp 81–102

Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Mach Learn 9:309–347

Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley Interscience, New York

Elvira-Consortium (2002) Elvira: an environment for probabilistic graphical models. In: Proceedings of the first European workshop on probabilistic graphical models (PGM'02). PGM, Cuenca, pp 222–230

EPA (1991) Compendium of ERT ground water sampling procedures. EPA 540/P-91-007. Technical Report Office, Washington, DC

Evin G, Favre AC (2012) Further developments of a transient Poisson-cluster model for rainfall. Stoch Environ Res Risk Assess. doi:10.1007/s00477-012-0612-y

Fernández A, Nielsen JD, Salmerón A (2010) Learning Bayesian networks for regression from incomplete databases. Int J Uncertain Fuzz Knowl Based Syst 18:69–86

Fernández A, Gámez JA, Rumí R, Salmerón A (2011) Data clustering using hidden variables in hybrid Bayesian networks. In: Book of abstracts of the 4th international conference of the ERCIM working group on computing & statistics (ERCIM'11). ERCIM, Senate House, p 19

Fernández A, Rumí R, Salmerón A (2012) Answering queries in hybrid Bayesian networks using importance sampling. Decis Support Syst 53:580–590

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163

Gámez JA, Rumí R, Salmerón A (2006) Unsupervised naïve Bayes for data clustering with mixtures of truncated exponentials. In: Proceedings of the 3rd European workshop on probabilistic graphical models (PGM'06). PGM, Granada, pp 123–132

García-Díaz JC (2011) Monitoring and forecasting nitrate concentration in the groundwater using statistical process control and time series analysis: a case study. Stoch Environ Res Risk Assess 25:331–339

Ghorban M (2012) Testing the weak stationarity of a spatial-temporal point process. Stoch Environ Res Risk Assess. doi:10.1007/s00477-012-0597-6

Henriksen HJ, Barlebo HC (2008) Reflections on the use of Bayesian belief networks for adaptive management. J Environ Manag 88:1025–1036

Henriksen HJ, Rasmussen P, Brandt G, von Bülow D, Jensen FV (2007) Public participation modelling using Bayesian networks in management of groundwater contamination. Environ Model Softw 22:1101–1113

Jain A, Murty M, Flynn P (1999) Data clustering: a review. ACM COmput Surv 31(3):264–323

Jensen FV, Nielsen TD (2007) Bayesian networks and decision graphs. Springer, New York

Jensen FV, Lauritzen SL, Olesen KG (1990) Bayesian updating in causal probabilistic networks by local computation. Comput Stat Q 4:269–282

Langseth H, Nielsen TD, Rumí R, Salmerón A (2009) Inference in hybrid Bayesian networks. Reliab Eng Syst Saf 94:1499–1509

Langseth H, Nielsen TD, Rumí R, Salmerón A (2010) Parameter estimation and model selection in mixtures of truncated exponentials. Int J Approx Reason 51:485–498

Langseth H, Nielsen TD, Rumí R, Salmerón A (2012) Mixtures of truncated basis functions. Int J Approx Reason 53(2):212–227

Larrañaga P, Moral S (2011) Probabilistic graphical models in artificial intelligence. Appl Soft Comput 11:1511–1528

Lauritzen SL (1992) Propagation of probabilities, means and variances in mixed graphical association models. J Am Stat Assoc 87:1098–1108

Lauritzen SL, Jensen F (2001) Stable local computation with conditional Gaussian distributions. Stat Comput 11:191–203

Liao Y, Wang J, Guo Y, Zheng X (2010) Risk assessment of human neural tube defects using a Bayesian belief network. Stoch Environ Res Risk Assess 24:93–100

Lischeid G (2009) Non-linear visualization and analysis of large water quality data sets: a model-free basis for efficient monitoring and risk assessment. Stoch Environ Res Risk Assess 23:977–990

Liu K, Lu C, Chen C, Shen Y (2012) Applying Bayesian belief networks to health risk assessment. Stoch Environ Res Risk Assess 26:451–465

Liu W, Yu H, Chung C (2011) Assessment of water quality in a subtropical Alpine lake using multivariate statistical techniques and geostatistical mapping: a case study. Int J Environ Res Public Health 8:1126–1140

Lu KL, Liu CW, Wang SW, Jang CS, Lin KH, Liao VHC, Liao CM, Chang FJ (2011) Assessing the characteristics of groundwater quality of arsenic contaminated aquifers in the blackfoot disease endemic area. J Hazard Mater 185:1458–1466

Martínez-Santos P, Henriksen HJ, Zorrilla P, Martínez-Alfaro PE (2010) Comparative reflections on the use of modelling tools in conflictive water management settings: the Mancha Occidental aquifer, Spain. Environ Model Softw 25:1439–1449

Molina J, García-Aróstegui J, Benavente J, Varela-Ortega C, Hera A, Lópe-Geta J (2009a) Aquifers overexploitation in SE Spain: a proposal for the integrated analysis of water management. Water Resour Manag 23:2737–2760

Molina J, Farmani R, Bromley J (2011) Aquifers management through evolutionary Bayesian networks: the Altiplano case study (Spain). Water Resour Manag 25:3883–3909

Molina JL, Bromley J, García-Aróstegui JL, Sullivan C, Benavente J (2009b) Integrated water resource management of overexploited hydrogeological systems using object-oriented Bayesian networks. Environ Model Softw 25:383–397

Molina L (1998) Hidroquímica e intrusión marina en el Campo de Dalías (Almería). PhD thesis, Universidad de Granada, Granada

Moral S, Rumí R, Salmerón A (2001) Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In: Benferhat S, Besnard P (eds) Symbolic and uantitative approaches to reasoning with uncertainty. Lecture notes in artificial intelligence, vol 2143. Springer, New York, pp 156–167

Moral S, Rumí R, Salmerón A (2002) Estimating mixtures of truncated exponentials from data. In: Gámez J, Salmerón A (eds) Proceedings of the first European workshop on probabilistic graphical models. PGM, Granada, pp 156–167

Moral S, Rumí R, Salmerón A (2003) Approximating conditional MTE distributions by means of mixed trees. In: Symbolic and quantitative approaches to reasoning with uncertainty. Lecture notes in artificial intelligence, vol 2711. Springer, New York, pp 197–183

Morales M, Rodríguez C, Salmerón A (2007) Selective naïve Bayes for regression using mixtures of truncated exponentials. Int J Uncertain Fuzz Knowl Based Syst 15:697–716

Nyberg JB, Marcot BG, Sulyma R (2006) Using Bayesian belief networks in adaptive management. Can J For Res 36:3104–3116

Papaioannou A, Dovriki E, Rigas N, Plageras P, Rigas I, Kokkora M, Papastergiou P (2010) Assessment and modelling of groundwater quality data by environmetric methods in the context of public health. Water Resourc Manag 24:3257–3278

Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan-Kaufmann, San Mateo

Pérez-Parra J, Molina L, Vallejos A, Danielle L, Zaragoza G, Pulido-Bosch A (2007) Los acuíferos costeros: retos y soluciones, IGME, Madrid, chap Evolución del estado de intrusión marina en el Campo de Dalías (Almería). IGME, Almería, pp 781–789

Pulido-Bosch A, Navarrete F, Molina L, Martínez-Vidal JL (1991) Quantity and quality of groundwater in the Campo de Dalías (Almería, SE Spain). Water Sci Technol 24:87–96

Refsgaard JC, van der Sluijis J, Hojberg A, Vanrolleghem PA (2007) Uncertainty in the environmental modelling process—a framework and guidance. Environ Model Softw 22:1543–1556

Romero V, Rumí R, Salmerón A (2006) Learning hybrid Bayesian networks using mixtures of truncated exponentials. Int J Approx Reason 42:54–68

Rumí R, Salmerón A (2007) Approximate probability propagation with mixtures of truncated exponentials. Int J Approx Reason 45:191–210

Rumí R, Salmerón A, Moral S (2006) Estimating mixtures of truncated exponentials in hybrid Bayesian networks. Test 15:397–421

Santa Olalla FM, Domínguez A, Artigao A, Fabeiro C, Ortega JF (2005) Integrated Water Resourc Manag of the hydrogeological unit "Eastern Mancha" using Bayesian belief networks. Agric Water Manag 77:21–36

Santa Olalla FM, Domínguez A, Ortega F, Artigao A, Fabeiro C (2007) Bayesian networks in planning a large aquifer in Eastern Mancha, Spain. Environ Model Softw 22:1089–1100

Shenoy PP, Shafer G (1990) Axioms for probability and belief functions propagation. In: Shachter R, Levitt T, Lemmer J, Kanal L (eds) Uncertainty in artificial intelligence, vol 4. North Holland, Amsterdam, pp 169–198

Shenoy PP, West JC (2011) Inference in hybrid Bayesian networks using mixtures of polynomials. Int J Approx Reason 52(5):641–657

Spirtes P, Glymour C, Scheines R (1993) Causation, prediction and search. Lecture notes in statistics, vol 81. Springer, New York

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Stat Assoc 82:528–550

Uusitalo L (2007) Advantages and challenges of Bayesian networks in environmental modelling. Ecol Model 203:312–318

Voinov A, Bousquet F (2010) Modelling with stakeholders. Environ Model Softw 24:1268–1281

Vousoughi F, Dinpashoh Y, Aalami MT, Jhajharia D (2012) Trend analysis of groundwater using non-parametric methods (case study: Ardabil plain). Stoch Environ Res Risk Assess. doi: 10.1007/s00477-012-0599-4

Wang H, Jin X (2012) Characterization of groundwater contaminant source using Bayesian method. Stoch Environ Res Risk Assess. doi:10.1007/s00477-012-0622-9

Zorrilla P, Carmona G, De la Hera A, Varela-Ortega C, Martínez-Santos P, Bromley J, Henriksen HJ (2010) Evaluation of Bayesian networks in participatory water resource management, Upper Guadiana Basin, Spain. Ecol Soc 15(3). http://www.ecologyandsociety.org/vol15/iss3/art12/