Continuous Bayesian networks for probabilistic environmental risk mapping

Ana D. Maldonado, Pedro A. Aguilera, Antonio Salmerón

Noname manuscript No.
(will be inserted by the editor)

Continuous Bayesian networks for probabilistic environmental risk mapping

A.D. Maldonado · P.A. Aguilera · A. Salmerón

Received: date / Accepted: date

Abstract Bayesian networks (BNs) are being increasingly applied to environmental research. Nonetheless, most of the literature related to environmental sciences use discrete or discretized data, which entails a loss of information. We propose a novel methodology based on continuous BNs to predict the probability that surface waters do not meet the standards, in relation to nitrate concentration, established by the European Water Framework Directive. In order to achieve our purpose, a Tree Augmented Naive Bayes (TAN), was developed and applied to estimate and map the risk of failing to meet the European standards established. The TAN models were tested by means of the k-fold cross validation method. The results revealed that the TAN model performed proper risk maps and suggested that poor water quality is highly probable in watersheds dominated by irrigated herbaceous crops. On the contrary, "good surface water status" is more likely to occur in areas where forest is notably present.

Keywords Risk mapping · Regression · Continuous Bayesian networks · Good surface water status

A.D. Maldonado · A. Salmerón
Department of Mathematics, University of Almería, Almería, Spain
E-mail: {amg457,antonio.salmeron}@ual.es

P.A. Aguilera
Informatics and Environment Laboratory, Department of Biology and Geology, University of Almería, Almería, Spain
E-mail: aguilera@ual.es

1 Introduction

It is widely recognized that agricultural practices can put surface waters at risk of pollution, mainly by increasing the concentration of nitrogenous compounds resulting from the consumption of fertilizers (Tilman et al 2001; Foley et al 2005; Moreno et al 2006; Scalon et al 2007; Lee et al 2009). Risk assessment methods can quantify resultant risks (surface water pollution), which can be mapped in order to characterize a given area in terms of risk levels. According to Lahr and Kooistra (2010) there is a enormous diversity of risk maps, ranging from contamination maps to others showing the outcome of complex predictive models. An interesting approach to the concept of risk is given by the probability of exceeding a threshold concentration value of a pollutant (Passarella et al 2002). This probabilistic approach allows to estimate risk straightly from probabilistic tools, whose results can be plotted in order to obtain risk maps.

Bayesian networks (BNs) are probabilistic tools that belong to the so-called *probabilistic graphical models*, which use directed acyclic graphs to represent the joint probability distribution over a set of variables, with the aim of resolving complex problems (Larrañaga and Moral 2011), including characterization, inference, classification and regression. Characterization refers to the analysis of the dependence or independence relationships between variables in the model by means of presence or absence of links in the graph. Inference refers to using a BN to either predict the response or determine the cause of any variable given new values (evidence) of one or more variables in the model. Classification and regression are specific inference problems, in which one of the variables takes the role of interest. The aim of a classification problem is to predict the value of a discrete variable of interest - called the class variable - given the values of other discrete or continuous variables - known as feature variables. The aim of a regression model, on the other hand, is to predict the value of a continuous response variable, given some values of the explanatory continuous or discrete variables.

The review by Aguilera et al (2011) showed that BNs have been applied within the scope of Environmental Sciences, mainly to solve inference issues (Ames et al 2005; Fienen et al 2013; Quinn et al 2013; Dyer et al 2014; Shenton et al 2014), but also for classification problems (Palmsten et al 2013), while papers aimed at resolving regression tasks were uncommon (Borsuk et al 2004). Likewise, while there has been widespread use of discrete and discretized data (Wang et al 2009; Chan et al 2010), few investigations have used continuous data (Pérez-Miñana et al 2012). The infrequent use of

continuous data may be due to the fact that, although most available environmental data are either continuous or hybrid (both discrete and continuous), BNs were originally designed to deal with discrete data, therefore environmental variables have usually been discretized, which implies a loss of information (Uusitalo 2007). Recently, a number of alternative solutions to this problem of information loss have been proposed, including *Mixtures of Polynomials* (MoP) (Shenoy and West 2011), *Mixtures of Truncated Basis Functions* (MoTBFs) (Langseth et al 2012) and *Mixture of Truncated Exponentials* (MTE) (Moral et al 2001). These solutions are able to handle discrete and continuous variables simultaneously, without imposing restrictions on the structure of the network.

In the case of classification and regression problems it is possible to use certain fixed structure models such as naive Bayes (NB) or tree augmented naive Bayes (TAN). These fixed structure models emphasize the importance of one of the variables in the model, with the remaining variables being conditionally dependent on the variable of interest. The NB model is the simplest BN, which assumes the feature/explanatory variables to be conditionally independent of each other given the class/response variable (Friedman et al 1997). The NB model has been used both in classification (Bressan et al 2009; Markus et al 2010; Aguilera et al 2010; Fytilis and Rizzo 2013; Aguilera et al 2013) and regression tasks (Ropero et al 2014). However, feature/explanatory variables are highly correlated at times and the accuracy of the prediction would be improved if dependence relationships between variables could be taken into account. The TAN model allows links between feature/explanatory variables, which results in an increase in complexity. However, a TAN can provide greater accuracy than a NB model (Friedman et al 1997). In spite of its potential, the literature indicates that a TAN model has been applied just once to solve a classification problem (Aguilera et al 2010) in the Environmental Sciences area.

Applications of BNs to risk assessment and risk mapping are scarcely found in the Environmental Sciences area and we are not aware that continuous data were used. Pollino et al (2007) propose a methodology for BN parameterization by means of combining expert knowledge and data and, afterwards, the model is utilized to assess the risk of a native fish in Victoria, Australia. Dlamini (2011) uses BNs to estimate and map fire risk in Swaziland. Aalders et al (2011) take the assessment and mapping of vulnerability and risk of peat deposits in Scotland to erosion as an example for applying BNs. Troldborg et al (2013) apply BNs to estimate and map the vulnerability and risk of soil compaction across Scotland. However, BNs have not so far been applied to estimate and map pollution risk in surface waters.

The objective of this paper is to develop a methodology based on continuous Bayesian networks - more precisely, on a TAN regression model - in order to predict and map the probability of exceeding a threshold value of nitrate concentration in surface waters, taking several land use and environmental variables into account. The threshold value is determined by the EU Water Framework Directive (WFD, 2000/60/EC), which is aimed at meeting "good surface water status" for each river basin in Europe by 2015. In the case of nitrate concentration, the threshold value is set in 25 mg/L for Spain.

2 Methodology

2.1 Study area

The study area occupies roughly 60 000 km^2 of Andalusia, a region of southern Spain (Fig. 1). The study area borders the Sierra Morena mountain range on the north, the Penibaetic System (which is the southernmost mountain range in the Baetic System) on the south-southeast-east, the Guadalquivir Marshes on the southwest and Portugal on the west.

As far as elevation is concerned, the study area ranges from 0 to 3460 meters above sea level, with the highest regions corresponding to the Baetic Systems and the Sierra Morena mountain range, and the lowest corresponding to the Baetic Depression, a vast plain situated between the aforementioned mountain ranges and created by the Guadalquivir river. This orographical diversity enables a wide variation in volume of rainfall, with the rainiest area being located in western foothills of the Subbaetic System (exceeding 2000mm per year), whereas the driest regions lie on southeastern inner depressions of the Baetic Systems (being less than 500 mm per year).

Regarding soil, the study area is widely varied. The Sierra Morena mountain range predominantly presents siliceous lithology, developing acidic and shallow soils with low permeability. In contrast, the Baetic Depression is composed of a variety of detrital and calcareous materials, leading to deep and fertile soils, with the most permeable ones laying along the banks of the Guadalquivir river. The Baetic Systems present a rich soil diversity, dominating calcareous materials (mainly limestone and dolomite) that contribute to alkaline soils.

Concerning land use, (Fig. 2, Appendix A) almost half of the study area (47.52%) corresponds to forest cover, which is mainly located along the Sierra Morena mountain range and the Prebaetic System. Agricultural

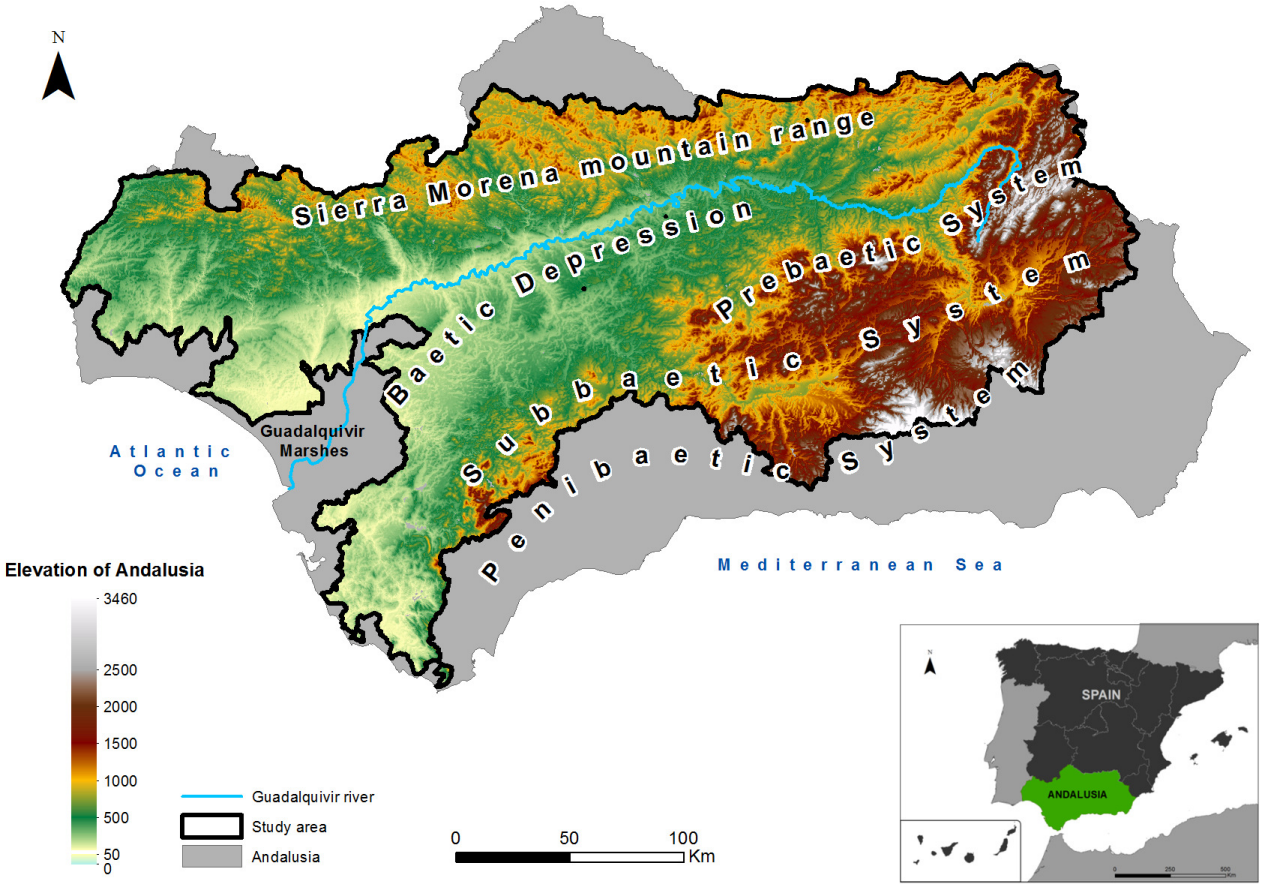

Fig. 1 Study area, which occupies most part of Andalusia (37 °23'00"N 5 °59'00"W)

land uses represent an important part of the study area (47.84%), with olive groves and rainfed herbaceous crops being the most copious ones and being mainly located in the Baetic Depression. Irrigated herbaceous crops are predominantly present on the banks of the Guadalquivir river.

2.2 Data pre-processing

Data from different thematic maps, such as nitrate concentration in surface waters, land use and land cover, temperature, precipitation, potential evapotranspiration and permeability, were obtained from the Andalusian Environmental Information Network [1] and incorporated into a geographic information system - the so-called ArcGis (ESRI®ArcMap^{TM}10.0). The coordinate system for all these datasets is based on the European Terrestrial Reference System 1989 (ETRS89).

First of all, sampling points containing nitrate concentration (N) values were selected and, where possible, 4 measurements taken at different times (1 per season) were collected. N ranged from < 0.1 mg/L to 124.63 mg/L, with a mean of 13.88 mg/L and a standard deviation of 18.3 mg/L. In the analyses carried out in this paper, we used the measurements taken at the sampling points, rather than averaged values. Hence, 971 observations included in 312 sampling points were obtained.

To determine the contributing area of each sampling point, ArcGis *Hydrology* toolset was used (Zhang et al 2012), obtaining as a result a layer composed of 312 watersheds. Watersheds were utilized to delimit the value of the remaining variables. The percentage of each land-use within each watershed was calculated by dividing the area occupied by the land-uses by the watershed area (A). Soil permeability data (K) were incorporated for each given watershed in terms of $1 = low$, $2 = medium$ and $3 = high$.

Moreover, both the average annual potential evapotranspiration (PET) and the average daily precipitation within each watershed were calculated. The daily precipitation data were used to calculate the volume of rainfall in a week (Rain vol.), the number of days

[1] http://www.juntadeandalucia.es/medioambiente/site/rediam

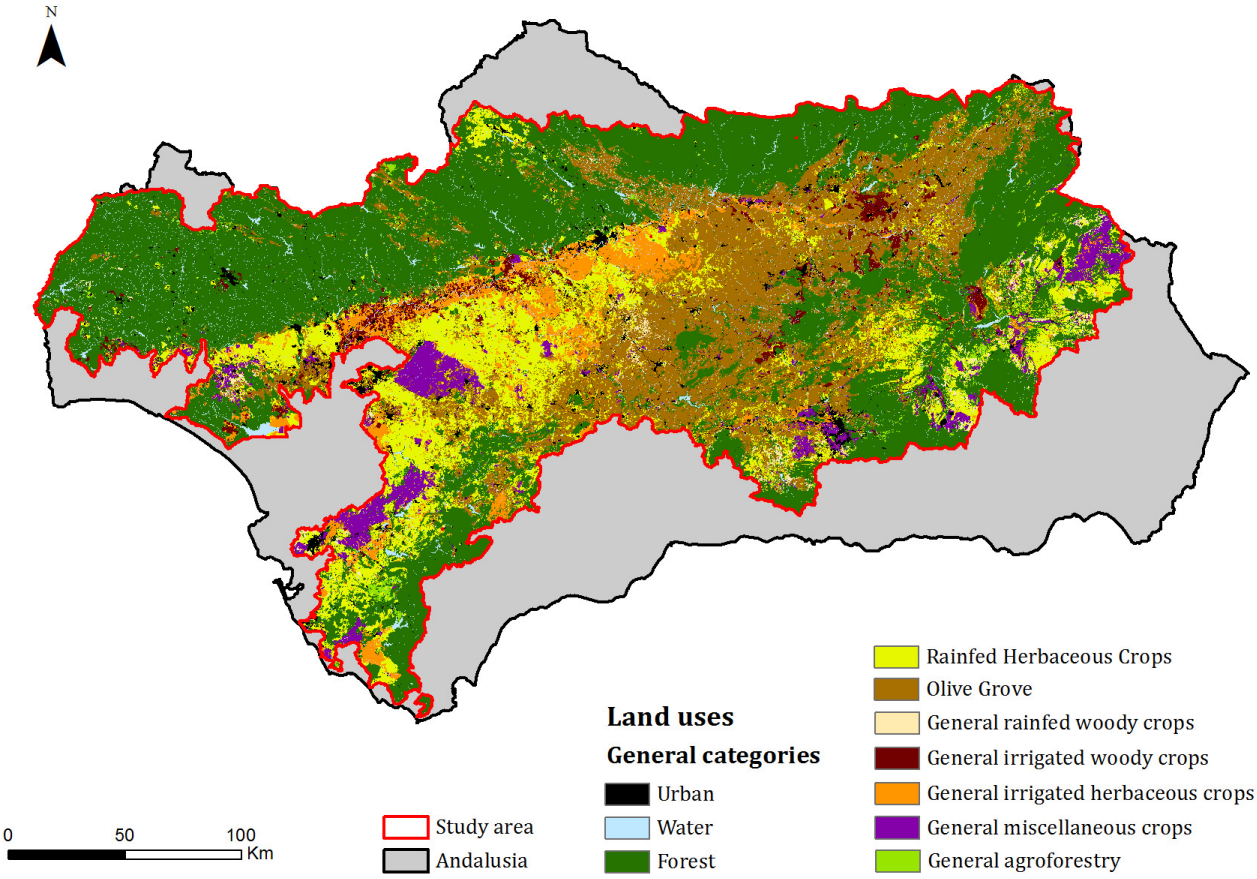

Fig. 2 Land uses selected. For graphical reasons, general categories (Appendix A) are displayed

since the last rainfall event (Last event) and the number of rainy days during a week (Rainy week) for each watershed.

Apart from N, sampling points presented temperature (T) records taken *in situ*. Elevation data (Z) were added to each sampling point from the Andalusian Digital Terrain Model[2] (DTM). Besides, Season (S) variable was selected and obtained from the sampling dates.

Finally, these data yielded a matrix of 971 observations and 44 variables (Table 1), where each observation represents a watershed, at different seasons, characterized by 44 features (34 land uses and another 10 variables: N; A; K; PET; Rain vol.; Last event; Rainy week; T; Z and S).

Once the data matrix was built, data were rescaled in order to make their values range between 0 and 1 by using the transformation

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

[2]The DTM, with grid width 200 m, was provided by the Spanish National Geographic Institute (http://www.ign.es/ign/layoutIn/modeloDigitalTerreno.do)

2.3 Probabilistic Graphical Models. Bayesian Networks

A Bayesian network (BN) is a statistical multivariate model for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, which is defined in terms of two components:

1. Qualitative component, comprising a directed acyclic graph of random variables (vertices, $\mathbf{X}$), with the links between them representing the relationships (of dependence or independence) between variables in the model.
2. Quantitative component, where a set of conditional probability functions represents the strength of the relationships between the variables. The probability distribution of each variable, given its parents, is defined by

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | pa(x_i)) \qquad \forall x_1, \ldots, x_n \in \Omega_{x_1,\ldots,x_n} \tag{2}$$

Table 1 Summary of variables

Variable kind	Description
N (mg/L)[a]	Nitrate concentration. Data collected from sampling stations belonging to the Physico-Chemical Quality Network of the EU WFD in several rivers in Andalusia.
A (km^2)	Watershed area. Upslope area that drains water to each sampling point. These watersheds were delineated by using ArcGis *Hydrology* toolset.
Land uses (%) [b]	Percentage of occupation of each selected land-use within each watershed (31 out of 34 are agricultural land uses. Fig. 2). Note that minority land uses can be handled as *General* categories (Appendix A) to facilitate the analysis of the results and discussion.
K[c]	Permeability, encoded as 1 = *low*, 2 = *medium* and 3 = *high* by the Andalusian Environmental Information Network. The average of soil permeability was calculated for each watershed.
PET (mm)[d]	Average of potential evapotranspiration calculated for each watershed.
Rain vol. (mm)[e]	Average of total rainfall in a week calculated for each watershed.
Last event (days)[e]	Number of days since last rainfall event, calculated for each watershed on average.
Rainy week (days)[e]	Number of days having rainfall during a week, calculated for each watershed on average.
T (°C)[a]	Temperature taken *in situ* at the sampling stations.
Z (m a.s.l.)[f]	Elevation of each sampling point.
S	Season in which each sample was collected.

[a] Variable obtained from the Andalusian Dataset of Surface Waters
[b] Variables obtained from the Andalusian Land Use and Land Cover Map (1:25,000)
[c] Variable obtained from the Andalusian Dataset of Groundwaters
[d] Variable obtained from the Annual PET Dataset of Andalusia (500 m spatial resolution)
[e] Variables calculated from the Daily Precipitation Dataset of Andalusia (500 m spatial resolution)
[f] Variable obtained from the Andalusian DTM (200 m spatial resolution)

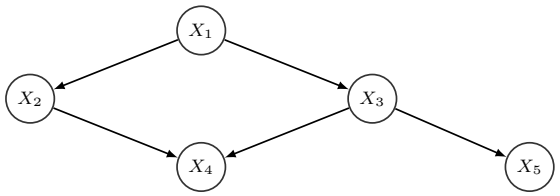

Fig. 3 An example of a Bayesian network

where Ω_{x_i} represents the set of all possible values of variable x_i and $pa(x_i)$ denotes an instantiation of the parents of X_i. Fig. 3 shows an example of a Bayesian network. It represents a joint distribution for variables $X_1, \ldots, X_5$ factorized as

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_3)p(x_4|x_2, x_3). \quad (3)$$

Note that the joint distribution is specified in terms of smaller distributions involving fewer variables. In this way, the network facilitates the specification of complex distributions as it is done in a structured way. In this work, we focus on modeling a particular variable, namely the nitrate concentration (N). With this aim, we will describe the use of BNs to solve regression problems in the next subsection.

2.3.1 Bayesian Networks for Regression: NB and TAN

A Bayesian network can be used as a regression model. Assume that Y is the response variable and $X_1, \ldots, X_n$ are the explanatory variables. Then, in order to predict the value for Y given the observations $x_1, \ldots, x_n$, the conditional density $f(y|x_1, \ldots, x_n)$, is computed to give the numerical prediction for Y, denoted as $\hat{y}$. More specifically, we use the conditional expectation of the response variable (given the observed explanatory variables), which means that the regression model is (Fernández and Salmerón 2008)

$$\hat{y} = g(x_1, \ldots, x_n) = E[Y|x_1, \ldots, x_n] = \int_{\Omega_Y} y f(y|x_1, \ldots, x_n) dy. \quad (4)$$

As $f(y|x_1, \ldots, x_n)$ is proportional to $f(y) \times f(x_1, \ldots, x_n|y)$, solving the regression problem requires the specification of an n dimensional density for $X_1, \ldots, X_n$ given Y. However, using the factorization encoded by the Bayesian network, this problem is simplified depending on the structure of the network. The extreme case is the NB structure, where all the explanatory variables are considered independent given Y (see Fig. 4(a)).

The strong assumption of independence behind NB models is somehow compensated by the reduction on

the number of parameters to be estimated from data, since in this case, it holds that

$$f(y|x_1,\ldots,x_n) \propto f(y)\prod_{i=1}^{n} f(x_i|y) \ , \quad (5)$$

which means that, instead of one n-dimensional conditional density, n one-dimensional conditional densities have to be estimated.

The impact of relaxing the independence assumption has been studied for regression oriented Bayesian networks (Fernández et al 2007), employing the so-called *tree augmented naive Bayes* (TAN) (Friedman et al 1997). In TAN models, more dependencies are allowed, expanding the naive Bayes structure by permitting each feature to have one more parent besides Y (see Fig. 4(b)). In terms of the representation of the distribution over $Y, X_1, \ldots, X_n$, we are interested in modeling problems where discrete and continuous variables coexist. During the last decade, the model based on *mixtures of truncated exponentials (MTEs)* (Moral et al 2001) has perhaps been the most successfully employed in this context.

The MTE model is characterized by a function defined as follows. Let $\mathbf{W} = (W_1, \ldots, W_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be sets of discrete and continuous parts respectively. A *Mixture of Truncated Exponentials* is a function $f()$ defined for each fixed value of the discrete variables as

$$f(z_1,\ldots,z_c) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{\sum_{j=1}^{c} b_i^{(j)} z_j\right\} \quad (6)$$

where a_i, $i = 0,\ldots,m$ and $b_i^{(j)}$, $i = 1,\ldots,m$, $j = 1,\ldots,c$ are real numbers.

An MTE function f is an *MTE density* if it integrates to 1. A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and giving an MTE density of the conditioned variable for each configuration of splits of the other variables. An example of conditional MTE density for a continuous variable Y given a continuous variable X is (Aguilera et al 2013)

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

To estimate the parameters of MTE densities, we followed the approach recently introduced in Langseth et al (2014), which is based on least squares optimization.

One of the advantages of using MTEs in BNs instead of Gaussian densities (Lauritzen 1992) is that the resulting model is more expressive. From the regression point of view, it means that the conditional expectation in Eq. (4) is not necessarily a linear model.

Algorithm 1: Selective MTE-TAN regression model

Input: A database D for variables $X_1,\ldots,X_n,Y$.
Output: Selective TAN regression model for variable Y.

1 **for** $i \leftarrow 1$ to n, compute $\hat{I}(X_i, Y)$.
2 Let $X_{(1)},\ldots,X_{(n)}$ a decreasing order of the independent variables according to $\hat{I}(X_i, Y)$.
3 Divide D into two sets: D_l, for learning the model, and D_t for testing its accuracy.
4 Construct a TAN regression model M with variables Y and $X_{(1)}$ from database D_l.
5 Let $RMSE(M)$ the estimated accuracy of model M using D_t.
6 **for** $i \leftarrow 2$ *to* n **do**
7 Let M_1 be the TAN regression model for the variables in M and $X_{(i)}$.
8 Let $RMSE(M_1)$ be the estimated accuracy of model M_1 using D_t.
9 **if** $(RMSE(M_1) \le RMSE(M))$ $M \leftarrow M_1$.
10 **return** M.

2.3.2 Construction of TAN models using MTEs

The TAN structure (see Fig. 4 (b)) is not unique for a given set of variables. The dependence structure among the explanatory variables is obtained by constructing a maximum spanning tree where the arcs are labelled with the mutual information between the linked variables, conditional on the response variable (Friedman et al 1997; Fernández et al 2007). More precisely, the conditional mutual information between two explanatory variables X_i and X_j given Y is

$$I(X_i, X_j|Y) = \iiint f(x_i, x_j, y) \log \frac{f(x_i, x_j|y)}{f(x_i|y)f(x_j|y)} dx_i dx_j dy \ . \quad (7)$$

The integral above cannot be obtained in closed form for MTE densities, and therefore it has to be approximated. We adopt here the solution proposed in (Fernández et al 2007), consisting of estimating it from

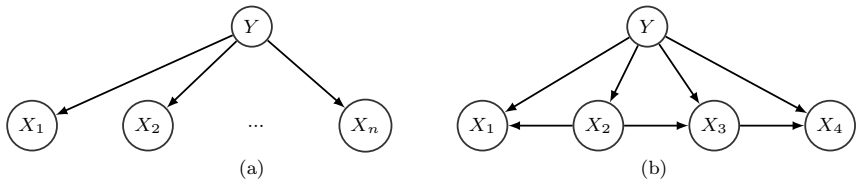

Fig. 4 Structure of a naive Bayes model (a) and a TAN model (b)

a sample of size m, $\{(X_i^{(k)}, X_j^{(k)}, Y^{(k)})\}_{k=1}^m$ drawn from the joint distribution $f(x_i, x_j, y)$, as

$$\hat{I}(X_i, X_j|Y) = \frac{1}{m}\sum_{k=1}^{m}\left(\log f(X_i^{(k)}|X_j^{(k)}, Y^{(k)}) - \log f(X_i^{(k)}|Y^{(k)})\right) . \quad (8)$$

Not necessarily the best strategy for obtaining an accurate regression model conveys the inclusion of all the available explanatory variables in the model. Instead, we decided to use the variable selection procedure described in (Morales et al 2007; Fernández and Salmerón 2008), where the *filter-wrapper* approach described by Ruiz et al (2006) is followed. It consists of first ordering the explanatory variables according to their mutual information with the response variable and then including them one by one according to the initial ranking, whenever the inclusion of a new variable increases the accuracy of the previous model. The accuracy of the model is measured by the *root mean squared error* between the actual values of the response variable and those ones predicted by the model for the records in a test database. If we call $\hat{y}_1, \ldots, \hat{y}_n$ the corresponding estimates provided by the model, the root mean squared error is obtained as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} . \quad (9)$$

The details of the construction of the selective TAN regression model are given in Algorithm 1.

2.3.3 Constructing structured TAN models

When modeling scenarios where the response variable shows a high variability, more accurate models can be obtained if the sample space of the explanatory variables is partitioned and a different model is fit within each split. This is the motivation of successful regression models like the so-called *model trees* (Wang and Witten 1997). Our proposal in this paper is to follow a similar idea, taking as a basis the TAN regression model. More precisely, our goal is to obtain a partition $D_1, \ldots, D_l$ of the sample space of the explanatory variables $X_1, \ldots, X_n$, i.e. $\cup_{i=1}^{l} D_i = \Omega_{x_1,\ldots,x_n}$ and $D_i \cap D_j = \emptyset$, $i \neq j$, and construct a regression model expressed as

$$\hat{y} = m(x_1, \ldots, x_n) = \sum_{i=1}^{l} g_i(x_1, \ldots, x_n) I_{D_i}(x_1, \ldots, x_n), \quad (10)$$

where $I_{D_i}(x_1, \ldots, x_n) = 1$ if $(x_1, \ldots, x_n) \in D_i$ and 0 otherwise, and each g_i, $i = 1, \ldots, l$ is a regression model as in Eq. (4), estimated as a TAN using the data in D_i according to Algorithm 1.

Several methods can be applied to obtain the partition of the sample space. A simple idea is to carry out a hierarchical clustering identifying each individual by the full set of variables, i.e. the response and explanatory variables. A TAN model g_i is then fit within each set D_i in the partition and the *centroid*, denoted as c_i of D_i is attached to g_i. The centroids are computed taking into account only the explanatory variables, as they will be used to decide which TAN to use for prediction purposes, when of course the value of the response variable is not known. The details on constructing the structured TAN regression model are given in Algorithm 2.

When the learnt model is to be used for predicting a value $\hat{y}$ for a given configuration of the explanatory variables, $(x_1, \ldots, x_n)$, the first step is to decide the set in the partition where the configuration will be allocated. It is achieved by computing the distance from $(x_1, \ldots, x_n)$ to each one of the centroids attached to the models g_i. This corresponds to assigning value 1 to the indicator functions I_{D_i} in Eq. (10).

Note that, unlike classic ensemble techniques, the prediction is made from a single model instead of from a weighted average of all of them. The underlying idea is inspired by random-effects models. However, the estimation procedure is completely different, as the TAN model is aimed at estimating the parameters of the conditional distributions.

Algorithm 2: Structured TAN regression model

Input: A database D for variables $X_1, \ldots, X_n, Y$.
Output: A structured TAN regression model for variable Y.

1 Obtain a partition $D_1, \ldots, D_l$ of D using a hierarchical clustering.
2 Let $c_1, \ldots, c_l$ be the centroids of $D_1, \ldots, D_l$ computed using only variables $X_1, \ldots, X_n$.
3 **for** $i \leftarrow 1$ *to* l **do**
4 Obtain a TAN regression model g_i for Y using the data in D_i.
5 Attach centroid c_i to model g_i.
6 Let $I_{D_i}(x_1, \ldots, x_n)$ be a function returning value 1 if c_i is closer to $(x_1, \ldots, x_n)$ than any other centroid c_j, and 0 otherwise.
7 **return** $m(x_1, \ldots, x_n) = \sum_{i=1}^{l} g_i(x_1, \ldots, x_n) I_{D_i}(x_1, \ldots, x_n)$.

2.4 Validation of the Model

A k-fold cross validation (Stone 1974) was carried out in order to test the structured TAN regression model. This technique randomly splits the dataset into k subsets and the method is repeated k times. In each step, one subset is used to test the model built from the remaining k-1 subsets (training subset). Then, the RMSE (Eq. 9) is computed in each step. Finally, the mean of the RMSE is computed to measure the accuracy of the model. In this paper, a k-value of 5 was applied.

2.5 Nitrate risk mapping

A remarkable advantage of regression models based on Bayesian networks is their versatility, in the sense that they not only provide a numerical prediction for a configuration of the explanatory variables, but they also give a full specification of the posterior distribution of the response variable, which is the distribution used to compute the conditional expectation in Eq. (4).

In a scenario where the target variable (*Nitrate concentration* in this case) may take different values for different measurements in the same location, one may be interested in determining the risk that the target variable surpasses a given threshold, rather than giving a fixed prediction.

With regard to Nitrate, water is considered to fail to meet the WFD "good surface water status" if the concentration goes above 25 mg/L. The TAN regression models described above can be used to plot *risk maps*, where the term risk refers to the probability of surpassing the aforementioned threshold, in which case the water body is considered to be below "good status". Hence, for a given point in a map for which the explanatory variables of the TAN regression model are observed to be equal to $(x_1, \ldots, x_n)$, we define the *risk of failing to meet the "good surface water status" due to Nitrate* as

$$R_N(x_1, \ldots, x_n) = P(Y > 25 | x_1, \ldots, x_n) = \int_{25}^{\infty} f(y|x_1, \ldots, x_n) dy, \quad (11)$$

where Y is the Nitrate concentration and $f(y|x_1, \ldots, x_n)$ is the posterior distribution of Y given $(x_1, \ldots, x_n)$, obtained from the TAN regression model described in section 2.3.3.

The computation of the posterior distribution can be carried out over the TAN model by means of the *probability propagation* process, which consists of computing the posterior distribution of some variables in a Bayesian network given that some other variables have been observed. We have implemented the computation of the risk function in Eq. (11) in the Elvira software (Elvira-Consortium 2002), using the variable elimination algorithm (Zhang and Poole 1996) for probability propagation.

Note that function R_N takes values on $[0, 1]$. A risk map can be easily conformed by taking points inside it and evaluating each one of them according to the risk function. Each point can then be colored with an intensity varying from dark green for low risk values (i.e. close to 0) to dark red for high risk values (close to 1). For graphical reasons, watersheds contributing to each sampling point are colored on the risk map instead. A simplified scheme of the methodology followed to obtain the risk map is shown in Fig. 5.

3 Results

3.1 Regression models

The structures of the components of the TAN regression model are shown in Figs. 6, 7 and 8, where each variable is influenced by the response variable (*Nitrate concentration*) and by another explanatory variable directly linked with it in the network. Each component of the TAN model selects the variables that best explain the nitrate concentration in the given data set. It is important to note that the number of variables selected by each component progressively increases from TAN 1 (8) to TAN 2 (11) and TAN 3 (15). The weighted average RMSE of the 3 component of the structured TAN model, calculated as described in section 2.4, is 5.19.

TAN 1 selected 7 land use variables and 1 environmental variable (T). TAN 2 selected 9 land use variables and 2 environmental variables (PET and Z). TAN

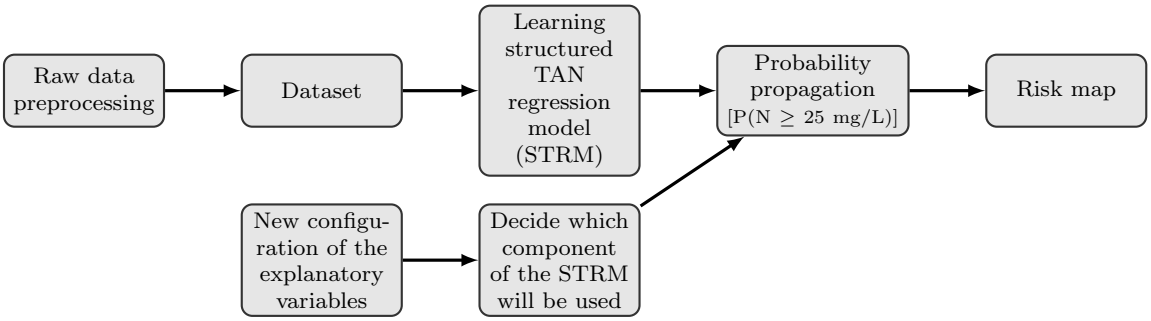

Fig. 5 Methodology followed to obtain the risk map

3 selected 11 land use variables and 4 environmental variables (A, K, S and T). Counting them all, 27 out of the 44 variables were chosen. With reference to the 17 unchosen remaining variables, it is important to note that none of the variables related to precipitation (*Rain vol.*, *Last event*, *Rainy week*) explain *Nitrate concentration*, for these datasets.

Forest was the only land use variable chosen by the 3 components of the TAN model. Furthermore, TAN 1 and TAN 2 coincided on 4 other variables: *Olive grove*, *Vineyard*, *Rainfed olive grove and vineyard crops* and *Herbaceous and woody crops and natural vegetation*. Meanwhile, TAN 1 and TAN 3 coincided only on *Temperature* and TAN 2 and TAN 3 just coincided on *partly Irrigated herbaceous crops*.

Regarding *General* categories (Appendix A), variables included in the Irrigated herbaceous crop, Miscellaneous and Agroforestry general categories were chosen by the 3 models. On the other hand, variables included in the General Rainfed woody crop category were chosen only by TAN 1 and 2 while only TAN 3 chose variables included in the General Irrigated woody crop category.

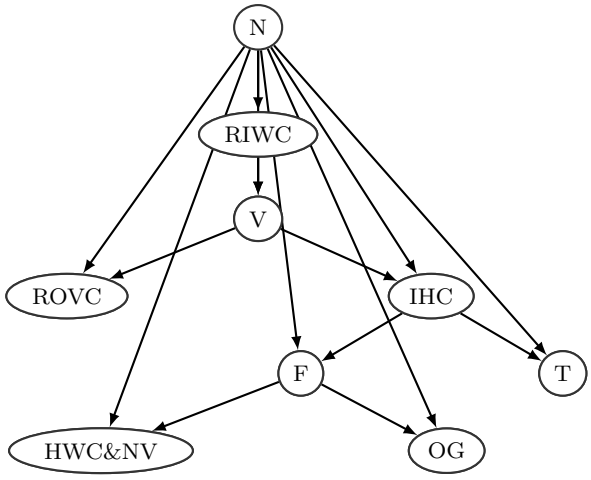

Fig. 6 TAN 1. First component of the structured TAN model used for constructing the risk map. N: Nitrate; RIWC: Rainfed and Irrigated Woody Crops; V: Vineyard; ROVC: Rainfed Olive grove and Vineyard Crops; IHC: Irrigated Herbaceous Crops; F: Forest; T: Temperature; HWC&NV: Herbaceous and Woody Crops and Natural Vegetation; OG: Olive grove

3.2 Analysis of risk map obtained through TAN regression models

The risk map (Fig. 9) was obtained from the the structured TAN regression model, as explained in section 2.5. The map shows the estimated probability that nitrate concentration will exceed 25 mg/L in surface waters, i.e., the risk of failing to meet the WFD. Table 2 shows the average value of the variables within each risk level class[3](*precipitation* variables are included in the table, even though none of the components of the regression model selected them).

Very low risk probabilities mainly correspond to watersheds located in the Sierra Morena mountain range and eastern Baetic Systems. These watersheds are mainly occupied by forest (77.62% on average), with olive groves and rainfed herbaceous crops covering a low percentage (7.27% and 5.60% respectively) and the remaining land uses being scarcely represented. Moreover, these watersheds are located at the highest elevations (351.40 m a.s.l., on average) and present the lowest both watershed area (573.15 km^2) and soil permeability value (1.54), indicating a silty clay soil texture.

Low risk probabilities mainly correspond to watersheds bordering the Sierra Morena mountain range on the south. On average, Forest does not dominate (34.94%), Olive grove increases dramatically (up to 30.31%) and both Rainfed herbaceous crops (16.52%) and General Irrigated herbaceous crops (3.10%) become more patent. The area of these watersheds is the largest (8340.19 km^2, on average).

[3]Risk level classes: *very low*=[0-0.1], *low*=(0.1-0.3], *moderate*=(0.3-0.5], *high*=(0.5-0.8] and *very high*=(0.8-1] (Dlamini 2011)

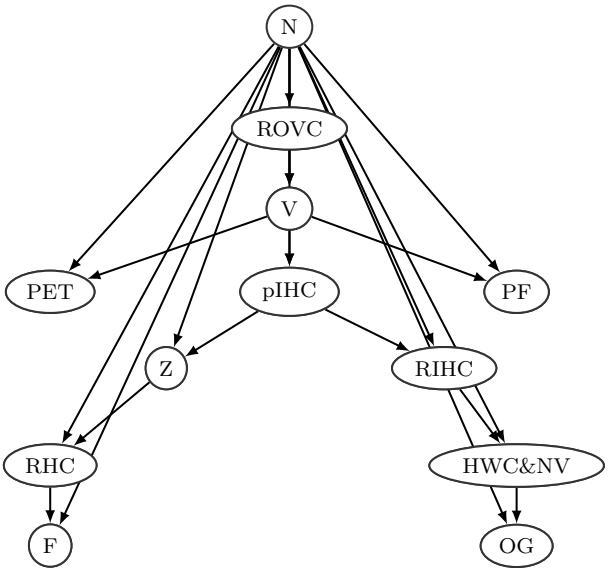

Fig. 7 TAN 2. Second component of the structured TAN model used for constructing the risk map. N: Nitrate; ROVC: Rainfed Olive grove and Vineyard Crops; V: Vineyard; PET: Potential Evapotranspiration; PF: Paddy Field; pIHC: Partly Irrigated Herbaceous Crops; Z: Elevation; RIHC: Rainfed and Irrigated Herbaceous Crops; RHC: Rainfed Herbaceous Crops; HWC&NV: Herbaceous and Woody Crops and Natural Vegetation; F: Forest; OG: Olive grove

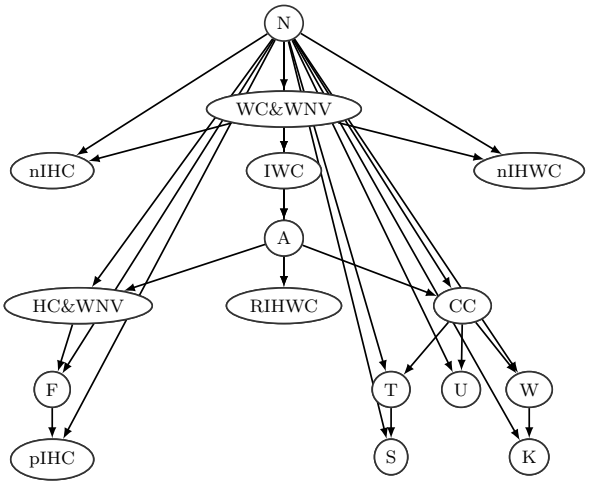

Fig. 8 TAN 3. Third component of the structured TAN model used for constructing the risk map. N: Nitrate; WC&WNV: Woody Crops and Woody Natural Vegetation; nIHC: Non-Irrigated Herbaceous Crops; nIHWC: Non-Irrigated Herbaceous and Woody Crops; IWC: Irrigated Woody Crops; A: watershed Area; HC&WNV: Herbaceous Crops an Woody Natural Vegetation; RIHWC: Rainfed and Irrigated Herbaceous and Woody Crops; CC: Citrus Crops; F: Forest; T: Temperature; U: Urban; W: Water; pIHC: Partly Irrigated Herbaceous Crops; S: Season; K: Permeability

Moderate risk probabilities mainly correspond to watersheds that border the Subbaetic System on the north. There is a mixture of different crops, with dominance of Forest (36.51%), Olive grove (23.64%) and Rainfed herbaceous crops (18.40%). On average, General Miscellaneous crops increase up to 9.49% and General Irrigated herbaceous crops represent 3.83%.

High risk probabilities mainly correspond to watersheds located in the Baetic Depression. In these watersheds, the most important change in terms of land use is the increase of General Irrigated herbaceous crops, occupying up to 8.00%, on average.

Very high risk probabilities mainly correspond to watersheds located in the lowest areas (145.04 m a.s.l., on average) of the Baetic Depression. On average, these watersheds present the lowest percentage of Forest (11.33%) and the highest of Rainfed herbaceous crops (35.75%), General Irrigated herbaceous crops (14.71%) and General Irrigated woody crops (5.55%). In addition, these watersheds show the highest soil permeability value (2.3), indicating a sandy soil texture.

4 Discussion

4.1 TAN as a new methodology for environmental risk mapping

Many methodologies used to assess risk in surface waters are commonly based on indexes (Gillentine 2000; Eimers et al 2000; Giupponi et al 1999; Verro et al 2002, 2009) and, even though there is no agreement on the definition, risk is usually understood as the result of combining vulnerability with hazard (Huang 2009; Zou et al 2013). However, index methodologies based on expert opinion may yield completely opposed results (Diamantino et al 2005) since the parameters and weights introduced in the model can be different from one methodology to another (Payraudeau and van der Werf 2005). Probabilistic methodologies lead to a number of advantages over the traditional ones, such as the avoidance of estimating both components (vulnerability and hazard) of risk (Barca and Passarella 2008). This advantage prevents undesirable errors, such as misestimating one of those components (Moratalla et al 2011).

In this regard, we found that TAN models are useful for assessing risk in surface waters, especially when dealing with dynamical compounds such as nitrate. In general, BNs provide several advantages over traditional methods (Aguilera et al 2011). Instead of just providing a numerical prediction, the TAN is able to compute the probability of a particular hypothesis (for instance,

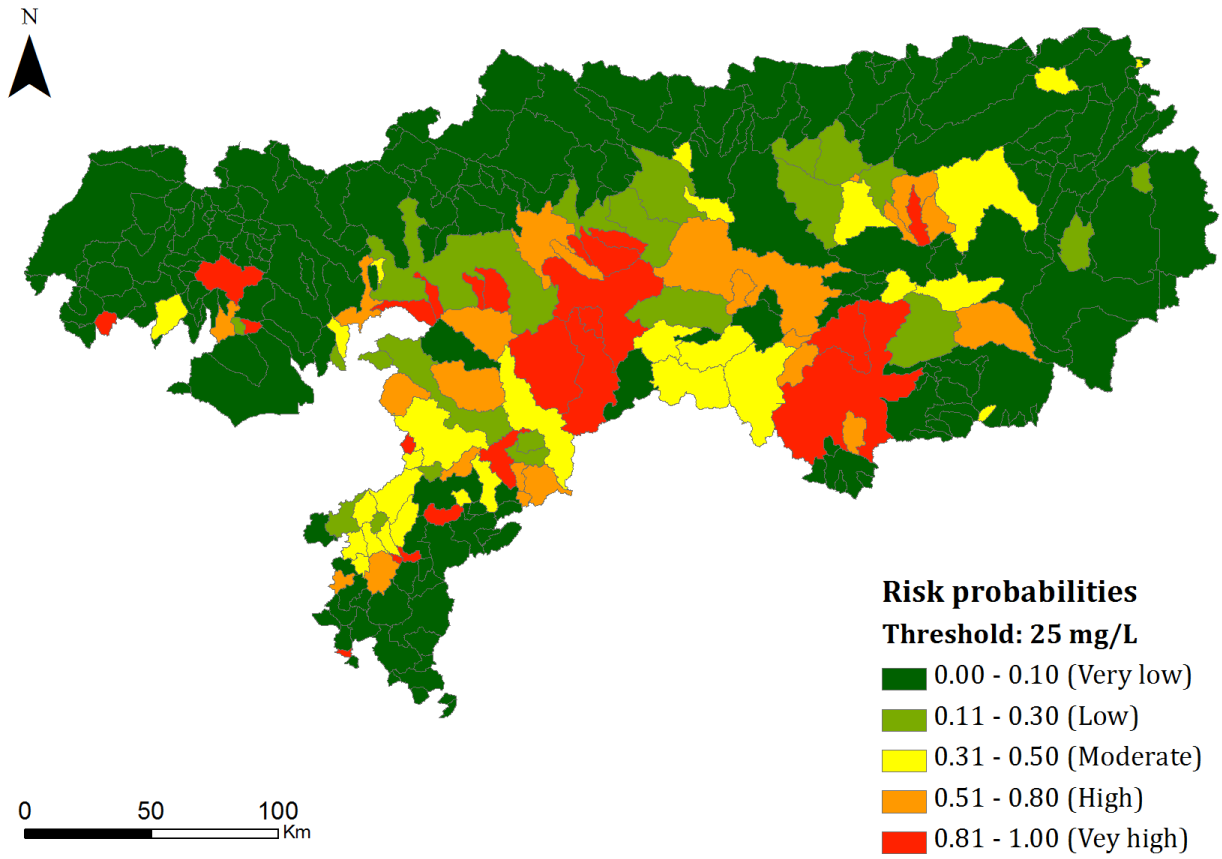

Fig. 9 Risk map depicting the probability of failing to meet the "good surface water status"

the probability that nitrate concentration will exceed 25 mg/L), which is useful for elaborating risk maps. Moreover, we may also be interested in predicting the posterior density of the response variable (*Nitrate concentration*) given some evidence of the explanatory variables (land uses, etc.), which is also possible since the variables are expressed in terms of their density functions. That means that the TAN provides us not only with a numeric prediction of the response, but also with a full specification of its posterior distribution. Knowing the distribution can be useful in practice for making inferences about the response variable as, for instance, computing confidence intervals or testing hypotheses.

On the other hand, BNs are capable of dealing with continuous data, which allows each variable to be expressed in terms of its density function. However, none of the found literature that performs BNs to estimate risk uses continuous data (Liao et al 2010; Aalders et al 2011; Dlamini 2011; Liang et al 2012; Rennie et al 2007; Troldborg et al 2013). On the contrary, discretizing continuous variables is a standard operating procedure, even though it entails a loss of information (Aguilera et al 2011).

In view of the qualitative component of the TAN model, it is worth noting that as heterogeneity of land use increases, the complexity of the structure augments by adding explanatory variables. The selection as well as the dependence relationships existing between explanatory variables allows the model to perform better.

4.2 Assessment of the fulfillment of the Water Framework Directive

The results highlight that areas dominated by forest display a *very low* risk of exceeding the threshold values established by the European legislation. According to Borin et al (2005) forested lands can act as buffer strips, providing an efficient way to decrease nutrient concentration from runoff. Moreover, these areas are located at the highest elevations, which means that water flow speeds downhill, reducing its residence time. Also, steepness, inaccessibility and and poor quality soils discourage the introduction of agricultural prac-

Table 2 Average value of the variables within each risk level class

		Risk level				
Variable		**Very low**	**Low**	**Moderate**	**High**	**Vey high**
Land uses (%)	Urban	1.43	4.89	2.82	1.72	2.94
	Water	2.15	1.66	2.08	1.68	1.71
	Forest	77.62	34.94	36.51	34.88	11.33
	Rainfed herbaceous crops	5.60	16.52	18.40	20.01	35.75
	Olive Grove	7.27	30.31	23.64	25.02	23.87
	Rainfed woody crops*	0.64	1.86	0.93	1.21	0.52
	Irrigated woody crops*	1.04	1.62	1.21	1.89	5.55
	Irrigated herbaceous crops*	0.97	3.10	3.83	8.00	14.71
	Miscellaneous crops*	1.79	3.87	9.49	4.29	3.07
	Agroforestry*	1.48	1.25	1.09	1.29	0.55
A (km^2)		573.15	8340.19	1624.84	4585.79	939.19
K		1.54	2.02	1.87	2.12	2.29
PET (mm)		63.48	66.70	66.03	60.83	66.08
Rain vol. (mm)		12.17	12.18	8.10	9.15	7.26
Last event (days)		3.28	2.52	2.42	2.33	2.81
Rainy week (days)		3.82	4.78	4.35	4.29	3.55
T (°C)		16.44	16.62	16.83	16.55	17.45
Z (m a.s.l.)		351.40	258.00	247.07	268.98	145.04

* General categories. See Appendix A.

tices, which benefits water quality in terms of nutrient concentration.

In contrast, watersheds located at the lowest elevations in the Baetic depression favors the existence of high yield crops because of the flat plains and the soil fertility. These watersheds are occupied by important percentages of irrigated (herbaceous and woody) crops, which show a *very high* probability of exceeding the "good quality" threshold. The results are consistent with previous research which agrees on the large negative impact of irrigated crops on hydrological cycles because of their high water requirement, fertilizer use (Scalon et al 2007) and irrigation return flow (Causapé et al 2006). Besides, the surface covered by rainfed herbaceous crops is considerably large. Rainfed herbaceous crops are an important contributor to water pollution, although to a lesser extent than irrigated crops, since the absence of irrigation reduces the potential negative impact on surface waters (Sun et al 2013). Regarding olive groves, which also represent a relevant percentage, they may affect water quality since they are treated with excess nitrogen fertilizer (Fernández-Escobar et al 2009). Moreover, forest covers a small percentage and, as a result, cannot act as buffer strips. Furthermore, these watersheds are located at the lowest elevations; hence, nutrients drain by runoff and accumulate at the lowest outlets.

The remaining watersheds, whose risk level is halfway between both previous situations (*very low* and *very high* risk levels), show a mixture of land uses, where at least 3 of them maintain a land-use balance in terms of extension. However, it is observable that watersheds classified as at *high* risk differentiate from the others mainly because of the the increase of irrigated herbaceous crops, which (as mentioned above) are a major source of nutrient enrichment to surface waters. On the other hand, watersheds with a *moderate* risk of exceeding the threshold present a notable percentage of *General* miscellaneous crops, which have an important component of irrigated crops (Appendix A). With regard to

low risk watersheds, even though the area occupied by olive grove is large, land-use surface having irrigated crops is low in percentage terms.

The results suggest that land uses, especially irrigated crops, are the most influential variables in nitrate concentration. Nevertheless, some of the environmental variables may play an active role. The average of *Elevation (Z)* decreases dramatically from *very low* to *very high* risk levels and keeps approximately constant in the intermediate (*low*, *moderate* and *high*) risk levels, where land uses determine the response of *Nitrate concentration*. It is widely known that elevation is inversely related to agriculture, due to accessibility and climate. Although some research has established a relationship between nitrate concentration and precipitation (Grimm and Lynch 2005), our selective structured TAN regression model discarded variables related to rainfall, which are not well correlated to *Nitrate concentration*, for our dataset.

5 Conclusions

The TAN model has not previously been applied to draw surface water risk maps. The results given by the structured TAN regression model accord with previous studies, showing that this novel methodology performs appropriate predictions, and also provides advantages over other regression models. In addition, the existence of efficient algorithms make learning, inference and validation processes convenient.

The analysis of the risk map highlights that water bodies running on intensive irrigated herbaceous farmlands are highly probable to exceed the trigger value of 25 mg/L, which indicates that those watersheds fail to meet the "good surface water status" established by the EU Water Framework Directive. Summing up, we have found the TAN model to be an appropriate tool for risk mapping in surface waters. Furthermore, this methodology can be applied to other environmental research areas.

A Land uses

Land use variables incorporated into the data matrix. Percentage of occupation of each land use in the study area is shown in brackets.

1. *Urban* (2.19%) includes urban, industrial and commercial areas, landfills, mining deposits, communication infrastructure, parks, recreational and sport facilities and areas under construction.
2. *Water* (2.45%) comprises surface waters in Andalusia, including rivers, artificial channels, lakes and reservoirs. For the study purposes, only waterbodies corresponding to rivers were taken into account.
3. *Forest* (47.52%) includes forest, shrub and grassland cover.
4. *Rainfed herbaceous crops* (13.89%) consist of non-irrigated herbaceous monocultures, with cereals (wheat, barley, oats) and leguminous crops (peas, chickpeas, beans) being the most copious crops.
5. *Olive grove* (21.08%) is the main crop of inland Andalusia and consists of non-irrigated monocultures, excluding wild olive trees.
6. *Vineyard*[1] (0.28%) consists of non-irrigated monocultures, devoted to grape production.
7. *Rainfed woody crops*[1] (0.67%) comprise woody monocultures under rainfed conditions, such as almond, carob, fig, walnut or chestnut trees, excluding plots dedicated to logging activities.
8. *Rainfed olive grove and vineyard crops*[1] (0.03%) are composed of mixtures of vine and olive trees under dry farming conditions.
9. *Rainfed woody heterogeneous crops*[1] (0.03%) comprise vine and olive tree associations with other rainfed woody crops, where no dominance of any of the crops exists.
10. *Abandoned olive grove*[1] (0.19%) comprises abandoned plots of woody crops, patently dominated by olive grove.
11. *Abandoned woody crops*[1] (0.03%) include abandoned plots of undifferentiated woody crops.
12. *Paddy fields*[2] (0.07%) consist of flooded parcels devoted to rice cultivation.
13. *Greenhouse crops*[2] (0.10%) consist of high yield crops under controlled conditions.
14. *Irrigated herbaceous crops*[2] (1.96%) comprise permanently irrigated intensive herbaceous monocultures, including lettuce, asparagus, carrot, onion and garlic crops.
15. *Partly irrigated herbaceous crops*[2] (2.07%) comprise both irrigated and non-irrigated (but liable to be irrigated) plots where herbaceous crops are grown.
16. *Non-irrigated herbaceous crops*[2] (0.49%) consist of irrigated herbaceous crop areas that were not being watered at the moment of taking the image.
17. *Partly irrigated woody crops*[3] (0.14%) are composed of both irrigated and non-irrigated (but liable to be irrigated) plots where woody crops are grown.
18. *Citrus crops*[3] (0.64%) include orange, lemon, mandarin and grapefruit trees, among other irrigated woody species.
19. *Irrigated olive grove*[3] (1.17%) consists of irrigated olive tree monocultures.
20. *Tropical crops*[3] (0.00003%) include avocado, cherimoya, mango and medlar trees, among other irrigated woody species.
21. *Irrigated woody crops*[3] (0.13%) include other irrigated woody crops not aforementioned.
22. *Irrigated woody heterogeneous crops*[3] (0.1%) consist of undifferentiated woody crop mixtures under irrigated conditions.
23. *Rainfed herbaceous and woody crops*[4] (0.63%) consist of annual herbaceous crops associated with permanent woody crops under dry farming conditions.
24. *Irrigated herbaceous and woody crops*[4] (0.23%) consist of annual herbaceous crops associated with permanent woody crops under irrigated conditions.

[1]Land uses included into the "General rainfed woody crop" category

[2]Land uses included into the "General irrigated herbaceous crop" category

[3]Land uses included into the "General irrigated woody crop" category

[4]Land uses included into the "General miscellaneous crop" category

25. *Partly irrigated herbaceous and woody crops*[4] (0.02%) comprise woody and herbaceous crop mixtures under either dry farming or irrigated conditions.
26. *Non-irrigated herbaceous and woody crops*[4] (0.01%) are composed of woody and herbaceous crop mixtures which are situated on non-irrigated plots at the moment of taking the image.
27. *Rainfed and irrigated herbaceous crops*[4] (2.58%) comprise undifferentiated herbaceous crop mixtures under either rainfed or irrigated conditions.
28. *Rainfed and irrigated herbaceous and woody crops*[4] (0.31%) are composed of undifferentiated woody and herbaceous crop mixtures under either rainfed or irrigated conditions.
29. *Rainfed and irrigated woody crops*[4] (0.01%) comprise undifferentiated woody crop mixtures under either rainfed or irrigated conditions.
30. *Herbaceous crops and grasslands*[5] (0.25%) consists of land mainly occupied by undifferentiated herbaceous crops, with significant areas of grassland.
31. *Herbaceous crops and woody natural vegetation*[5] (0.19%) is composed of land mainly covered by herbaceous crops, with a important portion occupied by woody natural vegetation.
32. *Woody crops and grasslands*[5] (0.02%) consists of land mainly occupied by undifferentiated woody crops, with significant areas of grassland.
33. *Woody crops and woody natural vegetation*[5] (0.43%) is composed of land mainly covered by woody crops, with a important portion occupied by woody natural vegetation.
34. *Herbaceous and woody crops and natural vegetation*[5] (0.19%) includes other undifferentiated crop mixtures associated with natural vegetation, not aforementioned.

Acknowledgements This work has been supported by the Spanish Ministry of Economy and Competitiveness, through project TIN2013-46638-C3-1-P , by Junta de Andalucía through project P11-TIC-7821 and by ERDF-FEDER funds. A.D. Maldonado is being supported by the Spanish Ministry of Education, Culture and Sport through an FPU research grant, FPU2013/00547.

Author conflict of interest declaration

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from amg457@ual.es.

The authors;

A.D. Maldonado P.A. Aguilera A. Salmerón

[5]Land uses included into the "General agroforestry" category

References

Aalders I, Hough RL, Towers W (2011) Risk of erosion in peat soils - an investigation using Bayesian belief networks. Soil Use and Management 27:538–549

Aguilera PA, Fernández A, Reche F, Rumí R (2010) Hybrid Bayesian network classifiers: Application to species distribution models. Environmental Modelling & Software 25(12):1630–1639

Aguilera PA, Fernández A, Fernández R, Rumí R, Salmerón A (2011) Bayesian networks in environmental modelling. Environmental Modelling & Software 26:1376–1388

Aguilera PA, Fernández A, Ropero RF, Molina L (2013) Groundwater quality assessment using data clustering based on hybrid Bayesian networks. Stochastic Environmental Research & Risk Assessment 27(2):435–447

Ames DP, Neilson BT, Stevens DK, Lall U (2005) Using Bayesian networks to model watershed management decisions: an East Canyon Creek case study. Journal of Hydroinformatics 7:267 – 282

Barca E, Passarella G (2008) Spatial evaluation of the risk of groundwater quality degradation. a comparison between disjunctive kriging and geostatistical simulation. Environmental Monitoring Assessment 137:261–273

Borin M, Vianello M, Morari F, Zanin G (2005) Effectiveness of buffer strips in removing pollutants in runoff from a cultivated field in North-East Italy. Agriculture, Ecosystems and Environment 101:101–114

Borsuk ME, Stow CA, Reckhow KH (2004) A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling 173:219–239

Bressan GM, Oliveira VA, Hruschka ER, Nicoletti MC (2009) Using Bayesian networks with rule extraction to infer risk of weed infestation in a corn-crop. Engineering Applications of Artificial Intelligence 22:579–592

Causapé J, Quílquez D, Aragüés R (2006) Irrigation efficiency and quality of irrigation return flows in the Ebro river basin: an overview. Environmental Monitoring Assessment 117:451–461, DOI DOI: 10.1007/s10661-006-0763-8

Chan T, Ross H, Hoverman S, Powell B (2010) Participatory development of a Bayesian network model for catchment-based water resource management. Water Resources Research 46:doi: 10.1029/2009WR008,848

Diamantino C, Henriques MJ, Oliveira MM, Lobo-Ferreira JP (2005) Methodologies for pollution risk assessment of water resources systems. In: The Fourth Inter-Celtil Colloquium on Hydrology and Management of Water Resources

Dlamini WM (2011) Application of Bayesian networks for fire risk mapping using GIS and remote sensing data. GeoJurnal 76:283–296

Dyer F, ElSawah S, Croke B, Griffiths R, Harrison E, Lucena-Moya P, Jakeman A (2014) The effects of climate change on ecologically-relevant flow regime and water quality at-

tributes. Stochastic Environmental Research & Risk Assessment 28:67–82

Eimers L, Weaver JC, Terziotti S, Midgette RW (2000) Methods of rating unsaturated zone and watershed characteristics of public water supplies in North Carolina. Tech. rep., U.S. Geological Survey. Water-Resources Investigations

Elvira-Consortium (2002) Elvira: An environment for probabilistic graphical models. In: Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02), pp 222–230

Fernández A, Salmerón A (2008) Extension of Bayesian network classifiers to regression problems. In: Geffner H, Prada R, Alexandre IM, David N (eds) Advances in Artificial Intelligence - IBERAMIA 2008, Springer, Lecture Notes in Artificial Intelligence, vol 5290, pp 83–92

Fernández A, Morales M, Salmerón A (2007) Tree augmented naïve Bayes for regression using mixtures of truncated exponentials: Applications to higher education management. IDA'07 Lecture Notes in Computer Science 4723:59–69

Fernández-Escobar R, Marin L, Sánchez-Zamora MA, García-Novelo JM, Molina-Soria C, Parra MA (2009) Long-term effects of N fertilization on cropping and growth of olive trees and on N accumulation in soil profile. European Journal of Agronomy 31:223–232, DOI 10.1016/j.eja.2009.08.001

Fienen MN, Masterson JP, Plant NG, Guitierrez BT, Thieler ER (2013) Bridging groundwater models and decision support with a Bayesian network. Water Resources Reseach 49:6459–6473

Foley JA, DeFries R, Asner GP, Barford C, Bonan G, Carpenter SR, Chapin FS, Coe MT, Daily GC, Gibbs HK, Helkowski JH, Holloway T, Howard EA, Kucharik CJ, Monfreda C, Patz JA, Prentice IC, Ramankutty N, Snyder PK (2005) Global Consequences of Land Use. Science 309:570–574, DOI 10.1126/science.1111772

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

Fytilis N, Rizzo DM (2013) Coupling self-organizing maps with a Naïve Bayesian classifier: Stream classification studies using multiple assessment data. Water Resources Reseach 49:7747–7762

Gillentine J (2000) Source Water Assessment and Protection Program. Tech. rep., New Mexico Environmental Department. Drinking Water Bureau, Appendix E - WRASTIC Index: Watershed vulnerability estimation using WRASTIC

Giupponi C, Eiselt E, Ghetti P (1999) A multicriteria approach for mapping risks of agricultural pollution for water resources: The Venice Lagoon watershed case study. Journal of Environmental Management 56:259–269

Grimm JW, Lynch J (2005) Improved daily precipitation nitrate and ammonium concentration models for the Chesapeake Bay Watershed. Environmental Pollution 135:445–455

Huang C (2009) Integration degree of risk in terms of scene and application. Stochastic Environmental Research & Risk Assessment 23:473–484

Lahr J, Kooistra L (2010) Environmental risk mapping of pollutants: State of the art and communication aspects. Journal of the Total Environment 408:3899–3907

Langseth H, Nielsen TD, Rumí R, Salmerón A (2012) Mixtures of Truncated Basis Functions. International Journal of Approximate Reasoning 53(2):212–227

Langseth H, Nielsen T, Pérez-Bernabé I, Salmerón A (2014) Learning mixtures of truncated basis functions from data. International Journal of Approximate Reasoning 55:940–956

Larrañaga P, Moral S (2011) Probabilistic graphical models in artificial intelligence. Applied Soft Computing 11:1511–1528, DOI 10.1016/j.asoc.2008.01.003

Lauritzen SL (1992) Propagation of probabilities, means and variances in mixed graphical association models. Journal of the American Statistical Association 87:1098–1108

Lee SW, Hwang SJ, Lee SB, Hwang HS, Sung HC (2009) Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. Landscape and Urban Planning 92:80–89

Liang Wj, Zhuang Df, Jiang D, Pan Jj, Ren Hy (2012) Assessment of debris flow hazards using a Bayesian Network. Geomorphology 171-172:94–100

Liao Y, Wang J, Guo Y, Zheng X (2010) Risk assessment of human neural tube defects using a Bayesian belief network. Stochastic Environmental Research & Risk Assessment 24:93–100

Markus M, Hejazi MI, Bajcsy P, Giustolisi O, Savic DA (2010) Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. Journal of Hydroinformatics 12.3:251–261, DOI 10.2166/hydro.2010.064

Moral S, Rumí R, Salmerón A (2001) Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In: Benferhat S, Besnard P (eds) Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, Lecture Notes in Artificial Intelligence, vol 2143, pp 156–167

Morales M, Rodríguez C, Salmerón A (2007) Selective naïve Bayes for regression using mixtures of truncated exponentials. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 15:697–716

Moratalla A, Gómez-Alday JJ, Sanz D, no SC, de las Heras J (2011) Evaluation of a GIS-Based integrated vulnerability risk assessment for the Mancha Oriental System (SE Spain). Water Resources Management 25:3677–3697

Moreno JL, Navarro C, las Heras JD (2006) Abiotic ecotypes in south-central Spanish rivers: Reference conditions and pollution. Environmental Pollution 143:388–396, DOI 10.1016/j.envpol.2005.12.012

Palmsten ML, Holland KT, Plant NG (2013) Velocity estimation using a Bayesian network in a critical-habitat reach of the Kootenai River, Idaho. Water Resources Reseach 49:5865–5879

Passarella G, Vurro M, D'Agostino V, Giuliano G, Barcelona M (2002) A probabilistic methodology to assess the risk of groundwater quality degradation. Environmental Monitoring Assessment 79:57–74

Payraudeau S, van der Werf HM (2005) Environmental impact assessment for farming region: a review of methods. Agriculture, Ecosystems and Environment 107:1–19

Pérez-Miñana E, Krause PJ, Thornton J (2012) Bayesian Network fot the management of greenhouse gas emissions in the British agricultural sector. Environmental Modelling & Software 35:132–148

Pollino CA, Woodberry O, Nicholson A, Korb K, Hart BT (2007) Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. Environmental Modelling & Software 22:1140–1152

Quinn JM, Monaghan RM, Bidwell VJ, Harris SR (2013) A Bayesian Belief Network approach to evaluating complex effects of irrigation-driven agricultural intensification scenarios on future aquatic environmental and economic values in a New Zealand catchment. Marine and Freshwater Resarch 64:460–474, DOI 10.1071/MF12141

Rennie SE, Brandt A, Plant N (2007) A probabilistic expert system approach for sea mine burial prediction. IEEE Journal of Oceanic Engineering 32:260–272

Ropero RF, Aguilera PA, Fernández A, Rumí R (2014) Regression using hybrid Bayesian networks: Modelling landscape-socioeconomy relationships. Environmental Modelling & Software 54:127–137

Ruiz R, Riquelme J, Aguilar-Ruiz JS (2006) Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition 39:2383–2392

Scalon BR, Jolly I, Sophocleous M, Zhang L (2007) Global impacts of conversions from natural to agricultural ecosystems on water resources: Quantity versus quality. Water Resources Reseach 43:W03,437

Shenoy PP, West JC (2011) Inference in hybrid Bayesian networks using mixtures of polynomials. International Journal of Approximate Reasoning 52(5):641–657

Shenton W, Hart BT, Chan TU (2014) A Bayesian network approach to support environmental flow restoration decisions in the Yarra River, Australia. Stochastic Environmental Research & Risk Assessment 28:57–65

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B (Methodological) 36 (2):111–147

Sun R, Chen L, Chen W, Ji Y (2013) Effect of land use patterns on total nitrogen concentration in the upstream regions of the Haihe River Basin, China. Environmental Management 51:45–58

Tilman D, Fargione J, Wolf B, D'Antonio C, Dobson A, Howarth R, Schindler D, Schlesinger WH, Simberloff D, Swackhamer D (2001) Forecasting agriculturally driven global environmental change. Science 292:281–284, DOI 10.1126/science.1057544

Troldborg M, Aalders I, Towers W, Hallett PD, McKenzie BM, Bengough AG, Lilly A, Ball BC, Hough RL (2013) Application of Bayesian Belief Networks to quantify and map areas at risk to soil threats: Using soil compaction as an example. Soil & Tillage Research 132:56–68

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

Verro R, Calliera M, Maffioli G, Auteri D, Sala S, Finizio A, Vighi M (2002) GIS-Based system for surface water risk assessment of agricultural chemicals. 1. methodological approach. Environmental Science & Technology 36:1532–1538

Verro R, Finizio A, Otto S, Vighi M (2009) Predicting pesticide environmental risk in intensive agricultural areas I: Screeneing level risk assessment of individual chemicals in surface waters. Environmental Science & Technology 43:522–529

Wang QJ, Robertson DE, Haines CL (2009) A Bayesian network approach to knowledge integration and representation of farm irrigation : 1. Model development. Water Resources Reseach 45:doi:10.1029/2006WR005,419

Wang Y, Witten IH (1997) Induction of model trees for predicting continuous cases. In: Proceedings of the Poster Papers of the European Conference on Machine Learning, pp 128–137

Zhang NL, Poole D (1996) Exploiting causal independence in Bayesian network inference. Journal of Artificial Intelligence Research 5:301–328

Zhang W, Li H, Sun D, Zhou L (2012) A statistical assessment of the impact of agricultural land use intensity on regional surface water quality at multiple scales. Environmental research and public health 9:4170–4186, DOI 10.3390/ijerph9114170

Zou Q, Zhou J, Zhou C, Song L, Guo J (2013) Comprehensive flood risk assessment based on set pair analysis-variable fuzzy sets model and fuzzy AHP. Stochastic Environmental Research & Risk Assessment 27:525–546