Environmetrics 00, 1–22

DOI: 10.1002/env.XXXX

Applications of hybrid dynamic Bayesian networks to water reservoir management

R.F. Ropero[a*] , M.J. Flores[b], R. Rumí[c] and P.A. Aguilera[a]

Summary:

Bayesian networks (BNs) have been widely applied in environmental modelling to predict the behaviour of an ecosystem under conditions of change. However, this approximation doesn't take time into consideration. To solve this issue, an extension of BNs, the dynamic Bayesian networks (DBNs), has been developed in mathematics and computer science areas but has scarcely been applied in environmental modelling. This paper presents the application of DBN to water reservoir systems in Andalusia, Spain. The aim is to predict changes in the percent fullness of the reservoirs under the irregular rainfall patterns of Mediterranean watersheds. In comparison to static BNs, DBNs provide results that can be extrapolated to a particular time so that a climate change scenario can be studied in detail over time. Since results are expressed by density functions rather than unique values, several metrics are obtained from the results, including the probability of certain values. This allows the probability that water level in a reservoir reaches a certain level to be directly computed.

Keywords: Continuous variables; time series; TAN; naïve Bayes; Water Reservoir

1. INTRODUCTION

Over the last few decades, several new approaches have been developed and applied to environmental modelling (Kelly et al. 2013, Aguilera et al. 2011). Bayesian networks (BNs) are a novel statistic tool, which have demonstrated their ability to solve a wide range of

[a] *Informatics and Environment Laboratory, Dept. of Biology and Geology, University of Almería, Spain*
[b] *Dept. of Informatic Systems SIMD i^{3}A, University of Castilla-La Mancha, Campus Universitario, Albacete, Spain*
[c] *Dept. of Mathematics, University of Almería, Spain*
[*] *Correspondence to: R. F. Ropero. E-mail: rosa.ropero@ual.es*

This paper has been submitted for consideration for publication in *Environmetrics*

environmental problems under a framework of uncertainty (Landuyt et al. 2013, Barton et al. 2012). Defined in the beginning of the 1990s (Jensen & Andersen 1990), they have been extensively developed and applied in several scientific areas and a broad and consolidated literature is available. BNs have demonstrated their ability to solve several challenges in environmental modelling, such as the inclusion of information from several sources. Since their structure is based on Graph Theory, they can be directly interpreted by non-specialists and stakeholders who play an important role in the model learning and validation process (Voinov & Bousquet 2010). Besides, BNs have been developed to deal with large datasets and missing data, providing robust and accurate results (Fernandes et al. 2013). Even though they were first defined for discrete variables, real applications need to deal with data that are originally continuous. The most common solution for this is the discretization of these values, which implies a loss of accuracy (Uusitalo 2007). To solve this issue, BNs including the use of Gaussian models (Lauritzen 1992) were proposed, but these impose certain restrictions during the structural learning. Other models, such as the *Mixture of Truncated Exponential* models (Moral et al. 2001), were proposed to overcome such limitations and are able to deal with both discrete and continuous variables simultaneously.

Aguilera et al. (2011) pointed out that the most usual environmental problem BNs to which are applied is the study of the behavior of ecosystems under scenarios of change in which time is not taken into account. Nowadays, it is widely recognized that incorporating time in models is an important challenge in the field of data mining, reasoning and decision support systems (Russel & Norvig 2002). In the environmental sciences, time series analysis has a wide range of applications, and some models have been successfully applied (Davidson et al. 2016, Arya & Zhang 2015, Lagona et al. 2015). However, these models are usually based on specific technical concepts and notation that experts in environmental and ecological science not always deal with. This makes them hard to apply and specific literature is often difficult to find (von Asmuth et al. 2012).

Using traditional BNs, conclusions obtained cannot be extrapolated to a particular time, nor can time series be handled. For these reasons, the extension of BNs, the so-called Dynamic Bayesian networks (DBNs), has begun to be applied in environmental sciences to face this new challenge. In the work of Hill (2013), DBNs are applied to the control of streaming climatic data, in an attempt to detect anomalies and errors in the data. Zhang et al. (2012) used DBNs to integrate data from different times series into a model to accurately estimate the Leaf Area Index in a region of China. In both cases, the application of DBNs is focused on the pre-processing step, trying to correctly collect the data, or merge different data sets. In the paper of Molina et al. (2013), DBNs are learnt as a Decision Support System to predict, for the 2070-2100 period, the effects of Climate Change scenarios in a groundwater systems in Spain.

In general, static BNs have been extensively applied in water management (Aguilera et al. 2011) as a tool for decision support (Fienen et al. 2013), to make future predictions of changes under new management plans (Lowe et al. 2014) or scenarios of Climatic Change (Dyer et al. 2014). This is explained by the advantages that BNs provide as a tool for Decision Support System, which encouraged researchers to apply them to the Integrated Water Resource Management context (Castelletti & Soncini-Sessa 2007). This supposed the application of BNs in some European projects as the NeWater (Henriksen & Barlebo 2008). Modelling water resources, both superficial or groundwater systems, is a wide field of researching. In the case of Mediterranean areas, where the water is scarce and irregular, this topic becomes remarkable.

More specifically, in the case of Mediterranean watershed the irregular temporal and spatial rainfall patterns provokes periods of drought followed by events of strong storms and flood risk. As a solution, historically dams were constructed for water reservoir management, not only for human and agricultural consumption, but also for flood control and the maintenance of ecological flow in the river bed during drought periods. Currently, water management plans

need new tools able to provide information and prediction about the water reservoir under certain scenarios of change, even more under the current framework of Climatic Change.

The aim of this paper is to model the temporal behavior of the Water Reservoir System in Andalusia (Spain) through DBNs and show its applicability in environmental modelling. This is the first time that hybrid domains have been included in a DBN (both discrete and continuous variables in the same model) for environmental modelling.

2. METHODOLOGY

With the aim of modelling the behavior of the reservoir system (measured with the variable *Percent Fullness*), both static and dynamic BN models were learnt and compared in terms of error. Besides, to show the advantages of this methodology, a simple scenario of change was included (assuming an increase in the temperature accompanied by a decrease in rainfall) and some metrics were calculated for a better understanding of the results.

2.1. Bayesian networks and Dynamic Bayesian networks

A Bayesian network (BN) (Jensen & Nielsen 2007) is a statistical multivariate model for a set of random variables $X = \{X_1, \ldots, X_n\}$, which is defined in terms of two components;

- Qualitative component, a Direct Acyclic Graph (DAG) where each vertex represents one of the variables in the model, and the presence of an edge linking two variables (*i.e.* from variable X to variable Y, where X is a parent of Y) indicates the existence of statistical (in)dependence between them. It allows the model structure to be easily understood by experts which also can be included as a significant part of the model learning and validation processes (Voinov & Bousquet 2010).
- Quantitative component quantifying the relationships between the variables through the conditional distribution for each variable, given its parents in the graph. In the case

of discrete variables, this is expressed as a conditional probability table, while in the case of continuous variables it is expressed as a conditional density function.

BNs have the ability to represent the independencies between variables in the graph in a natural way through out the *d-separation* concept (Lauritzen 1996, Pearl 1988). This provides information about which variables are relevant or not for some other variable of interest and simplify the probability distribution of the variables necessary to specify the model. In general, in a DAG three types of relationships between variables are possible (Figure 1):

- Serial connections: Variable X_1 has a causal influence on variable X_3, which in turns, has an influence on variable X_5. So, information flows from X_1 to X_5 and viceversa. But, if we have information about X_3, any value of X_1 is irrelevant to our belied about X_5.
- Diverging connections: In that case, variable X_2 has a direct influence on both variables X_3 and X_4, and information flows from X_3 to X_4 and viceversa. But, again, if new information about X_2 is available, the state of variable X_3 is irrelevant to our belied about X_4, and viceversa.
- Converging connections: Variable X_3 is influenced by both X_1 and X_2, but they are irrelevant to each other.

[Figure 1 about here.]

Apart from providing information about the relevance of variables, these conditional independencies allow BNs to compact the representation of the joint probability and, therefore, facilitate the inference process.

BNs were initially proposed to deal with discrete or categorical data, thus when data include continuous values, they are commonly discretized to transform into categorical. However, this usually implies a loss in the statistical information included in the dataset. For that reason,

several methodologies have been developed in the BNs framework to deal with continuous variables. The first proposal was the *Gaussian models* (Lauritzen 1992) that yield appropriate results when all variables are continuous and follow a normal distribution. Even though this option can be applied where there are discrete variables as well, the discrete variables cannot have a continuous parent.

These constraints have led to the development of other alternatives such as the *Mixture of Truncated Exponentials* (MTEs) model, the *Mixtures of Polynomials* model and the *Mixtures of Truncated Basis Functions* model (for more information see Langseth et al. (2012), Shenoy & West (2011) and Moral et al. (2001)). Discretization is usually carried out by splitting the domain of the variable into several intervals, where the corresponding density function is approximated by a constant function that can also be seen as approximating the density function by a mixture of uniforms. However, if instead of constants, other functions were used, the accuracy of the approximation could be improved. This is the idea behind the MTE model, where exponential functions are used to estimate the density functions (for a detail information about MTE see Cobb et al. (2007), Rumí & Salmerón (2007), Rumí et al. (2006)). Another advantage of MTEs is that they are closed under restriction, marginalization and combination, so standard BNs inference processes can be applied. Up to now, MTEs and Gaussian models are the only alternatives that have been applied in environmental sciences (Maldonado et al. 2016, Meineri et al. 2015, Ropero, Rumí & Aguilera 2014, Aguilera et al. 2010).

BNs can cope with four different aims depending on the number and nature of the target variable(s) (Aguilera et al. 2011): Characterization, Inference, Classification and Regression. When the focus is on the behavior of one continuos variable, we are dealing with a *Regression* problem. The purpose is to predict this continuous goal variable as precisely as possible, returning its correct value, rather than trying to accurately model the joint probability of all the variables in the BN. To achieve this, fixed and constrained structures were mainly

developed for classification and regression tasks to accurately predict the class variable, so reducing the number of parameters that need to be estimated. These include the fixed structure naïve Bayes (NB) (Minsky 1963), and constrained structures like TAN (Friedman et al. 1997).

A NB is a fixed structure consisting of a BN with a single root node and a set of feature variables having only the root node as a parent. Its name comes from the fact that the features variables are independent given the root (Friedman et al. 1997). A step beyond this is to allow each feature to have one more parent besides the target variable, configuring a TAN structure. To learn this structure, the first step is to learn a directed tree structure with the features variables, using the mutual information with respect to the target variable. In the second step, the relationships between the target variable and each feature are included (Chow & Liu 1968). These relationships between features are not based on an ecological interpretation but on the amount of information they share with the target variable.

One of the main advantages of BNs is their ability to carry out efficient reasoning for a given scenario under conditions of uncertainty, called probability propagation or probabilistic inference. The objective is to obtain information about a set of variables of interest given known values (or evidences) of other variables (Shenoy & Shafer 1990). This provides an important advantage in environmental studies since a scenario of change can be introduced into the model to study the behaviour of the ecological system (Ropero et al. 2015). Furthermore, since not all variables must be evidenced to obtain a prediction, it is possible to introduce information about just a subset of variables, and update the probability of the remainder. In the case of fixed and constrained structures, their simpler structure and the *d-separation* concept allows to carry out the inference process in a direct and intuitive way.

However, time is not properly represented in static BNs even when links between variables can imply a temporal relationship in a certain way (Korb & Nicholson 2011), and future scenarios of change can be predicted with the inference process. For these reasons, static

BNs were extended to the so-called Dynamic Bayesian network in which a BN representing the static model is replicated in different time steps. Therefore, time is represented by means of links added between variables in different time steps.

The first attempt to deal with time using BNs appears in Provan (1993), which proposed their use for modelling a generic system in each time step, joining the BNs with links which represent the transition from one time to the next. In the DBN framework time is not included in the model as an input variable, instead, it is taken into account by representing the evolution of some variables over time, in a similar role as a transition matrix in a stochastic process. Even though timeline is continuous by nature, some approximations are carried out to build a DBN:

- The database or matrix the DBN is learnt from is composed by a sequence of observations time-sorted, with some predefined or constants time steps between them. This way, we focus on the case that observations happen in a discrete timeline.
- A basic assumption in this model is the *Markov assumption* (Murphy 2012, 2002). That is, the state of the world at a particular time depends on only a finite history of previous states. In the simplest case, the current state of the system depends only on the previous state, which is called a *first-order Markov process* (Figure 2)
- A last assumption comes from the characteristic of the probability distributions linking different time steps. In a DBN, it is assumed that it remains constant over time, which leads to a stationary, time-invariant or homogeneous model.

Using these 3 assumptions, a DBN can be formally defined as a a pair $(B_0, B_{\rightarrow})$, where B_0 is a BN over $X^{(0)}$ representing the initial distribution over states, and $B_{\rightarrow}$, is a 2-time-slice BN for the process (Koller & Friedman 2009). Thus, the term dynamic means that the system is changing over time, not that the network and the relationships between variables change (Murphy 2012, 2002). Therefore, DBN is composed of the following items (Korb & Nicholson 2011):

- Time slice: the state of the system at a particular time t, represented by a static BN identical in each time step, where the relationships between variables (*i.e.*, in Figure 2 links between X_t, Y_t and Z_t) are called intra-slice arcs.
- Inter-slice arcs: also called temporal arcs, they represent the relationships between variables at successive, or not successive, time slices both *(i)* the same variable over time (*i.e.* in Figure 2 links between Y_0 and Y_1) or *(ii)* between different variables over time (*i.e.* in Figure 2 links between Z_0 and Y_1). In order to reduce the potential number of temporal parents in the network, the *Markov assumption* is followed in its simplest case, the current state of the system depends only on the previous state, called a *first-order Markov process* (Figure 2). Thus, B_0 is represented twice, and the inter-slice or temporal arcs, are included to incorporate the evolution of the variables in time.

[Figure 2 about here.]

In Figure 2 the temporal aspect for X is captured by this probability distribution: $P(X_1|Y_1, X_0)$ in which variable X is influenced both by the current time slice and the previous one, following the *Markov assumption*.

In this way DBNs can be represented and solved by "static" models divided into different sub-models (model for time 0, model for time 1, and so on), which allows the available software and algorithms developed for static BNs to be used (for detailed information about learning and validation methodologies in BN, see for example Bookholt et al. (2014), Chen & Pollino (2012), Marcot (2012), Aguilera et al. (2011)). Moreover, both continuous and discrete data can be included since several models have been developed to represent this type of data within the BN framework. In the literature, there are several examples of DBN with hybrid data that uses *Gaussian* models (Wu et al. 2014). Even though they provide accurate results, the limitation they impose restrict their expansion to other areas and problems.

As in static BNs, evidence about a set of variables can be included into the model and update the posterior probability distributions for all the non-evidenced variables both in the

current and later time-slices. This is called *probabilistic projection* (Korb & Nicholson 2011). Given the special topology of DBNs, this *probability projection* is sometimes impossible to obtain due to the complexity of the problem. In those cases, the network can be *rolled out* over sequences of any length and visualize the system in more than two times slices (current situation and more than one step forward or backward). However, when the DBN is large, and the time interval between time slices is short, this process can be unfeasible. Another option is the *slide window* approach (Korb & Nicholson 2011) in which just the current and the following time slices are visualized.

2.2. Study area

Andalusia (Figure 3) is the second largest Autonomous Region of Spain, and the most-densely populated. The main characteristic of its annual water cycle is its irregularity. Rainfall patterns range from extremely heavy storms to prolonged periods of drought. For that reason, historically, reservoir construction has been the main solution to water scarcity and irregularity. Apart from urban water supply and agricultural consumption, the current system of reservoir has been designed to control and avoid the danger and loss from flood. Reservoir management also allows for the provision of a minimum water flow to maintain the natural river regime downstream during drought periods (Spanish Environmental Goverment 2007).

[Figure 3 about here.]

2.3. Data collection

Data were collected from the Water Quality Dataset from the Andalusian Environmental Information Network[†] (Andalusian Regional Government). A total of 61 reservoirs located

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

in the Guadalquivir and Guadalete-Barbate watershed were selected.

Data consist of 6 continuous and 1 discrete variables provided by the Regional Government as a monthly summary of data collected from the automated data network from October 1999 to September 2008. *Temperature* (°C) and *Rainfall* (m^3/m^2) represent the climatic conditions in the vicinity of the reservoir. *Percentage Evaporation* is the percentage of the reservoir capacity that evaporates. *Water level* indicates the height of the water column in m.a.s.l., whilst *Percent Fullness* expresses the percentage of the reservoir capacity that is currently used, from 0 to more than 100% (following a storm event, the reservoir can exceed the dam capacity). Finally, reservoir management is represented by Amount Discharge and Amount Transfer in. *Amount Discharge* (m^3) refers to the amount of water that is released for ecological, water consumption or regulation purposes. By contrast, *Amount Transfer in* (expressed as a discrete variable with three states: No transfer, less than $0.5m^3$ and more than $0.5m^3$) is the amount of water deliberately added to the reservoir, e.g., pumped in from another reservoir.

With this information two different datasets were created:

- Dataset organization for static models learning. Once the data are collected, variables for each month are merged into unique variables (*e.g.* in Figure 4(a), the variable *Temperature* is configured by taking the temperature data for october 1999, november 1999 and so on). This static dataset has 7 variables and 6588 observations and it was used for static models learning and validation.
- Dataset organization for dynamic models learning. For each reservoir data are organized into two-time slices, comprising pair of months (Figure 4(b)). This temporal dataset has 14 variables (temperature at time 0, temperature at time 1, rainfall at time 0, rainfall at time 1, and so on) and 6526 observations [‡]. This dataset was used for dynamic models learning and validation.

[‡]Note that the difference in the sample size in both dataset is due to the different organization of the data.

[Figure 4 about here.]

2.4. DBN learning

The objective is to predict, as accurately as possible, the behavior of the continuous variable *Percent Fullness*, which represents a *regression* task. Since features variables are both discrete and continuous, MTE models are used. Fixed and constrained structures, such as a NB and TAN, respectively, were used for both static and dynamic BNs. The static BN models consist of a single NB and TAN in which *Percent Fullness* variable is the root node, and the features are the rest of the variables. In the case of the DBN, these structures are repeated and connected through a temporal link between *Percent Fullness* at time 0 and *Percent Fullness* at time 1. Elvira software (Elvira-Consortium 2002) was used to learn and validate both static BN and DBN based on MTEs model.

Cross Validation (Stone 1974) was carried out to compute the *root mean square error (rmse)* from the test folds. It is a widely applied technique, in which the dataset is divided into two complementary datasets k times, one for learning, one for testing. Following this procedure, k different models are obtained, one for each different learning-dataset, and each one of them is validated with its complementary testing-dataset. In this way the whole dataset is used both for learning the model and for validating it, avoiding the overfitting problem Finally, the average error across all k folds is computed, using the *rmse* according to Equation 1. In both static and dynamic dataset, a *10-fold CV* is applied, so static and dynamic datasets are divided into 10 different pairs of train-test subdataset.

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

where n is the sample size; y_i, is the real value and $\hat{y}_i$ the predicted value obtained from the model.

2.5. Scenario of change

One of the main advantages of BNs is their ability to perform an scenario of change in which new information is included into the model, and the behavior of the rest of variables is studied. However, conclusions obtained from static BNs can not be extrapolated to a particular time. For that reason, a simple scenario is used to demonstrate the ability of DBN over BN in terms of future predictions.

Under the current climatic change framework, we used the model for predicting the behavior of *Percent Fullness* variable assuming that the temperature will rise by 10% and rainfall decrease by 15% in each time step (these values are quite drastic in order to see significant differences in the density function in only 2 steps). Firstly, the inference process is carried out over the model without any evidence observation to obtain the density functions *a priori*. Finally, we include the observed values (increase in temperature and decrease in rainfall) as evidences in variables *Temperature* and *Rainfall* both at time 0 and 1 at the NB dynamic model. Note that the rest of the feature values do not need to be evidenced. Once the evidences are included and the inference process is carried out, the updated density functions are obtained *a posteriori*.

From the water management point of view, it is often interesting to compute the probability that a reservoir reaches a certain level of Percent Fullness, both in the lowest and highest values. As an example, we compute the probability of values below 25% (left tail) and over 80% (right tail) of Percent Fullness (for a detailed explanation of how to compute the probability of a range of values, see Ropero, Rumí & Aguilera (2014)).

Since fixed and constrained structures were used, our DBN structure is not large or complex, so inference process, or *probabilistic projection*, can be carried out with standard

inference algorithms applied to static BNs. In our case, just the current (t) and the next ($t+1$) time slices are evaluated so neither *roll out* not *sliding window* approaches need to be applied.

2.6. Results

Figure 5 and 6 show the structure of both static BN and DBN for the case study based on NB and TAN structures. Table 1 shows the average *rmse* value of each model, obtained from the *10-fold Cross Validation*. Note that for the static models, *rmse* values are similar, but not in the case of the dynamic ones. Friedman's Test was performed for both static and dynamic models (Figure 7) to detect significant differences, returning that dynamic NB outperforms the rest of the models. Furthermore, results show that for the static models no significant differences are found. Comparing static and dynamic TAN even if the *rmse* is slightly lower for the dynamic model, the difference is not significant.

[Table 1 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

Even when static models seems to provide accurate results, dynamic models add an important advantage. Since results are expressed as a density function, not a unique value, and through the inference process, a scenario of change can be included and the probability function of the goal variable updated. This allows results to be deeply studied and compared between the situation *a priori* and under the scenario proposed (*a posteriori*) and their evolution over time. From these density functions, several metrics can be calculated, for example the mean, standard deviation, or even the probability of a certain range of values.

Figure 8 and Table 2 show the density function and the metrics obtained from *Percent Fullness* variable at time 0 and 1, both in the current situation (*a priori*), and under this scenario (*a posteriori*). *A priori*, both variables show a similar behavior, with a probability of both extreme values over 0.5 (in PF0, 0.25 and 0.33; in PF1, 0.06 and 0.50). However, when the proposed scenario is included, the probability of highest values (right tail) at time 0, increases from 0.33 to 0.43, and also the mean (from 59.74 to 69.77). By contrast, at time 1 the values tend to be more probable in the middle of the function, with a decrease in the probability of both right and left tails. This information is also confirmed by the behavior of the rest of the metrics in which standard deviation is reduced and the values are more concentrated around the mean.

From the environmental point of view, in the case of a rise in temperature and fall in rainfall, (which can be interpreted as a drought situation), the reservoir will be initially distributed from the smaller and secondaries dams to those that can collect a high amount of water reservoir and satisfied the water demand. Accordingly, at time 0, the values over 80% of Percent Fullness are more probable. If the scenario proposed persists, this would provoke a fall in the amount of water stored in the reservoir of the system being modelled.

[Figure 8 about here.]

[Table 2 about here.]

3. DISCUSSIONS AND CONCLUSIONS

In this paper, the theory behind BNs and DBNs models is explained, and their use is proposed for modelling temporal problems in environmental sciences. Through the study of the water reservoir system in Andalusia (Spain), both static and dynamic models based on constrained structures have been compared in terms of *rmse*.

One of the main advantages of BNs is that they provide not only a numeric prediction of the class variable but also its probability distribution, which allows several metrics to be calculated (*i.e.* mean, median, probability of a certain range of values) (Ropero, Rumí & Aguilera 2014). As Figure 8 shows, the target variable *Percent Fullness* can be studied in detail, its probability distribution, mean, standard deviation, or even the probability of extreme (tail) values. This is quite interesting from the management point of view since it allows, for example, computing the probability of having a low level of water in the reservoir, or by contrast, an amount exceeding its capacity. Also, through the inference process certain future predictions can be studied and the differences with respect to the *a priori* situation calculated. In static BNs application, the inference process allows changes in certain variables to be included, not necessarily in every feature, to check the behavior of the class variable (Ropero, Aguilera, Fernández & Rumí 2014). This has been applied to model the behavior of ecosystems under different scenarios, *i.e.* climatic change scenarios, global environmental change scenarios, management decision scenarios, among others (Mantyka-Pringle et al. 2014, Webster & McLaughlin 2014), but the conclusions obtained from this inference process in the static BN cannot been extrapolated to a particular time. Using DBN we can expand the model and obtain a similar conclusion for a particular time. As in the case study, the behavior of a system is studied at different times (current time and one month later). Even when the scenario proposed was designed with a drastic change in climatic conditions (in order to see significant differences in the density functions), this methodology can be applied to several environmental cases and also, roll out the model and check the behavior of the system in more than two time steps. In spite of this advantage over static BNs, the study of scenarios in DBNs has been developed and applied in other areas, but not in environmental science. A further effort is needed in that field.

Another advantage that has been demonstrated is the ability to include in the same model both discrete and continuous variables through the use of the *MTE* model. There is so far

no application in environmental sciences in which DBN is learnt using hybrid domains. In other areas, DBN include both discrete and continuous variables through the use of *Gaussian* models, but this is not a suitable option when we need to have some freedom in the structural learning. In this paper, continuous and discrete variables have been used in both static and dynamic models with no prior limitation in the structure or the parameter estimation due to the MTE models. See for example, Figure 5, the TAN structure shows a discrete variable (*Amount Transfer in*) with two continuous variables as parents (*Percent Fullness* and *Amount Discharge*), which would not be possible to learn using *Gaussian* models.

However, some challenges of the DBN application in environmental sciences have been identified. First, timeline is discretized in a set of time steps, and following the *Markov assumption*, just the previous step has an impact on the current time step. In each case, we should consider whether this approximation is suitable or not for our data and the problem itself. Environmental data from different areas (ecology, biodiversity, water resources) differs on their properties and characteristics, so expert should, if possible, decide what is the best time discretization (*i.e.* one day, a week, a month). Secondly, during the inference process two approaches are available, the *slide window* and *roll out*. Again, depending on our environmental problem and available data, more than two time-steps would be necessary to study. But, if the network is large, and time interval short, this process can be unfeasible. Even when some algorithms have been proposed (*i.e.* the Kalman Filter (Kalman 1960) or a version of the junction tree algorithm for DBNs (Kjærulff 1995), for a detail explanation see Korb & Nicholson (2011)), a further effort is needed to find suitable options for the *probabilistic projection* process in environmental modelling.

In spite of these challenges, and the limited number of papers in the environmental field that have applied DBN (Hill 2013, Molina et al. 2013, Nicholson & Flores 2011), it is clear that this new tool comprises a promising methodology with clear applications. While it has been mainly applied to climatic and (ground) water data, it can be extended to any time

series in the environmental field.

ACKNOWLEDGEMENTS

Thanks to the anonymous reviewers and A. Salmerón for their comments which help to improve this manuscript. This work has been supported by the Spanish Ministry of Economy and Competitiveness through projects TIN2013-46638-C3-1-P, TIN2013-46638-C3-3P, by the Junta de Andalucía through project P12-TIC-2541, and from ERDF funds. R. F. Ropero is supported by the FPU research grant, AP2012-2117, funded by the Spanish Ministry of Education, Culture and Sport.

REFERENCES

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

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

Arya, F. K. & Zhang, L. (2015), 'Time series analysis of water quality parameters an Stillaguamish river using order series method', *Stochastic Environmental Research & Risk Assessment* **29**, 227–239.

Barton, D. N., Kuikka, S., Varis, O., Uusitalo, L., Henriksen, H. J., Borsuk, M., de la Hera, A., Farmani, R., Johnson, S. & Linnell, J. D. (2012), 'Bayesian Networks in Environmental and Resource Management', *Integrated Environmental Assessment and Management* **8**, 418–429.

Bookholt, F. D., Stuurman, P. & Hanea, A. M. (2014), 'Practical Guidelines for Learning Bayesian Networks from Smalls Data Sets', *Open Access Library Journal* **1**, 1–13.

Castelletti, A. & Soncini-Sessa, R. (2007), 'Coupling real-time control and socio-economic issues in participatory river basin planning', *Environmental Modelling & Software* **22**, 1114–1128.

Chen, S. H. & Pollino, C. A. (2012), 'Good practice in Bayesian network modelling', *Environmental Modelling & Software* **37**, 134–145.

Chow, C. K. & Liu, C. N. (1968), 'Approximating discrete probability distributions with dependence trees', *IEEE Transactions on Information Theory* **14**, 462–467.

Cobb, B., Rumí, R. & Salmerón, A. (2007), *Bayesian networks models with discrete and continuous variables*, Advances in probabilistic graphical models, chapter Studies in Fuzziness and Soft Computing, pp. 81–102.

Cuaya, G., Muñoz Meléndez, A., Nuñez Carrera, L., Morales, E. F., Quiñones, I., Pérez, A. I. & Alessi, A. (2013), 'A dynamic Bayesian network for estimating the risk of falls from real gait data.', *Med. Biol. Eng. Comput.* **51**, 29–37.

Davidson, J. E., Stephenson, D. B. & Turasie, A. A. (2016), 'Time series modeling of paleoclimate data', *Environmetrics* **27**, 55–65.

Dyer, F., ElSawah, S., Croke, B., Griffiths, R., Harrison, E., Lucena-Moya, P. & Jakeman, A. J. (2014), 'The effects of climate change on ecologically-relevant flow regime and water quality attributes', *Stochastic Environmental Research & Risk Assessment* **28**, 67–82.

Elvira-Consortium (2002), Elvira: An Environment for Creating and Using Probabilistic Graphical Models, *in* 'Proceedings of the First European Workshop on Probabilistic Graphical Models', pp. 222–230.
URL: *http://leo.ugr.es/elvira*

Fernandes, J. A., Lozano, J. A., Inza, I., Irigoien, X., Pérez, A. & Rodríguez, J. D. (2013), 'Supervised pre-procesing approaches in multiple class variables clasification for fish recruitment forecasting', *Environmental Modelling & Software* **40**, 245–254.

Fienen, M. N., Masterson, J. P., Plant, N. G., Gutierrez, B. T. & Thieler, E. R. (2013), 'Bridging groundwater models and decision support with a bayesian network', *Water Resource Research* **49**, 6459–6473.

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

Henriksen, H. J. & Barlebo, H. C. (2008), 'Reflections on the use of Bayesian belief networks for adaptive management', *Journal of Environmental Management* **88**, 1025–1036.

Hill, D. J. (2013), 'Automated Bayesian quality control of streaming rain gauge data', *Environmental Modelling & Software* **40**, 289–301.

Jensen, F. & Andersen, S. (1990), Approximations in Bayesian belief universes for knowledge-based systems, *in* 'Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence', pp. 162–169.

Jensen, F. V. & Nielsen, T. D. (2007), *Bayesian Networks and Decision Graphs*, Springer.

Kalman, R. (1960), 'A new approach to linear filtering and prediction problems', *Trans.ASME, J.Basic Engineering* **82**, 34–45.

Kelly, R., Jakeman, A. J., Barreteau, O., Borsuk, M., ElSawah, S., Hamilton, S., Henriksen, H. J., Kuikka, S., Maier, H., Rizzoli, E., Delden, H. & Voinov, A. (2013), 'Selecting among five common approaches for integrated environmental assessment and management', *Environmental Modelling & Software* **47**, 159–181.

Kjærulff, U. (1995), 'dhugin: A computational system for dynamic time-sliced bayesian networks', *International Journal of Forecasting. Special Issue on Probability Forecasting* **11**, 89–111.

Koller, D. & Friedman, N. (2009), *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press.

Korb, K. B. & Nicholson, A. E. (2011), *Bayesian Artificial Intelligence*, CRC Press.

Lagona, F., Picone, M. & Maruotti, A. (2015), 'A hidden mark model for the analysis of cylindrical time series', *Environmetrics* **26**, 534–544.

Landuyt, D., Broekx, S., D'hondt, R., Engelen, G., Aertsens, J. & Geothals, P. (2013), 'A review of Bayesian belief networks in ecosystem service modelling', *Environmental Modelling & Software* pp. 1–13.

Langseth, H., Nielsen, T. D., Rumí, R. & Salmerón, A. (2012), 'Mixtures of Truncated Basis Functions', *International Journal of Approximate Reasoning* **53**(2), 212–227.

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

Lauritzen, S. L. (1996), *Graphical Models*, London: Oxford University Press.

Lowe, C. D., Gilbert, A. J. & Mee, L. D. (2014), 'Human-environment interaction in the baltic sea', *Marine Policy* **43**, 46–54.

Maldonado, A. D., Aguilera, P. A. & Salmerón, A. (2016), 'Continuous Bayesian networks for probabilistic environmental risk mapping', *Stochastic Environmental Research & Risk Assessment* **30**, 1441–1455.

Mantyka-Pringle, C. S., Martin, T. G., Moffatt, D. B., Linke, S. & Rhodes, J. R. (2014), 'Understanding and predicting the combined effects of climate change and land-use change on freshwater macroinvertebrates and fish', *Journal of Applied Ecology* **51**, 572–581.

Marcot, B. (2012), 'Metrics for evaluating performance and uncertainty of Bayesian network models', *Ecological Modelling* **230**, 50–62.

Meineri, E., Dahlberg, C. J. & Hylander, K. (2015), 'Using Gaussian Bayesian Network to disentangle direct and indirect associations between landscape physiography, environmental variables and species distribution', *Ecological Modelling* **313**, 127–136.

Minsky, M. (1963), 'Steps towards artificial intelligence', *Computers and Thoughts* pp. 406 – 450.

Molina, J. L., Pulido-Veláquez, D., García-Aróstegui, J. & Pulido-Velázquez, M. (2013), 'Dynamic Bayesian Network as a Decision Support tool for assessing Climate Change impacts on highly stressed groundwater systems', *Journal of Hydrology* **479**, 113–129.

Moral, S., Rumí, R. & Salmerón, A. (2001), Mixtures of Truncated Exponentials in Hybrid Bayesian Networks, *in* 'ECSQARU'01. Lecture Notes in Artificial Intelligence', Vol. 2143, Springer, pp. 156–167.

Murphy, K. (2012), *Machine learning. A probabilistic Perspective*, The MIT Press.

Murphy, K. P. (2002), Dynamic Bayesian Networks: Representation, Inference and Learning, PhD thesis, University of California, Berkeley.

Nicholson, A. & Flores, J. (2011), 'Combining state and transition models with dynamic Bayesian networks', *Ecological Modelling* **222**, 555–566.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann.

Provan, G. M. (1993), Tradeoffs in Constructing and Evaluating Temporal Influence Diagrams, *in* 'Proceedings of the 9th Conference of the Uncertainty in Artificial Intelligence', pp. 40–47.

Ropero, R. F., Aguilera, P. A., Fernández, A. & Rumí, R. (2014), 'Regression using hybrid Bayesian networks: Modelling landscape-socioeconomy relationships', *Environmental Modelling & Software* **57**, 127–137.

Ropero, R. F., Aguilera, P. A. & Rumí, R. (2015), 'Analysis of the socioecological structure and dynamics of the territory using a hybrid Bayesian network classifier', *Ecological Modelling* **311**, 73–87.

Ropero, R. F., Rumí, R. & Aguilera, P. A. (2014), 'Modelling uncertainty in social-natural interactions', *Environmental Modelling & Software* **75**, 362–372.

Rumí, R. & Salmerón, A. (2007), 'Approximate probability propagation with mixtures of truncated exponentials', *International Journal of Approximate Reasoning* **45**, 191–210.

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

Russel, S. & Norvig, P. (2002), *Artificial Intelligence: A Modern Approach*, Pearson, chapter Probabilistic reasoning over time, pp. 542–583.

Shenoy, P. P. & Shafer, G. (1990), Axioms for probability and belief functions propagation, *in* R. Shachter, T. Levitt, J. Lemmer & L. Kanal, eds, 'Uncertainty in Artificial Intelligence, 4', North Holland, Amsterdam, pp. 169–198.

Shenoy, P. P. & West, J. C. (2011), 'Inference in hybrid Bayesian networks using mixtures of polynomials', *International Journal of Approximate Reasoning* **52**(5), 641–657.

Spanish Environmental Goverment (2007), Plan especial de actuación en situaciones de alerta y eventual sequía de la cuenca hidrográfica del guadalquivir, Technical report, Ministerio de Medio Ambiente.

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

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

Voinov, A. & Bousquet, F. (2010), 'Modelling with stakeholders', *Environmental Modelling & Software* **24**, 1268–1281.

von Asmuth, J. R., Maas, K., Knotters, M., Bierkens, M. F. P., Bakker, M., Olsthoorn, T., Cirkel, D. G., Lenunk, I., Schaars, F. & von Asmuth, D. C. (2012), 'Software for hydrogeologic time series analysis, interfacing data with physical insight', *Environmental Modelling & Software* **38**, 178–190.

Webster, K. L. & McLaughlin, J. W. (2014), 'Application of a bayesian belief network for assessing the vulnerability of permafrost to thaw and implications for greenhouse gas production and climate feedback', *Environmental Science /& Policy* **38**, 28–44.

Wu, X., Wen, X., Li, J. & Yao, L. (2014), 'A new dynamic bayesian network approach for determining effective connectivity from fMRI data', *Neural Computing & Applications* **24**, 91–97.

Zhang, Y., Qu, Y., Wan, J., Liang, S. & Liu, Y. (2012), 'Estimating leaf area index from MODIS and surface meteorological data using a dynamic bayesian network', *Remote Sensing of Environment* **127**, 30–43.

FIGURES

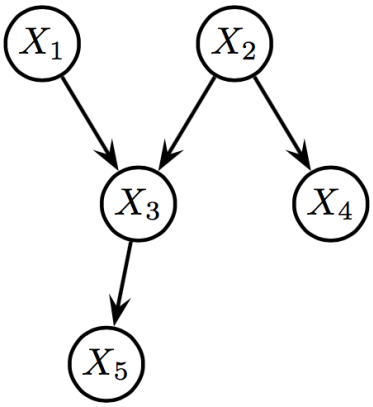


Figure 1. Example of the *d-separation* concept trough out a BN with five variables in which the three types of relationships are shown: serial, diverging and converging connections.

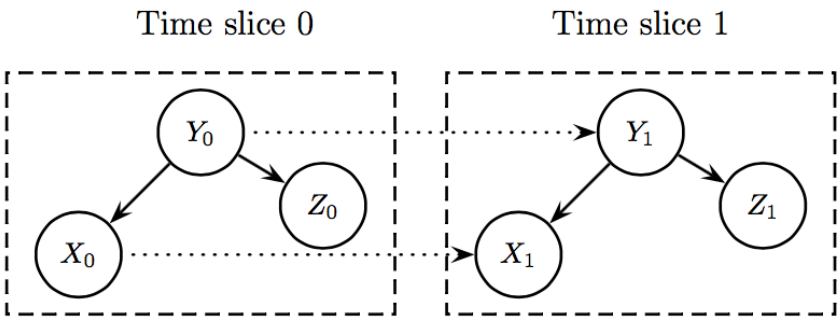


Figure 2. Example of a Dynamic Bayesian network following the *first-order Markov assumption* with a fixed naïve Bayes structure with two features, X and Z, and a class variable, Y composed of 2 time slices. Solid links represent intra-slice arcs, whilst dotted lines represent inter-slice arcs.

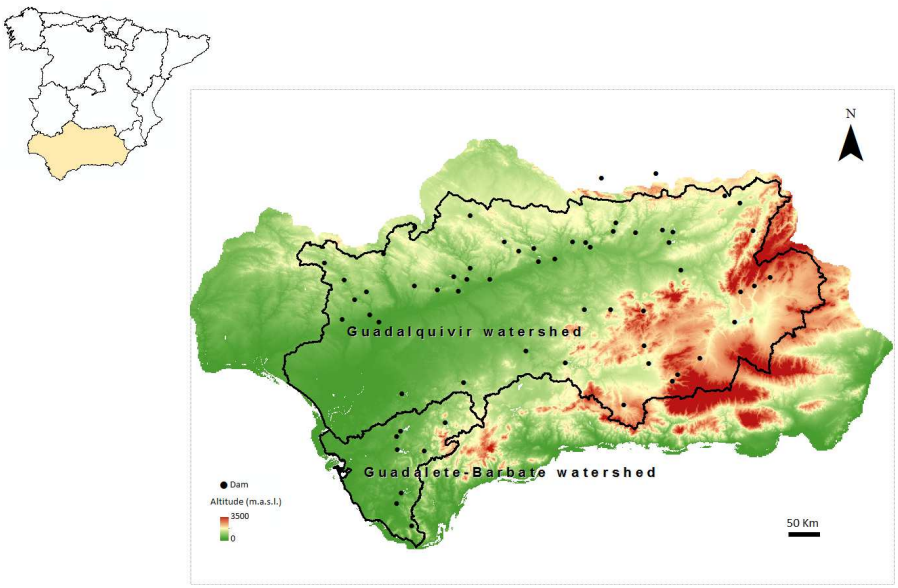


Figure 3. Relief map of Andalusia showing the watersheds and the reservoirs selected for the case study.

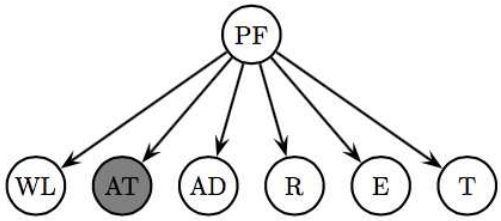
Dam	T	R	...
1	$T_{oct1999}$	$R_{oct1999}$	...
2	$T_{oct1999}$	$R_{oct1999}$	...
...	$T_{oct1999}$	$R_{oct1999}$	...
1	$T_{nov1999}$	$R_{nov1999}$	...
2	$T_{nov1999}$	$R_{nov1999}$	...
...	$T_{nov1999}$	$R_{nov1999}$	...
1	$T_{dec1999}$	$R_{dec1999}$	...
2	$T_{dec1999}$	$R_{dec1999}$	...
...	$T_{dec1999}$	$R_{dec1999}$	...

(a) Dataset for Static models

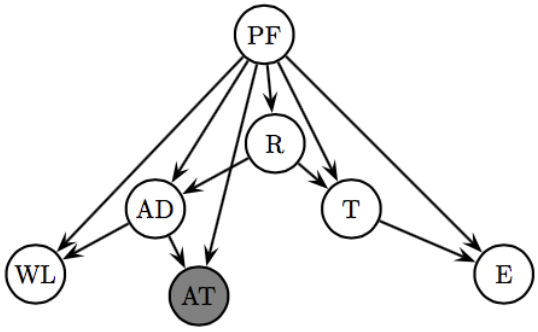
Dam	T_0	R_0	...	T_1	R_1	...
1	$T_{oct1999}$	$R_{oct1999}$	...	$T_{nov1999}$	$R_{nov1999}$	...
1	$T_{nov1999}$	$R_{nov1999}$	...	$T_{dec1999}$	$R_{dec1999}$	...
...	...	...	...	...	...	...
2	$T_{oct1999}$	$R_{oct1999}$	...	$T_{nov1999}$	$R_{nov1999}$	...
2	$T_{nov1999}$	$R_{nov1999}$	...	$T_{dec1999}$	$R_{dec1999}$	...
...	...	...	...	...	...	...

(b) Dataset for Dynamic models

Figure 4. Example of both datasets (for the static (a) and dynamic (b) models) for the *Temperature* (T) and Rainfall (R) variables.

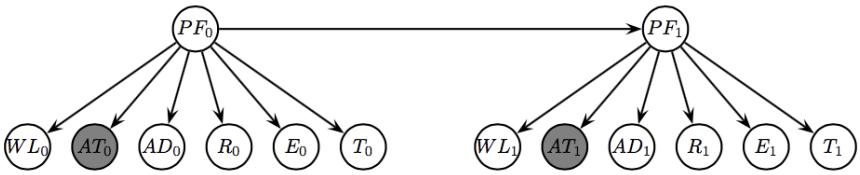


(a) Static NB

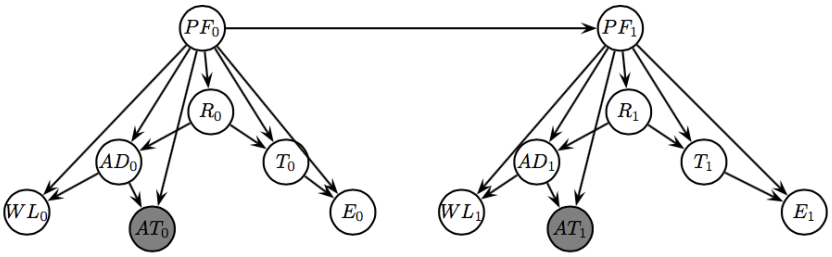


(b) Static TAN

Figure 5. Static naïve Bayes (a) and TAN (b) structures for the reservoir example. Discrete variable is filled in gray. PF, Percent Fullness; T, Temperature; R, Rainfall; E, Percentage Evaporation; AD, Amount Discharge; AT, Amount Transfer in; WL, Water Level.



(a) Dynamic NB



(b) Dynamic TAN

Figure 6. Dynamic naïve Bayes (a) and TAN (b) structures for the reservoir example. Discrete variables are filled in gray. PF, Percent Fullness; T, Temperature; R, Rainfall; E, Percentage Evaporation; AD, Amount Discharge; AT, Amount Transfer in; WL, Water Level.

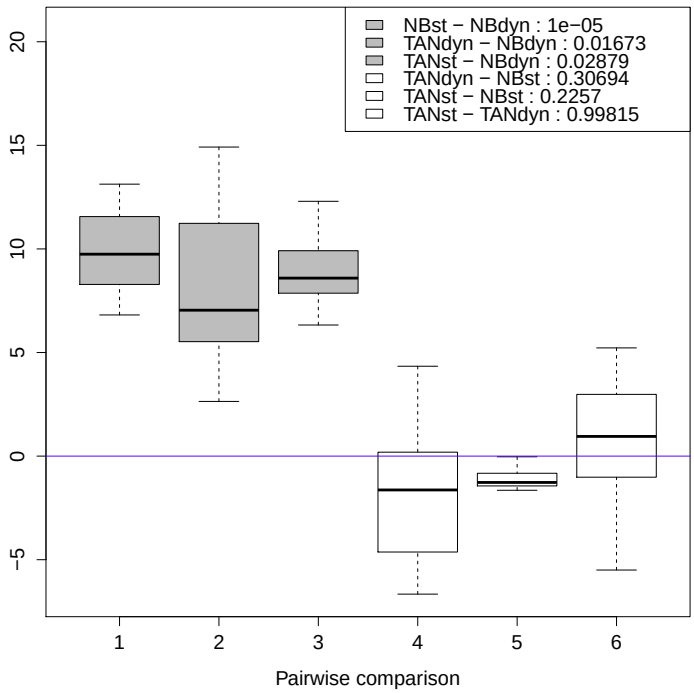
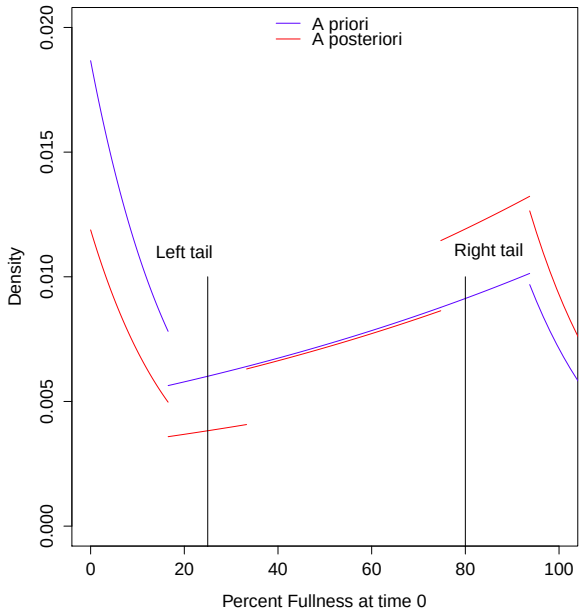
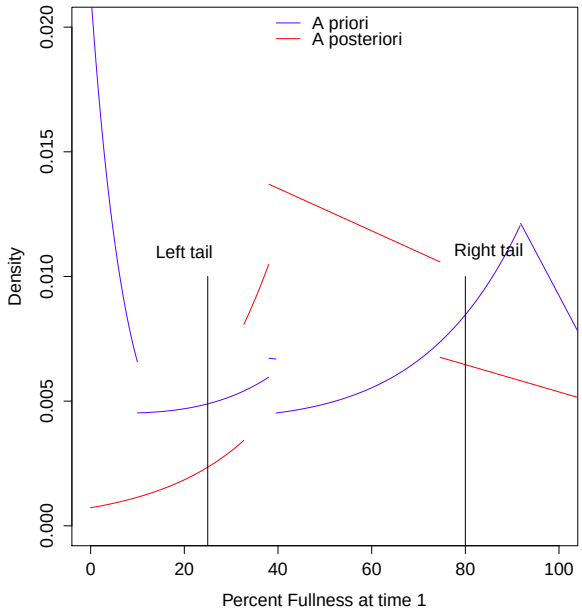


Figure 7. Box-plot summarizing the results of the pairwise comparison between static (a) and dynamic (b) regression models, p-values are shown in the legend. The gray-shaded boxes indicate significant differences between the corresponding models.



(a) Percent Fullness at time 0



(b) Percent Fullness at time 1

Figure 8. Probability distribution functions of *Percent Fullness* at time 0 (PF0) and 1 (PF1) variables in dynamic naïve Bayes (NB). Note that probability functions are defined as a piecewise function using MTEs.

TABLES

Table 1. Values for the *rmse* calculated by means of a *10-fold Cross Validation* for each method. NB, Bayesian networks based on naïve Bayes structure; TAN, Bayesian networks based on TAN structure.

Model	Static models	Dynamic models
NB	35.68	25.82
TAN	34.62	33.93

Table 2. Metrics calculated from the density functions of variables *Percent Fullness* at time 0 (PF0) and 1 (PF1) in both *a priori* and *a posteriori* situations. SD, Standard Deviation.

	A priori				A posteriori			
Variable	Mean	SD	$P(x \leq 25)$	$P(x \geq 80)$	Mean	SD	$P(x \leq 25)$	$P(x \geq 80)$
PF0	59.74	40.18	0.25	0.33	69.77	39.04	0.16	0.43
PF1	61.11	64.70	0.06	0.50	89.45	58.45	0.02	0.41