Running title: Bayesian networks in predictive ecology
Number of words: ∼6778
Number of tables: 5
Number of figures: 2
Number of references: ??

Exploiting the Full Potential of Bayesian Networks in Predictive Ecology

Pouria Ramazi[1,2,*], **Mélodie Kunegel-Lion**[3], **Russell Greiner**[2], **& Mark A. Lewis**[1,3]

1. *Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, T6G 2G1 Canada*
2. *Department of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8 Canada*
3. *Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E9 Canada*

*Correspondence author. ramazi@ualberta.ca

Date of submission: 5th January 2021

Summary

1. Although ecological models used to make predictions from underlying covariates have a record of success, they also suffer from limitations. They are typically unable to make predictions when the value of one or more covariates is missing during the testing. Missing values can be estimated but methods are often unreliable and can result in poor accuracy. Similarly, missing values during the training can hinder parameter estimation of many ecological models. Bayesian networks can handle these and other limiting issues, such as having highly correlated covariates. However, they are rarely used to their full potential.
2. Indeed, Bayesian networks are commonly used to evaluate the knowledge of experts by constructing the network manually and often (incorrectly) interpreting the resulting network causally. We provide an approach to learn a Bayesian network fully from observed data, without relying on experts and show how to appropriately interpret the resulting network, both to identify how the variables (covariates and target) are interrelated and to answer probabilistic queries.
3. We apply this method to the case-study of a mountain pine beetle infestation and find that the trained Bayesian network has a predictive accuracy of 0.88 AUC. We classify the covariates as primary and secondary in terms of contributing to the prediction and show that the predictive accuracy does not deteriorate when the secondary covariates are missing and degrades to only 0.76 when one of the primary covariates is missing.
4. As a complement to the previous work on constructing Bayesian networks by hand, we show that if instead, both the structure and parameters are learned only from data, we can achieve more accurate predictions as well as generate new insights about the underlying processes.

Keywords: Bayesian network, structure learning, automatic learning, pest, mountain pine beetle, invasive species, machine learning, risk modeling

1 Introduction

Predictions are essential in aquatic and terrestrial ecology, whether the focus lies in changes in ecosystem composition, structure, and richness to preserve the biodiversity and ecosystem function, or in the spatial distribution of individuals and species to inform conservation and invasive species policies. The field of predictive ecology focuses on how to make such predictions, particularly in the context of climate change, and has grown exponentially since the 1990s, given the quality and quantity of available ecological data (Purves *et al.*, 2013; Mouquet *et al.*, 2015). Simple and advanced statistical and machine-learning approaches have been used to this end, and some have reported great success. Commonly applied models include mechanistic equations, individual-based models, generalized linear models (Aukema *et al.*, 2008; Preisler *et al.*, 2012), generalized additive models, MaxEnt (Merow *et al.*, 2013), decision trees, support vector machines, and artificial neural networks, (Marmion *et al.*, 2009; Youssef *et al.*, 2016).

These standard models, however, lack some practical features, which questions their use as predictors. They are unable to make predictions when the value of a covariate is missing, a typical issue because some covariates are expensive or logistically impossible to collect. To impute the missing values can be unreliable as modelling assumptions are needed so as to "guess" them. The assumptions may even conflict with those posed by the original model using the imputed values. Another approach is to produce a model that does not involve any covariate that is ever missing. This can be problematic as well, because *(i)* those covariates are not fixed in the area of interest: the value of a covariate may be missing at location A, but present at location B, and the opposite may hold for another covariate; and *(ii)* even if a covariate is only measured in the lab and never on the field, incorporating it in the model can still reveal its effect on the response variable. Most models also cannot reveal the co-effect of more than one covariate on the response variable, and some do not allow for statistical inference. Moreover, those that are used for statistical inference cannot handle correlated covariates.

Bayesian networks (BNs) can deal with these issues. They are directed acyclic graphs, whose

nodes are the response variable and covariates, and the links between the nodes show how these nodes are related to each other. Both links from covariate to response and from covariate to covariate are allowed in the network. BNs are graphical, and hence often simpler to understand than complex systems of equations (*e.g.*, Bode *et al.*, 2017; Eklöf *et al.*, 2013; Troyanskaya *et al.*, 2003; Rish *et al.*, 2009), deepening our understanding of natural phenomena as well as allowing for accurate predictions. However, there are two main issues with how BNs are typically applied in practice: *(i)* they are rarely used to their full potential, and *(ii)* they are misinterpreted as *causal* networks. The common practice of applying BNs is to manually construct the structure (network), based on the knowledge of experts, then either set the parameters manually or learn them from data, and finally, read the links as causal relationships in the resulting BN. Although useful in assessing the qualitative descriptions of an ecological process, this approach relies heavily on our prior understanding of the process, and hence, is only as good as our understanding. If, instead, both the structure and parameters of the BN are learned only from the data, there will be room for more accurate predictions as well as new insights about the process. Moreover, BNs are not causal networks, but essentially a set of conditional (in)dependencies that factorize the joint probability distribution of all of the variables. Causal deductions, hence, may not be made, although some hypotheses may be tested.

We complement previous studies on BNs that used the knowledge of experts (Marcot *et al.*, 2006; Chen & Pollino, 2012) by focusing on learning the structure, and proper model interpretation in the form of conditional probabilistic inferences rather than causal deductions. The goal of this paper is *(i)* to discuss the advantages of different ecological modeling approaches, and highlight what BNs can offer in this context; *(ii)* to provide a systematic approach for training a BN completely from data, without incorporating the prior knowledge of experts, and then evaluating and interpreting the resulting BN; and *(iii)* to apply this method to the case study of a mountain pine beetle (MPB) outbreak.

2 Materials and Methods

2.1 Advantages of Bayesian networks

In what follows, we first briefly introduce BNs and then compare them with other modeling approaches in predictive ecology (Table 1). Here, we focus on the "typical" situation with each model; for example, the prediction accuracy of a properly trained BN being typically high does not imply that it is always higher or even as high as other highly accurate models.

Table 1 Comparison of models in predictive ecology. See Sections 2.1.2 to 2.1.9 for explanations of the model characteristics.

Model	Capable of generative learning	Handles missing values at training	Handles missing values at testing	Level of non-linearity that can be handled	Allows for statistical inference and model selection	Tolerates correlated variables in statistical inference	Provides insights on variables co-effects	Ease of incorporation of prior knowledge	Does not need prior knowledge	Predictive accuracy	Marginalizable
Mechanistic equation	•			●			●	●			•
Individual-based model				●			●	●			
Generalized linear model					●			•	•	•	
Generalized additive model				•	●			•	•		
MaxEnt			•					•		●	
Decision tree				●				•	●	●	
Support vector machine (linear)					●				•	•	
Neural network				●					●	●	
Bayesian network	●	●	●	●	●	●	●	●	●	●	●

(empty) typically low • typically medium ● typically high

2.1.1 *Introduction to Bayesian networks*

Given a set of n random variables $\mathcal{Z} = \{Z_i\}_{i=1}^{n}$ (consisting of the response variable and $n-1$ covariates), a BN factorizes the joint probability $P(\mathcal{Z})$ according to a specified directed acyclic

graph whose nodes are the variables $\mathcal{Z}$, following the equation

$$P(\mathcal{Z}) = \prod_{i=1}^{n} P(Z_i \,|\, \mathrm{Pa}_{Z_i}), \tag{1}$$

where Pa_{Z_i} denotes the *parents* of Z_i in the graph, *i.e.*, those nodes that have an outgoing edge that leads to Z_i (Fig. 1 in SI). The individual factors $P(Z_i \,|\, \mathrm{Pa}_{Z_i})$ are known as *conditional probability distributions (CPDs)* (Koller & Friedman, 2009). A BN encodes the claim that given the *Markov blanket* $\mathrm{MB}(Z_i)$ of a node Z_i – which is the set of its parents, children and the other parents of its children – the node becomes independent from the remaining of the nodes, written $Z_i \perp (\mathcal{Z} - Z_i - \mathrm{MB}(Z_i)) \,|\, \mathrm{MB}(Z_i)$. This provides the essentials for understanding how the variables relate to each other. We, therefore, refer to the nodes in the Markov blanket of the target node Y_i as *primary* covariates and to others as *secondary*. The estimation of the target node based on the values of the primary covariates does not change if the values of the secondary covariates are additionally known. The conditional independencies also reduce the number of parameters needed to represent the joint distribution $P(\mathcal{Z})$. It is possible to learn from data, both the graph and the CPDs, known as the *structure* and *parameters* of the BN (Section 2.2).

The factorization in Eq. (1) is sufficient to define BNs and draws a clear line between BNs and causal networks. To explain, assume that we are modeling the co-occurrence of two competitive species, with densities denoted by X_1 and X_2, each corresponding to a node in a BN. We could link these two distributions using either a directed edge from X_1 to X_2, decomposing the joint density distribution of the two species as $P(X_1, X_2) = P(X_1)P(X_2 \,|\, X_1)$, or a directed edge from X_2 to X_1, resulting in $P(X_1, X_2) = P(X_2)P(X_1 \,|\, X_2)$. The first relies on the distribution of X_1 and the conditional distribution of X_2 given X_1, and the reverse holds for the second. Both of these models can be used to make acceptable predictions if one can estimate parameters $P(X_2 \,|\, X_1)$ and $P(X_1 \,|\, X_2)$ effectively. However, none of the models are causal: neither of X_1 or X_2 is causing the other. The edge simply means probabilistic dependence and dictates the

factorization of the joint distribution. Now, assume that the species distributions are each partly "caused" by a third variable *vegetation*, denoted by V. Should we construct a BN based on this "causal understanding", we would add the node V and link it to both X_1 and X_2 without connecting the two. This results in the joint probability distribution of the two species and vegetation

$$P(X_1, X_2, V) = P(V)P(X_1 \mid V)P(X_2 \mid V).$$

However, this is not the only way to model the joint distribution. Depending on the training data, one may obtain a more accurate model in terms of data fitting by also linking X_1 to X_2 (or *vice versa*), resulting in

$$P(X_1, X_2, V) = P(V)P(X_1 \mid V)P(X_2 \mid X_1, V).$$

This might be because vegetation is not the only cause of the two, and another factor, say temperature, also plays a role, which is not included in our variable list but is highly correlated with X_1, and hence, provides a better estimation of X_2 by linking the two. One may yet use a different model, where X_1 and X_2 are not linked to each other but both linked to V, resulting in

$$P(X_1, X_2, V) = P(V \mid X_1, X_2)P(X_1)P(X_2).$$

This is particularly useful if we know the distributions of the species densities, *i.e.,* $P(X_1)$ and $P(X_2)$, but not that of vegetation $P(V)$, and we know how vegetation can be estimated based on the distribution of the two species, *i.e.,* $P(V \mid X_1, X_2)$. None of the links in this model are causal.

2.1.2 Generative vs discriminative learning

Consider the response variable Y and set of covariates (features) $\mathcal{X}$ that are used to estimate Y. One may pursue either of the two learning tasks with respect to these variables: *generative*, that is to learn the joint probability distribution $P(Y, \mathcal{X})$, or *discriminative*, that is to learn the conditional probability $P(Y \mid \mathcal{X})$. The joint probability distribution $P(Y, \mathcal{X})$ represents the probability of any given assignment to all of the variables Y and $\mathcal{X}$ in the data, or loosely speaking, how all the variables are related to each other. On the other hand, the conditional probability $P(Y \mid \mathcal{X})$ represents the probability of Y happening given $\mathcal{X}$, or in other words, in which cases does Y happen. So discriminative learning focuses only on the probability of the response variable whereas generative learning also reveals the probability of the covariates. For example, an ecologist may be interested in two species' co-occurrence, which is a generative question given by the distribution $P(X_1, X_2)$, were X_1 and X_2 are the density of the species. On the other hand, the same ecologist may be interested in whether the density of species X_1 (as a response variable) can be estimated using that of species X_2 (as the covariate), which is a discriminative question, given by $P(X_1 \mid X_2)$.

Note that knowing the "true" joint distribution $P(Y, \mathcal{X})$ allows knowing the conditional distribution $P(Y \mid \mathcal{X})$. However, because small errors in estimating $P(Y, \mathcal{X})$, which typically happen in practice, might lead to large errors in the associated values of $P(Y \mid \mathcal{X})$ (Ng & Jordan, 2002), each learning task deserves its own treatment. Although potentially capable of modelling the joint probability distribution, mechanistic models are not commonly used for this purpose as it would require a great deal of prior knowledge of the process. Roughly speaking, none of the models in Table 1, except for BNs, are effective at generative learning.

2.1.3 Missing data

Datasets often have many instances (observations) where the value of one or more of the covariates and/or response variables are missing. Missing values can occur both at the time of training and

testing of a model.

Should the training dataset contain missing values, most traditional statistical methods such as regressions would use *casewise deletion*, that is, to remove the entire instance (observation) from the dataset if the value of one or more variable is missing (Harrell, 2015). Casewise deletions can lead to bias in the estimated parameters if the degree to which the variable's value is likely to be missing is correlated with the actual range of values, *e.g.*, when a temperature sensor fails to record values below $-10°C$. Casewise deletions also result in losing the information provided by the remaining variables in the instance with missing values. Therefore, *imputation* is often used to estimate the missing values, which can be as simple as using the variable's mean or the variable's value from a similar instance, or can be more complex, such as using the *chained equation method* (Harrell, 2015). However, in essence, imputation is presuming a model for the variables with missing values, which may conflict the actual model that is going to be trained on the imputed dataset, resulting in a poor predictor. As with BNs, methods such as *expectation maximization (EM)* and *structural EM* can be used to learn the parameters and structure, without imputation or casewise deletion (Koller & Friedman, 2009).

Should the testing dataset contain missing values, almost all models fail to make predictions as each covariate has to take some value, *i.e.*, they cannot be left with "NA"s (not available). Imputation comes with the above mentioned shortcomings. Another alternative is to use expert knowledge to obtain probable limits for the covariates with missing values, and run the model on those limits to get a probable range for the prediction. For example, in climate change models, the exact concentration of the pathway of a covariate such as greenhouse gas emission that will be followed in the future is unknown. Therefore, models use a series of scenarios ranging from best to worst case scenario in order to predict changes in CO_2 emissions and temperatures (Pachauri *et al.*, 2015). There is, however, no need of these rough approximations when applying BNs. By marginalizing over the unobserved covariates, BNs can predict the target variable based on any observed subset of the covariates.

2.1.4 *Nonlinearity of the relationship between the covariates and response variable*

In many real-world situations, the response variable may be related to the covariates in a highly nonlinear manner. Simple models such as linear regressions, however, assume a linear relationship. To capture some levels of non-linearity, generalized linear models extend the regressions by applying functions such as $\log(\cdot)$ and $\text{logit}(\cdot)$ to the covariates. Other extensions, such as generalized additive models, fit a smooth curve to the data for each covariate, thereby allowing complex nonlinear relationships (Guisan *et al.*, 2002). Another extension is the machine-learning method MaxEnt (Phillips *et al.*, 2006) that is able to link highly non-linear response curves and estimate the probability distribution of the response variable using maximum entropy. Likewise, support vector machines classify the covariate space using hyper-planes, and hence, are linear, yet can allow for some nonlinearity by first transforming the space using nonlinear kernels (Scholkopf & Smola, 2001). Process-based models can also build in highly complex nonlinear relationships. In all of these cases, the relationships between the response variable and covariates must be entirely described, based on *a priori* model, a constraint that is relaxed in some other machine learning models. For example, classification trees can represent any function over the set of discrete covariates, but does not need to be defined beforehand. Note, this may require a very deep classification tree. Moreover, the fact that a classification tree can represent a complex function does not mean it can be learned effectively. Likewise, BNs are flexible in dealing with nonlinear relationships. Over a set of discrete variables, BNs can represent an arbitrary joint probability distribution $P(Y, \mathcal{X})$, which can represent any arbitrary conditional distribution $P(Y|\mathcal{X})$.

2.1.5 *Hypothesis testing, statistical inference and model selection*

The objective of hypothesis testing is to make inference through deduction. It consists of devising one or more working hypotheses and challenging them with data for corroboration (Hilborn & Mangel, 1997; Stephens *et al.*, 2005). The hypothesis to test is translated into a mathematical equation and is verified using methods such as least squares and maximum likelihood. So to test

a hypothesis, one needs *(i)* a mathematical equation representing a biological hypothesis and *(ii)* a test statistic with a distribution that can be determined, representing the model accuracy when confronted to data. The complexity of machine-learning models usually prevents us from obtaining a simple equation representing the hypothesis, but this is not the case for BNs.

For example, consider a process with the response variable Y and covariates X_1 and X_2. One may hypothesize that the response variable Y depends on both X_1 and X_2 but becomes independent of X_2, given X_1. Namely, the response variable depends directly only on covariate X_1, and that X_1 itself depends only on X_2. This can be modelled by a BN with three nodes for the variables and two links: one from X_1 to Y and another from X_2 to X_1. The BN assigns the following likelihood to each observation of the above process:

$$P(Y, X_1, X_2) = P(Y \mid X_1)P(X_1 \mid X_2)P(X_2). \tag{2}$$

The *null hypothesis* in this case is that there is no dependence among the variables: they are mutually independent. This results in a BN without any links between the nodes, yielding the following likelihood:

$$P(Y, X_1, X_2) = P(Y)P(X_1)P(X_2). \tag{3}$$

Given an observation, each of the probabilistic terms on the right hand-side of the above equations is simply a parameter provided that the BNs are discrete. Hence, the likelihood of a specified dataset for each of the BNs will be a polynomial in the parameters, the maximum of which is straightforward to derive. This allows for classical hypothesis testing, *e.g.*, by employing the *likelihood ratio test*, to reject the null hypothesis. Alternatively, among all BNs with the nodes Y, X_1 and X_2, one may find 'the best' using *multiple working hypotheses*, based on the *Akaike information criterion* (AIC; Akaike, 1974) or *Bayesian information criterion* (BIC; Schwarz, 1978). Therefore, with BNs, we are able to make inferences and obtain insights on the ecologically relevant covariates (*e.g.* Cooper & Herskovits, 1992; Pollino *et al.*, 2007a; Milns *et al.*, 2010).

2.1.6 Prior knowledge of the processes

Unlike mechanistic models that typically need a comprehensive knowledge of the involved processes to make accurate predictions, phenomenological methods such as traditional statistics and especially machine learning have more leeway. One does not need to have any knowledge about the ecological process to train and test a support vector machine, or neural network, for example. Although one may argue that the functions used in a neural network or the number of nodes and layers are parameters to be determined beforehand, yet these too can be selected automatically based on the training data or general rules of thumb. The level of autonomous learning is even higher with BNs. The whole structure and parameters of a discrete BN can be completely learned from data (McCann *et al.*, 2006a). The same goes for decision trees.

Although they can be trained autonomously, BNs allow experts to incorporate their knowledge into the network by forcing or preventing links between the nodes and additionally adding *latent variables* that are unobservable and often abstract variables, such as habitat quality. Indeed, the spectrum of autonomous learning for BNs ranges from neither to both structure and parameters learned based on experts' knowledge.

2.1.7 Correlations

Often two or more of the covariates in a process are highly correlated. This hinders statistical inference as the effects of the correlated covariates on the response variable are difficult to separate (Stewart, 1987; Dormann *et al.*, 2013). This would happen if we were building a model, say a logistic regression, with two covariates that are both relevant to the response variable, and also are highly correlated with each other. Thus, typically one of the variables is eliminated beforehand, either randomly, based on ecological relevance, measurement feasibility, and proximity to the mechanisms (Harrell, 2015; Dormann *et al.*, 2013), or by using some autonomous technique such as *minimum-redundancy maximum-relevance* (Peng *et al.*, 2005). However, this prevents understanding the impact of both of the correlated covariates together on the response variable.

Process-based models do not suffer from correlation (except for parameter estimability), yet they require the mechanisms to be *a priori* known (Dormann *et al.*, 2013). Nevertheless, a BN whose structure is learned from data, does not require any prior knowledge, and reveals the differences of the correlated covariates in terms of their probabilistic dependence to other covariates as well as the response variable.

2.1.8 Predictive accuracy

Despite the complexity of ecological systems (Levin, 1992; Anand *et al.*, 2010), some machine-learning models are reported to make accurate predictions. In contrast, process-based and traditional statistical models are rarely able to reach the same level of accuracy (*e.g.* Elith *et al.*, 2006). Particularly, process-based models are known for their inability to make good predictions, although this has been challenged by, for example, (Håkanson, 2004), who presented an accurate mechanistic model for aquatic systems. Within machine-learning models, neural networks are acknowledged for accurate performance in highly complex tasks such as image recognition (Egmont-Petersen *et al.*, 2002). However, this does not mean that neural networks necessarily outperform simpler models in practice. Firstly, finding the optimal number of layers and nodes is not always practical due to limited computational resources. Secondly, proper estimation of the many parameters of a neural network often requires massive data. Hence, while asymptotically effective, neural networks may not be as successful as simple models when the available data is insufficient. Finally, if the system in question is actually simple, then neural networks, especially deep ones, can easily overfit the training data. Simpler models may be a better choice also in this case.

2.1.9 Marginalizability

The notion of marginalizability addresses the possibility of separately studying how a particular covariate or subset of the covariates informs us about the response variable. We call a model

marginalizable if it allows us to compute the probability of the response variable Y given any subset $\mathcal{Z}$ of the covariates $\mathcal{X}$; that is, $P(Y \mid \mathcal{Z})$. Most predictive models allow us to compute $P(Y \mid \mathcal{X})$, that is, the likelihood of the response given all of the covariates. However, only those that perform a generative task, *i.e.*, learning $P(Y, \mathcal{X})$, allow us to marginalize the likelihood over the variables $\mathcal{X} - \mathcal{Z}$, to obtain the likelihood conditioned on only those variables that we are interested in: $P(Y \mid \mathcal{Z})$. Therefore, only BNs and those mechanistic models developed to formulate the joint probability $P(Y, \mathcal{X})$, are marginalizable.

2.2 Learning Bayesian networks from data

We explain, step by step, how to learn and then use a BN to make predictions and acquire biological insights. Most steps are general enough to be applied by any statistical/machine-learning method in the context of model selection or prediction making.

2.2.1 Setup

Ecological processes are typically modelled by a response variable Y and a set of covariates $\mathcal{X}$. If the process is spatial and temporal, then each instance (observation) of the process has a unique pair of identities: *(i)* the time t of the instance, the unit of which indicates the frequency of the observations, *e.g.*, a year, month, or day, and *(ii)* a general index g, roughly to distinguish the instances location-wise. For example, if the process of interest is Cyanobacteria bloom in lakes, then g indicates the label of the lakes. If the interest is in the spread of an infestation over a given area, then we may divide the area into $r \times r$ squares for say $r = 1$km, and label them by $g = 1, 2, \ldots$. We may exclude time when modelling a stationary quantity, *e.g.*, the joint distribution of several species in a specific area. Similarly, we may exclude the index g, if all instances are taken from the same location, *e.g.*, from the same lake. Also, note that time and especially the index g are not necessarily two covariates of the process. Indeed, time must be excluded from the set of covariates if the goal is to obtain a model that can be applied to

times different from those in the available data, *e.g.*, to predict the future (see SI). Similarly, the index g may be excluded; however, one must acknowledge the possible performance loss when applying the model to areas far-away from the training area, with dramatically different geographic features.

For illustration purposes, in what follows, we consider a spatial and temporal process. For each index g and time t, let $\mathcal{X}_{g,t}$ denote the set of covariates and $Y_{g,t}$ denote the response variable (Table 2). Although the response variable can be continuous or integer, in order to use acknowledged performance measures such as AUC (Section 2.2.5), we restrict it to be binary. For example, given an index g and time t, the response variable $Y_{g,t}$ may represent the presence, $Y_{g,t} = 1$, or absence, $Y_{g,t} = 0$, of infestation or a species of interest. The covariates can be correlated with each other and may include variables that are not known *a priori* to contribute to the response variable. Our goal is to estimate (learn) the joint probability distribution $P(Y_{g,t}, \mathcal{X}_{g,t})$ using available data.

Table 2 Variable notation.

Notation	Variable
t	time
g	general index
$Y_{g,t}$	response variable
$\mathcal{X}_{g,t}$	set of covariates
$P(\cdot)$	probability function

2.2.2 Step 1: Data discretization

The random variables in a BN can be either continuous or categorical. However, if they are continuous, we must predetermine their distributional forms, *e.g.*, a Gaussian distribution. To avoid making such assumptions, we use discrete BNs where every variable is categorical. We discretize all continuous variables by considering various number of intervals or discretization levels (say $2, 3, \ldots, 10$) and using data to determine which number leads to a higher performance score. If a continuous variable's range does not have evident thresholds in terms of the biological

context, we use *Hartemink's information-preserving algorithm* (Hartemink, 2001) to quantify the values in a way that maximizes the mutual information shared by the variables (Cover & Thomas, 2012).

2.2.3 Step 2: Partitioning the dataset into train and test

The typical machine-learning approach to learn, then evaluate a model, is to randomly partition the dataset in two subsets, *train* and *test*, where the greater portion (train) is used to estimate the model, and the remaining portion (test) to evaluate the trained model. However, evaluation concerns are raised if the instances of the original data are randomly partitioned into train and test. Indeed, using this method, the train and the test datasets are extremely similar (see SI). For each instance of the test dataset, it is highly likely to have a matching instance in the train dataset due to correlations in time and space. The purpose of a test dataset is to simulate how the model performs when applied in practice to a new dataset. If the goal is to make predictions in the future, say next month, we set the train dataset to be the data from the final observations (instances) and let the train dataset be the remaining instances. Namely, we make the train and test datasets time-wise disjoint.

2.2.4 Step 3: Learning

Step 3.1. Learning the BN structure. For each of the k-level quantified training datasets, we find the structure that results in the lowest BIC or the lowest AIC. Although this can be done by performing an exhaustive search on all possible BN structures – *i.e.*, directed acyclic graphs, with the response variable and covariates as the node-set – we instead use efficient algorithms, *e.g.*, (Silander & Myllymaki, 2012), which is implemented in the `R` package `bnstruct` (Franzin *et al.*, 2017). Both BIC and AIC criteria penalize having more parameters, which reduces the chance of over-fitting to the training dataset. The choice of BIC or AIC depends on the main goal of the study, the model complexity, and the number of instances relative to the number of

parameters (Aho *et al.*, 2014).

Note this approach is computationally infeasible if there are too many variables, *e.g.*, more than 25, or too many discretization levels. Then one may, instead, either use a fixed (*a priori* known) BN structure, *e.g.*, naive Bayes, or learn a 'close-to-optimal' (*a priori* unknown) BN on the training dataset using acknowledged searching-algorithms (Table 3, Fig. 1). We learn the structure of the *a priori* unknown networks by the `bnlearn` package in `R` (Scutari, 2009). The input to each algorithm is the variables and the corresponding training dataset, and the output is a BN structure whose nodes are the variables. In case the learned structure contains undirected links, we randomly assign directions as long as directed cycles and v-structures do not appear. This is because BNs must not contain cycles by definition, and the introduction of v-structures can change the performance of the resulting BN (Koller & Friedman, 2009). So for each discretization level k, we obtain a BN structure according to one of the algorithms or fixed structures in Table 3.

Step 3.2. Learning the BN parameters. After finding the highest-scoring BN structure for each of the k-level quantified training datasets, we learn the associated CPD parameters on the same training dataset and denote the resulting BN by $\mathcal{B}_k^*$. We use the *Bayesian parameter estimation* approach (Koller & Friedman, 2009), implemented in `bnlearn`. To this end, for each quantization level k, we obtain a BN $\mathcal{B}_k^*$ that best fits the training data in terms of BIC, AIC, or other constraints listed in Table 3.

2.2.5 Step 4: Evaluation

How to choose among the different $\mathcal{B}_k^*$s from the previous step? Namely, what number of discretization levels results in 'the best' BN? We cannot compare them directly using a performance measure that involves the likelihood of the data, *e.g.*, log-likelihood, AIC and BIC, because the $\mathcal{B}_k^*$s do not use the same data but different discretized versions of it.

However, all BNs use the same number of discretization levels for the response variable. So

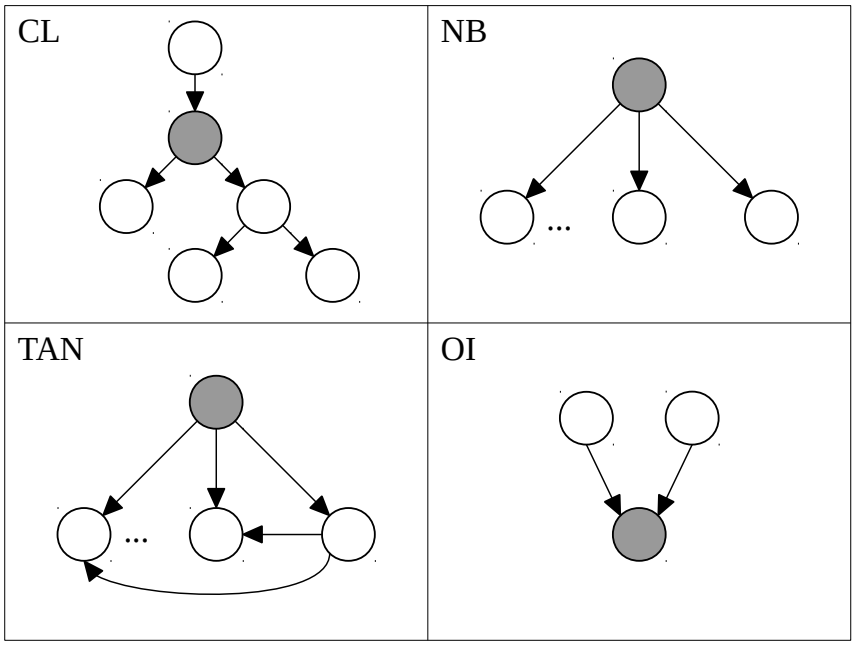

Fig. 1 Structure of different Bayesian networks CL: Chow-Liu, NB: naive Bayes, TAN: tree-augmented naive Bayes, and OI: one-memory infestation (Section 3). Grey and white circles represent the target $I_{g,t}$ and its covariates. In the OI case, the covariates are $I_{g,t-1}^{\text{Missed}}$ and $I_{g,t-1}^{\text{Managed}}$.

we can compare them based on how well they predict the response variable on the test dataset. Each network allows us to compute $P(Y_{g,t} \mid \mathcal{X}_{g,t})$, that is, the chances of the observed response variable given the covariates, for every instance in the dataset. Correspondingly, we compare the *area under receiver operating characteristic curve (AUROC* or simply *AUC))* (Metz, 1978; Bradley, 1997) score of the BNs on the test dataset (see SI). The choice of AUC is to make our results comparable with the huge body of literature using this performance score as the final performance of a classifier. For each discretization level k, we calculate the AUC score of $\mathcal{B}_k^*$ and pick the highest-scoring one as our final BN. If there is a tie between the top BNs, we break it by looking at the *area under precision-recall curve (AUPR)* (Raghavan *et al.*, 1989; Saito & Rehmsmeier, 2015) scores; that is, among the top BNs with a deficit of at most, say 0.01, from the top AUC, we pick the one with the highest AUPR. The AUPR score better handles unbalanced data by looking at precision rather than the false positive rate (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Given the temporal nature of our task, we evaluate the final model on a single test dataset,

Table 3 Bayesian networks to compare with the learned one.

Abbreviated name	Based on the algorithm/ structure	Type of the algorithm	Description
GS	Grow shrink	Constraint based	Uses conditional independence tests on the training dataset to detect the Markov blankets of the variables (Margaritis & Thrun, 1999)
IAMB	Incremental association Markov blanket	Constraint based	Detects Markov blankets with an attempt to avoid *false positives*, *i.e.*, fault infestation predictions (Tsamardinos *et al.*, 2003)
IIAMB	Interleaved incremental association Markov blanket	Constraint based	A variant of IAMB to maintain the size of the Markov blanket as small as possible (Tsamardinos *et al.*, 2003)
HC	Hill climbing	Local search	Starts from a random directed graph and adds or removes an edge only if it results in a higher score (BIC in our case) on the train dataset (Margaritis, 2003)
CL	Chow-Liu	Global search	Finds the undirected spanning tree of the variables to minimize the *Kullback-Leibler distance* from the actual distribution (Chow & Liu, 1968) (Fig. 1)
NB	Naive Bayes	-	The most basic yet often successful BN formed by the response variable ($Y_{g,t}$ in our case), linking to all of the covariates (Koller & Friedman, 2009) (Fig. 1)
TAN	Tree-augmented naive Bayes	-	A NB network with a spanning tree among the covariates that can be learned from the train dataset (Friedman *et al.*, 1997) (Fig. 1)

as explained in Section 2.2.3. If instead, one divides the original dataset into several yearly-separated folds and uses cross-validation to obtain the AUC and AUPR values for each fold, then one could also provide confidence intervals for the reported AUC and AUPR values.

2.2.6 Step 5: Interpretation

Given an index g and year t, the final BN obtained from the above steps, determines the joint probability distribution of the response variable $Y_{g,t}$ and covariates $\mathcal{X}_{g,t}$. Perhaps the most important implication of the obtained BN is the primary ($\mathcal{X}^1_{g,t}$) and secondary ($\mathcal{X}^2_{g,t}$) division of the covariates $\mathcal{X}_{g,t}$ with respect to the response variable. Namely, if we just know the primary

covariates, there is no need to know the secondary covariates, *i.e.*,

$$P(Y_{g,t} \,|\, \mathcal{X}^1_{g,t}, \mathcal{X}^2_{g,t}) = P(Y_{g,t} \,|\, \mathcal{X}^1_{g,t}). \tag{4}$$

Moreover, other conditional independencies between the covariates themselves can be identified based on the *d-separations* of the BN (Koller & Friedman, 2009).

Also, based on the CPDs, we can investigate knowing which covariates increases the probability of the response variable most. For example, consider the covariate temperature $T_{g,t}$, discretized into the two ranges $[20°\text{C}, 30°\text{C})$ and $[30°\text{C}, 40°\text{C})$. We can see how the response variable $Y_{g,t}$ depends on this covariate by sweeping through these quantified levels, *e.g.*,

$$P\left(Y_{g,t} = 1 \,|\, T_{g,t} \in [20°\text{C}, 30°\text{C})\right) = 0.3 \qquad \text{and} \qquad P\left(Y_{g,t} = 1 \,|\, T_{g,t} \in [30°\text{C}, 40°\text{C})\right) = 0.4. \tag{5}$$

Hence, the response variable being equal to 1 is most likely when temperature is in the range $[30°\text{C}, 40°\text{C})$. Note that this is only if other covariates are unknown. Now, comparing this with the similar probability conditioned on a different covariate, clarifies which is more informative to the response variable.

2.2.7 Step 6 (optional): Sensitivity analysis

We examine the prediction accuracy (AUC and AUPR) of the best model when a primary covariate becomes unobservable. This roughly shows the contribution of each covariate to the prediction, although it is, indeed, the co-effect of all the covariates that leads to accurate predictions.

2.2.8 Step 7 (optional): Comparison with simple Bayesian networks

To further assess the prediction performance of the final BN, we may compare its AUC (or AUPR) with that of simple BNs consisting of a single or two covariates linked to the response

variable. These BNs might be considered as the 'null model'.

Recall that our final BN is designed to perform a generative task, that is to reveal the relationships between the variables, not a discriminative task, that is to predict the response variable. However, if the BN performs well on the first, it is likely to also do well on the second. Yet, the opposite does not hold (Ng & Jordan, 2002). So even if any of these simple BNs predicts the response variable better than our final BN, it does not question the capability of our BN in explaining the probabilistic relationships between the variables. The same may hold in the previous optional step: the AUC score of the BN may increase after removing some of the covariates. This can also be explained by the fact that our final BN is the best fit to the data under the performance score that we used, which is BIC (or AIC) not AUC.

Nevertheless, in such cases, we may train a BN with a different set of covariates for prediction purposes. For example, we may find that subset of the covariates that results in a BN scoring the highest AUC on the training dataset.

3 The mountain pine beetle case study

We illustrate the learning and interpretation of BNs via the data on the MPB infestation in the Cypress Hills park–an interprovincial park located in Alberta and Saskatchewan (Fig. 2 in SI). Endemic-level populations of MPB have existed in Cypress Hills since the 1980's. However, a MPB outbreak started in 2006 and propagated in the park, where it continues until now.

3.1 Biology and management

MPB presents two main population phases: an *endemic phase* with small population size where beetles attack weak and stressed pines with the help of other bark beetles, and an *epidemic phase* where the number of individuals is large enough to overcome the defences of large and healthy pines (Safranyik & Carroll, 2006). In summer, beetles will emerge from a tree, mate, and attack new pines to lay eggs in galleries under the bark. New MPB infestations are reported

to frequently appear in south and west-facing slopes (Safranyik, 2004). During the tree growing season, water-stress negatively impacts the pine's ability to build its defense against bark beetles (Safranyik, 1978; Lusebrink *et al.*, 2016). Indeed, pines use water to make a toxic resin that is exuded during a beetle attack to prevent beetles from attracting conspecifics and inhibit the formation of galleries and oviposition (Raffa & Berryman, 1983; Erbilgin *et al.*, 2017). MPB emergence and flights are reduced with high temperatures during the dispersal season (Safranyik & Carroll, 2006). MPB can disperse at short distances within a stand or, more rarely, fly above the canopy to use the wind to travel long distances of the order of tens to thousands of kilometers (Safranyik & Carroll, 2006; Robertson *et al.*, 2007). Once the eggs are laid, the adults die. Over the fall, winter, and spring, eggs become larvae then pupae before finishing their transition to adult and emerging in the summer. Individuals need a minimum of 833 degree days to complete their transition to adult (Safranyik *et al.*, 1975, 2010).

The Forest Service Branch of the Saskatchewan Ministry of Environment follows a strict direct control approach. At the start of every fall, the park is surveyed aerially to collect geo-referenced data on red-top trees – *i.e.*, trees that are dead or dying from a MPB infestation at the previous year. Then, on the ground, managers survey 50 meter-radius circular plots around each red-top tree to find recently infested trees during the summer. The newly-found infestations are later controlled in late fall/winter using a fell and burn method.

Our goal is to provide a set of covariates that potentially impact the MBP infestation in Cypress Hills area, understand how they are related to each other and to the infestation, and find which covariates are sufficient for an accurate prediction. We also are interested to test some of the claims in the literature, *e.g.*, lower humidity increases the chances of infestation (Lusebrink *et al.*, 2016), and to find what values of the highly correlated covariates degree-days and maximum temperature, that are typically not included together in a model, makes infestation most likely. These objectives are well suited to BNs.

3.2 Methods

We divide the studying area into 100m × 100m squares and label them by $g = 1, 2, \ldots$. We choose one year as our time unite and define the response variable $I_{g,t}$ as the presence or absence of infestation in pixel g at the fall of year t. We use the covariates listed in Table 4 and quantify them into $k = 1, 2, \ldots, 7$ levels. Our data includes the values of $\mathcal{X}_{g,t}$ and $I_{g,t}$ over the years $t = 2006, 2007, \ldots, 2018$ and for 18 317 different pixels g in Cypress Hills, resulting in a total of 238 121 instances (see SI for an instance of the data).

We compare the AUC and AUPR scores of our final model with those of what we call the *one-memory infestation (OI)* Bayesian network, consisting of $I^{\text{Managed}}_{g,t-1}$ and $I^{\text{Missed}}_{g,t-1}$, being linked to the target $I_{g,t}$, considered as the null model (Fig. 1).

3.3 Resulting Bayesian network

We find the BN with the best BIC score on the train dataset with 6 discrete levels, *i.e.*, $\mathcal{B}^*_6$ (Fig. 3), as our "best model" to explain the MPB infestation, with AUC = 0.88 and AUPR = 0.28. The OI model scores 0.75 for AUC and 0.19 for AUPR – both lower than our selected model. According to the structure of $\mathcal{B}^*_6$, the infestation $I_{g,t}$ in location g at year t is directly connected to $I^{\text{Missed}}_{g,t-1}$, $I^{\text{Managed}}_{g,t-1}$, $I^{\text{Missed}}_{\mathcal{N}_g,t-1}$, $I^{\text{Managed}}_{\mathcal{N}_g,t-1}$, B_g, and $T^{\max}_{g,t-1}$. These together with $C_{g,t-1}$, form the Markov blanket of the infestation node, and hence, are the primary covariates and sufficient for estimating infestation with 0.88 AUC score. Other covariates are all indirectly linked to infestation and are secondary covariates. Given $\mathcal{B}^*_6$, one can obtain conditional independencies of the covariates to infestation using d-separations and plot the CPDs (see SI).

3.4 Sensitivity to missing covariates

The prediction accuracy of $\mathcal{B}^*_6$ does not deteriorate when the values of any of the secondary covariates are missing. Upon missing values for the primaries, the model can still accurately predict infestation as it can use some of the secondary covariates (Table 5).

Table 4 Description of the covariates $\mathcal{X}_{g,t}$.

Name	Symbol	Description	Unit
Aspect	A_g	Compass direction that the slope at pixel g faces	$\circ$
Distance to infested border	B_g	Distance of the centre of pixel g to the border of the whole area of interest that was initially infested (Fig. 2 in SI)	km
Degree days	$D_{g,t-1}$	Sum of daily temperatures above 5.5°C from fall of year $t-1$ to summer of year t	Celsius degree-day
Maximum temperature	$T^{\max}_{g,t-1}$	Highest maximum daily temperature in July and August of year t	°C
Wind speed	$W_{g,t-1}$	Average daily wind speed in July and August of year t	km/h
Relative humidity	$R_{g,t-1}$	Average daily relative humidity in spring of year t	%
Cold tolerance	$C_{g,t-1}$	An index in $[0,1]$ representing the ability of the larvae to survive the cold season of year $t-1$, as defined in (Régnière & Bentz, 2007)	
Pine cover	$P_{g,t-1}$	Pine density in summer of year t	%
Managed last year infestation	$I^{\text{Managed}}_{g,t-1}$	Defined to be 1 if pixel g includes at least one tree that was infested and managed (controlled) at year $t-1$, and 0 otherwise (Fig. 2)	–
Missed last year infestation	$I^{\text{Missed}}_{g,t-1}$	Defined to be 1 if pixel g includes at least one tree that was infested and missed (not controlled) at year $t-1$, and 0 otherwise	–
Missed neighbors' last year infestation	$I^{\text{Missed}}_{\mathcal{N}_g,t-1}$	MPB's ability to disperse at short distances within a stand, defined as $I^{\text{Missed}}_{\mathcal{N}_g,t-1} = \sum_{i=1}^{3} \frac{1}{2^i} \sum_{g' \in \mathcal{N}_g^i} I^{\text{Missed}}_{g',t-1}, \quad I^{\text{Missed}}_{\mathcal{N}_g,t-1} \in [0,6],$ where $\mathcal{N}_g^i$ are those pixels that are essentially at a distance of $i \times 100$m from g (Fig. 3 in SI).	–
Managed neighbors' last year infestation	$I^{\text{Managed}}_{\mathcal{N}_g,t-1}$	Defined similarly to $I^{\text{Missed}}_{\mathcal{N}_g,t-1}$, with the difference that $I^{\text{Missed}}_{g',t-1}$ is replaced by $I^{\text{Managed}}_{g',t-1}$	–

3.5 Discussion

The final model we have chosen to explain the MPB infestation in the Cypress Hills area is the BN $\mathcal{B}_6^*$ with 6 discretization levels, scoring 0.88 AUC on the test dataset. For a managed MPB outbreak in the Cypress Hills area, the model postulates the following covariates as primary (and hence sufficient for an 0.88 AUC prediction) at each location, at each time: *(1,2)* presence of infestation in last year, both managed and missed, *(3,4)* neighbors' degree of infestation in last

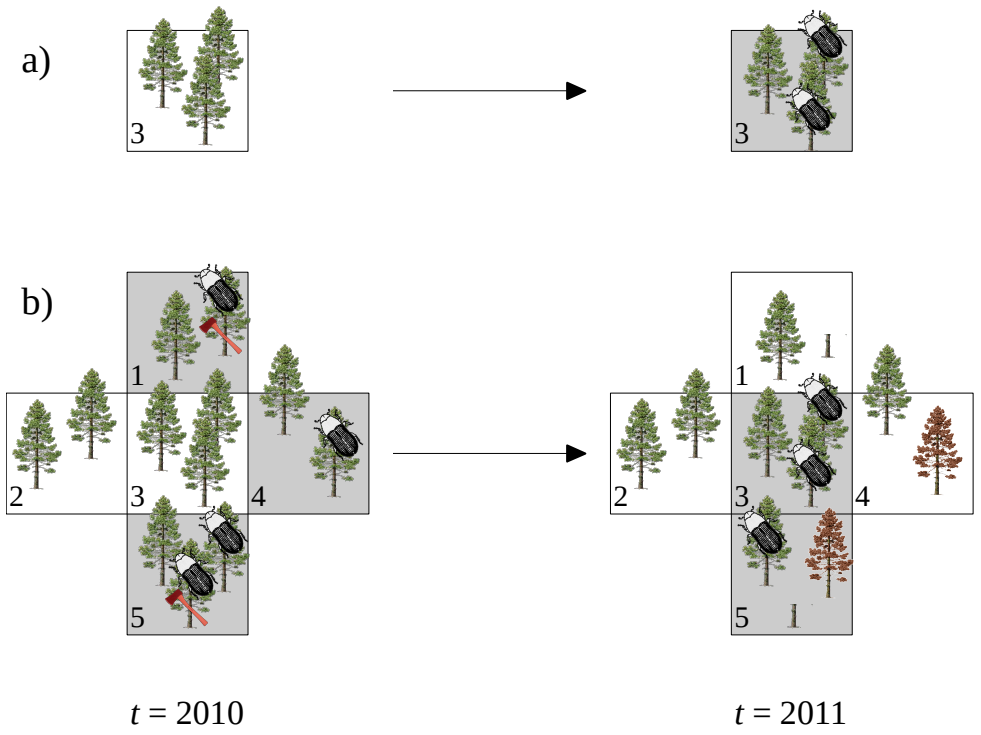

Fig. 2 Infestation status. Gray and white are used to indicate the presence and absence of infestation in a pixel. **(a)** None of the trees in pixel 3 were infested at year 2010 ($I_{3,2010} = 0$); however, at least one tree was infested at year 2011 ($I_{3,2011} = 1$). **(b)** All infested trees in pixel 1 that were infested at year 2010 were managed at the same year ($I^{\text{Managed}}_{1,2010} = 1$, $I^{\text{Missed}}_{1,2010} = 0$), there were no infested trees in pixel 2 at year 2010 ($I^{\text{Managed}}_{2,2010} = 0$, $I^{\text{Missed}}_{2,2010} = 0$), all infested trees in pixel 4 that were infested at year 2010, were missed at the same year, and hence, turned red in the following year ($I^{\text{Managed}}_{4,2010} = 0$, $I^{\text{Missed}}_{4,2010} = 1$), some infested trees were missed and some were managed in pixel 5 ($I^{\text{Managed}}_{5,2010} = 1$, $I^{\text{Missed}}_{5,2010} = 1$). Missed and managed neighbors' last year infestation for pixel 3 at year 2010 are, thus, $I^{\text{Missed}}_{\mathcal{N}_3,2010} = I^{\text{Managed}}_{\mathcal{N}_3,2010} = \frac{1}{2} + \frac{1}{2}$, presuming that $\mathcal{N}^2_3 = \mathcal{N}^3_3 = \emptyset$.

Table 5 AUC and AUPR scores of 'the best' BN $\mathcal{B}^*_6$, when one of the primary covariates is missing.

Missing covariate	AUC	AUPR
Nothing missing	0.882	0.277
Maximum temperature	0.889	0.350
Cold tolerance	0.881	0.290
Distance to infested border	0.890	0.309
Missed neighbors' past infestation	0.760	0.220
Managed neighbors' Past infestation	0.879	0.284
Missed last year infestation	0.811	0.103
Managed last year infestation	0.869	0.206
Last year infestation (both missed and managed)	0.784	0.068

year, both managed and missed, *(5)* distance to the border where the infestation was initiated, *(6)* maximum temperature in July and August of that year, and *(7)* cold tolerance in the cold season

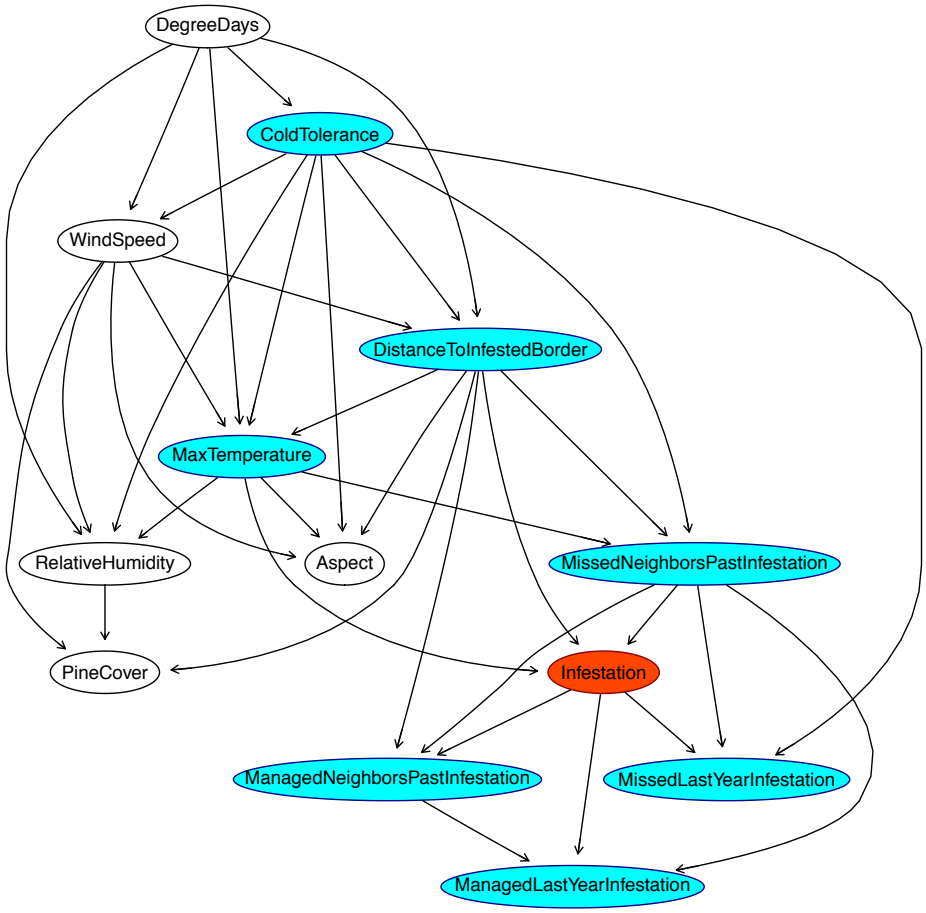

Fig. 3 The structure of 'the Best' Bayesian network ($\mathcal{B}_6^*$). We choose this structure as the one to explain and predict MPB infestation. The response variable and its Markov blanket are in red and Cyan.

of that year; *n.b.*, the remaining covariates are secondary and are used to predict infestation if one or more of the primary covariates are missing.

Given this BN, we can provide a wide range of *ceteris paribus* claims revealing the co-effects of the covariates on the presence of infestation (see SI). For example, if we know maximum daily temperature is high (above 31.2°C), the interval of relative humidity that results in the highest infestation risk sharply changes from medium to low. This is in line with the claim in (Safranyik, 1978; Lusebrink *et al.*, 2016) that lower humidity increases the infestation probability. However, for maximum daily temperatures lower than 31.2°C, the infestation likelihood is high for both low and high relative humidity. This inconsistency can be solved by looking at maximum temperature and relative humidity together. We find that humid areas require low maximum

daily temperature, while dry areas require high maximum daily temperature for a considerable risk of infestation (above 20%).

As another example, a MPB needs 833 degree days to complete its transition to adults and the minimum number of degree-days in the data is 1054 (Safranyik *et al.*, 1975, 2010). Therefore, degree-day never prevents infestation in our data and just reflects the negative impact of high summer temperatures. This, however, does not mean that degree day is useless in our model. First of all, as mentioned earlier, in the absence of some of the primary covariates, the model effectively estimates infestation via the information on degree day and other present covariates. Secondly, although highly correlated, degree day and maximum temperature are different, and the model reveals their coexistence effect on the infestation: for low (resp. high) degree-days, infestation becomes more likely as maximum temperature increases (resp. decreases; see SI).

We emphasize that one may not make causal conclusions based on the structure of the model. Clearly, the edge from infestation to managed-last-year-infestation does not imply that this year's infestation has caused last year's (managed) infestation. It only means that the two are probabilistically dependent. The same holds for all other links, such as the one from maximum temperature to infestation: although temperature may be "causing" infestation, one may not conclude so just based on the BN. One may refer to the literature on causality and the corresponding tests in order to verify the causality of a link in a BN (Pearl, 2009; Pearl & Mackenzie, 2018). On the other hand, the absence of an edge between, for example, degree day and infestation does not necessarily mean that the two are independent. They may be dependent but become conditionally independent if some other covariates are known here.

In summary, the learned BN contributes to the prediction and understanding of MPB infestations by *(1)* accurately predicting MPB infestations, *(2)* identifying the primary set of covariates that are sufficient for making these predictions, *(3)* making acceptable predictions when data on some of the primary covariates are unavailable, *(4)* revealing the previously unknown co-effects of the covariates on infestation likelihood, *(5)* identifying the most informative covariate(s) to

infestation likelihood, and *(6)* proposing a BN structure that can serve as the basis for future causality tests between the variables. Points 1, 2, 3, and 5 are particularly useful to forest managers to plan ahead of time and know what data to collect. See SI for a more elaborate discussion on the MPB case study.

Nevertheless, as with almost all other machine-learning models, BNs are generally constructed under the stationary assumption, implying fixed structure and parameters over time. This may result in poor performance when the model is used to make predictions at a time different from those in the training dataset, provided that the "true ecological process" is non-stationary. For example, a BN trained on data collected during the beginning of an outbreak may not accurately predict the declining phases of the outbreak. Similar concerns are raised when using the learned BN in environmental situations, where the ranges of the covariates are very different from those in the training dataset. We refer the reader to (Zhu & Wang, 2015; Zhou *et al.*, 2008; Robinson *et al.*, 2010) for relaxing the stationary assumption.

4 Discussion

Although traditional models used to make ecological predictions from underlying covariates have a record of success, they also suffer from limitations. They cannot make predictions when one or more covariates are missing; unless the missing values are imputed using other methods which can be unreliable and result in low prediction accuracy. They also do not allow for statistical inference when some of the covariates are highly correlated. BNs can handle these issues. Specifically, they provide a primary and secondary ordering of the covariates, where primary covariates are essential to predicting the target variable and secondary covariates, while not always essential, can be helpful in making predictions when the values of some covariates are missing.

However, BNs are not used to their full potential in the literature as their structure is typically constructed based on the knowledge of experts. Moreover, the obtained BN is often read causally, a questionable practice as BNs are different from causal networks.

We have complemented previous work by providing a systematic approach to obtain a BN fully from data. We have demonstrated the approach via a MPB case study, where no knowledge of experts was involved in finding either the structure or CPDs. The resulting BN predicts infestations fairly accurately, even in the absence of any of the selected covariates that are involved in the model.

Researchers have utilized BNs to visualize their understanding of the causal relationships between the variables involved in ecological processes (Borsuk *et al.*, 2004; Amstrup *et al.*, 2008; Johnson *et al.*, 2010; Newton, 2010; Aps *et al.*, 2009; Pollino *et al.*, 2007b). The resulting networks have been often used as predictors and sometimes reported to be fairly successful on a test dataset. This is an acceptable approach to assess the *a priori* knowledge of the experts or when there is no data available to learn the BN structure. However, by means of the results for our MPB case study, we challenge claims that put forward this approach as "the (only) right one" for constructing a BN. Examples include synthesizing existing knowledge into the model is necessary and structural learning is only for modeling poorly understood systems or those difficult to characterize (Chen & Pollino, 2012), modellers must demonstrate causal relations (McCann *et al.*, 2006b), models based on theories about causal relations are generally better (Uusitalo, 2007), and network structure is a matter of judgement and should reflect expert knowledge and stakeholder needs (Gutierrez *et al.*, 2011). Some researchers have looked into fixed (naive Bayes) and partially learnable (Tree-augmented naive Bayes) structures (Aguilera *et al.*, 2010), yet this is different from learning fully based on data.

In general, for modelling the joint probability distribution of the variables involved in an ecological process, *i.e.*, a generative task, BNs seem to be the first and often best candidate, especially if the governing dynamics are yet unknown to be mechanistically modeled. However, if the sole purpose is to predict the response variable, *i.e.*, a discriminative task, other models may show a higher prediction accuracy, although unlike BNs, they typically cannot deal with missing values in the covariates. We are currently exploring ways to use BNs as well as other

models, to predict infestation many years in the future (Ramazi *et al.*, accepted).

5 Acknowledgements

We thank Rory L. McIntosh for providing the mountain pine beetle data needed for the application section. Thank you to the Greiner and Lewis Research Groups for helpful feedback on ideas related to this research. The research was partly funded by Alberta Environment & Parks (AEP). This research was also supported by a grant to M.A.L. from the Natural Science and Engineering Research Council of Canada (grant no. NET GP 434810-12) to the TRIA Network, with contributions from Alberta Agriculture and Forestry, Foothills Research Institute, Manitoba Conservation and Water Stewardship, Natural Resources Canada-Canadian Forest Service, Northwest Territories Environment and Natural Resources, Ontario Ministry of Natural Resources and Forestry, Saskatchewan Ministry of Environment, West Fraser and Weyerhaeuser. M.A.L. is also grateful for the support through the NSERC Discovery and the Canada Research Chair programs. R.G. is grateful for funding from NSERC Discovery, and Alberta Machine Intelligence Institute.

6 Authors' contributions

All authors conceived the ideas, interpreted the results and drafted the manuscript. P.R. developed the methods and under-took the analysis. All authors gave final approval for publication.

7 Data Accessibility

The dataset analyzed in the current study is described in (Kunegel-Lion *et al.*, 2020a) and available from Dryad repository (https://doi.org/10.5061/dryad.70rxwdbt9) (Kunegel-Lion *et al.*, 2020b).

References

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

Aho, K., Derryberry, D. & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, **95**, 631–636. doi: 10.1890/13-1452.1.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723. doi: 10.1109/TAC.1974.1100705.

Amstrup, S.C., Marcot, B.G. & Douglas, D.C. (2008) A Bayesian network modeling approach to forecasting the 21st century worldwide status of polar bears. *Arctic sea ice decline: observations, projections, mechanisms, and implications*, **180**, 213–68.

Anand, M., Gonzalez, A., Guichard, F., Kolasa, J. & Parrott, L. (2010) Ecological systems as complex systems: Challenges for an emerging science. *Diversity*, **2**, 395–410. doi: 10.3390/d2030395.

Aps, R., Fetissov, M., Herkül, K., Kotta, J., Leiger, R., Mander, Ü. & Suursaar, Ü. (2009) Bayesian inference for predicting potential oil spill related ecological risk. *WIT Transactions on the Built Environment*, **108**, 149–159.

Aukema, B.H., Carroll, A.L., Zheng, Y., Zhu, J., Raffa, K.F., Dan Moore, R., Stahl, K. & Taylor, S.W. (2008) Movement of outbreak populations of mountain pine beetle: influences of spatiotemporal patterns and climate. *Ecography*, **31**, 348–358. doi: 10.1111/j.0906-7590.2007.05453.x.

Bode, M., Baker, C.M., Benshemesh, J., Burnard, T., Rumpff, L., Hauser, C.E., Lahoz-Monfort, J.J. & Wintle, B.A. (2017) Revealing beliefs: using ensemble ecosystem modelling to extrapolate expert beliefs to novel ecological scenarios. *Methods in Ecology and Evolution*, **8**, 1012–1021.

Borsuk, M.E., Stow, C.A. & Reckhow, K.H. (2004) A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modelling*, **173**, 219–239.

Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.

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

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

Cooper, G.F. & Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347. doi: 10.1007/BF00994110.

Cover, T.M. & Thomas, J.A. (2012) *Elements of information theory*. John Wiley & Sons.

Davis, J. & Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 233–240. ACM Press, Pittsburgh, Pennsylvania.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D. & Lautenbach, S. (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46. doi: 10.1111/j.1600-0587.2012.07348.x.

Egmont-Petersen, M., de Ridder, D. & Handels, H. (2002) Image processing with neural networks—a review. *Pattern Recognition*, **35**, 2279–2301. doi: 10.1016/S0031-3203(01)00178-9.

Eklöf, A., Tang, S. & Allesina, S. (2013) Secondary extinctions in food webs: a Bayesian network approach. *Methods in Ecology and Evolution*, **4**, 760–770.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151. doi: 10.1111/j.2006.0906-7590.04596.x.

Erbilgin, N., Cale, J.A., Hussain, A., Ishangulyyeva, G., Klutsch, J.G., Najar, A. & Zhao, S. (2017) Weathering the storm: how lodgepole pine trees survive mountain pine beetle outbreaks. *Oecologia*, **184**, 469–478.

Franzin, A., Sambo, F. & Di Camillo, B. (2017) bnstruct: An R package for Bayesian Network structure learning in the presence of missing data. *Bioinformatics*, **33**, 1250–1252.

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

Guisan, A., Edwards, T.C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100. doi: 10.1016/S0304-3800(02)00204-1.

Gutierrez, B.T., Plant, N.G. & Thieler, E.R. (2011) A Bayesian network to predict coastal vulnerability to sea level rise. *Journal of Geophysical Research: Earth Surface*, **116**.

Harrell, F.E.J. (2015) *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer Series in Statistics. Springer, second edition edn.

Hartemink, A.J. (2001) *Principled computational methods for the validation discovery of genetic regulatory networks*. Ph.D. thesis, Massachusetts Institute of Technology.

Hilborn, R. & Mangel, M. (1997) *The ecological detective: confronting models with data*. No. 28 Monographs in population biology. Princeton University Press, Princeton, NJ.

Håkanson, L. (2004) Break-through in predictive modelling opens new possibilities for aquatic ecology and management – a review. *Hydrobiologia*, **518**, 135–157. doi: 10.1023/B:HYDR.0000025064.26497.53.

Johnson, S., Fielding, F., Hamilton, G. & Mengersen, K. (2010) An integrated Bayesian network approach to Lyngbya majuscula bloom initiation. *Marine Environmental Research*, **69**, 27–37.

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

Kunegel-Lion, M., McIntosh, R.L. & Lewis, M.A. (2020a) Dataset of mountain pine beetle outbreak dynamics and direct control in Cypress Hills, SK. *Data in brief*, **29**, 105293.

Kunegel-Lion, M., McIntosh, R.L. & Lewis, M.A. (2020b) Dataset of mountain pine beetle outbreak dynamics and direct control in Cypress Hills, SK (dataset). *Dryad*. doi: 10.5061/dryad.70rxwdbt9.

Levin, S.A. (1992) The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, **73**, 1943–1967. doi: 10.2307/1941447.

Lusebrink, I., Erbilgin, N. & Evenden, M.L. (2016) The effect of water limitation on volatile emission, tree defense response, and brood success of *Dendroctonus ponderosae* in two pine hosts, lodgepole, and jack pine. *Frontiers in Ecology and Evolution*, **4**, 1–13. doi: 10.3389/fevo.2016.00002.

Marcot, B.G., Steventon, J.D., Sutherland, G.D. & McCann, R.K. (2006) Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, **36**, 3063–3074. doi: 10.1139/x06-135.

Margaritis, D. (2003) Learning Bayesian network model structure from data. , Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Margaritis, D. & Thrun, S. (1999) *Bayesian network induction via local networks*. School of Computer Science, Carnegie Mellon University.

Marmion, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009) The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, **220**, 3512–3520. doi: 10.1016/j.ecolmodel.2008.10.019.

McCann, R.K., Marcot, B.G. & Ellis, R. (2006a) Bayesian belief networks: Applications in ecology and natural resource management. *Canadian Journal of Forest Research*, **36**, 3053–3062. doi: 10.1139/x06-238.

McCann, R.K., Marcot, B.G. & Ellis, R. (2006b) Bayesian belief networks: applications in ecology and natural resource management. *Canadian Journal of Forest Research*, **36**, 3053–3062.

Merow, C., Smith, M.J. & Silander Jr, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Metz, C.E. (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 283–298. doi: 10.1016/S0001-2998(78)80014-2.

Milns, I., Beale, C.M. & Smith, V.A. (2010) Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, **91**, 1892–1899. doi: 10.1890/09-0731.1.

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G.J., Lavorel, S., Gall, L.L., Meslin, L., Morand, S., Morin, X., Morlon, H., Pinay, G., Pradel, R., Schurr, F.M., Thuiller, W. & Loreau, M. (2015) Predictive ecology in a changing world. *Journal of Applied Ecology*, **52**, 1293–1310. doi: 10.1111/1365-2664.12482.

Newton, A.C. (2010) Use of a Bayesian network for Red Listing under uncertainty. *Environmental Modelling & Software*, **25**, 15–23.

Ng, A.Y. & Jordan, M.I. (2002) On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. *Advances In Neural Information Processing Systems*, pp. 841–848.

Pachauri, R.K., Mayer, L. & on Climate Change, I.P., eds. (2015) *Climate change 2014: synthesis report*. Intergovernmental Panel on Climate Change, Geneva, Switzerland.

Pearl, J. (2009) *Causality*. Cambridge University Press.

Pearl, J. & Mackenzie, D. (2018) *The book of why: the new science of cause and effect*. Basic Books.

Peng, H., Long, F. & Ding, C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1226–1238.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259. doi: 10.1016/j.ecolmodel.2005.03.026.

Pollino, C.A., White, A.K. & Hart, B.T. (2007a) Examination of conflicts and improved strategies for the management of an endangered Eucalypt species using Bayesian networks. *Ecological Modelling*, **201**, 37–59. doi: 10.1016/j.ecolmodel.2006.07.032.

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

Preisler, H.K., Hicke, J.A., Ager, A.A. & Hayes, J.L. (2012) Climate and weather influences on spatial temporal patterns of mountain pine beetle populations in Washington and Oregon. *Ecology*, **93**, 2421–2434. doi: 10.1890/11-1412.1.

Purves, D., Scharlemann, J.P.W., Harfoot, M., Newbold, T., Tittensor, D.P., Hutton, J. & Emmott, S. (2013) Ecosystems: Time to model all life on Earth. *Nature*, **493**, 295–297. doi: 10.1038/493295a.

Raffa, K. & Berryman, A. (1983) The role of host plant resistance in the colonization behavior and ecology of bark beetles (Coleoptera: Scolytidae). *Ecological monographs*, **53**, 27–49.

Raghavan, V., Bollmann, P. & Jung, G.S. (1989) A Critical Investigation of Recall and Precision As Measures of Retrieval System Performance. *ACM Trans. Inf. Syst.*, **7**, 205–229. doi: 10.1145/65943.65945.

Ramazi, P., Kunegel-Lion, M., Russell, G. & Lewis, M.A. (accepted) Predicting insect outbreak using machine learning: A mountain pine beetle case study. *Ecology and Evolution*.

Régnière, J. & Bentz, B. (2007) Modeling cold tolerance in the mountain pine beetle, Dendroctonus ponderosae. *Journal of Insect Physiology*, **53**, 559–572. doi: 10.1016/j.jinsphys.2007.02.007.

Rish, I., Thyreau, B., Thirion, B., Plaze, M., Paillere-martinot, M.l., Martelli, C., Martinot, J.l., Poline, J.b. & Cecchi, G.A. (2009) Discriminative network models of schizophrenia. *Advances in Neural Information Processing Systems*, pp. 252–260.

Robertson, C., Nelson, T.A. & Boots, B. (2007) Mountain Pine Beetle Dispersal: The Spatial–Temporal Interaction of Infestations. *Forest Science*, **53**, 395–405.

Robinson, J.W., Hartemink, A.J. & Ghahramani, Z. (2010) Learning Non-Stationary Dynamic Bayesian Networks. *Journal of Machine Learning Research*, **11**.

Safranyik, L. (1978) Effects of climate and weather on mountain pine beetle populations. *Proceedings of Symposium on Theory and Practice of Mountain Pine Beetle Management in Lodgepole Pine Forests, April 25-27, 1978, Washington State University, Pullman, Washington* (eds. A.A. Berryman, G. Amman & R. Stark), pp. 77–84. College of Forestry, Wildlife and Range Sciences, University of Idaho, Moscow, Idaho.

Safranyik, L., Carroll, A., Régnière, J., Langor, D., Riel, W., Shore, T., Peter, B., Cooke, B., Nealis, V. & Taylor, S. (2010) Potential for range expansion of mountain pine beetle into the boreal forest of North America. *The Canadian Entomologist*, **142**, 415–442. doi: 10.4039/n08-CPA01.

Safranyik, L., Shrimpton, D.M. & Whitney, H.S. (1975) An interpretation of the interaction between lodgepole pine, the mountain pine beetle, and its associated blue stain fungi in western Canada. *Management of Lodgepole Pine Ecosystems Sympo sium Proceedings*, pp. 406 – 428. Washington State University Coop. Extension Service, Pullman, WA, Pullman, Washington, USA.

Safranyik, L. (2004) Mountain Pine Beetle Epidemiology in Lodgepole Pine. *Mountain Pine Beetle Symposium: Challenge and Solutions. October 30-31, 2003, Kelowna, British Columbia.* (eds. T. Shore, J. Brooks & J. Stones), Information Report BC-X-399, pp. 33 – 40. Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, B.C.

Safranyik, L. & Carroll, A.L. (2006) The biology and epidemiology of the mountain pine beetle in lodgepole pine forests. *The mountain pine beetle: a synthesis of biology, management and impacts on lodgepole pine* (eds. L. Safranyik & B. Wilson), pp. 3–66. Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, Canada.

Saito, T. & Rehmsmeier, M. (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, **10**, e0118432. doi: 10.1371/journal.pone.0118432.

Scholkopf, B. & Smola, A.J. (2001) *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464. doi: 10.1214/aos/1176344136.

Scutari, M. (2009) Learning Bayesian networks with the bnlearn R package. *arXiv preprint arXiv:0908.3817*.

Silander, T. & Myllymaki, P. (2012) A simple approach for finding the globally optimal Bayesian network structure. *arXiv preprint arXiv:1206.6875*.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Martínez Del Rio, C. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, **42**, 4–12. doi: 10.1111/j.1365-2664.2005.01002.x.

Stewart, G.W. (1987) Collinearity and least squares regression. *Statistical Science*, **2**, 68–84. doi: 10.1214/ss/1177013439.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. & Botstein, D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences*, **100**, 8348–8353.

Tsamardinos, I., Aliferis, C.F., Statnikov, A.R. & Statnikov, E. (2003) Algorithms for Large Scale Markov Blanket Discovery. *FLAIRS conference*, vol. 2, pp. 376–380.

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

Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S. & Al-Katheeri, M.M. (2016) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, **13**, 839–856. doi: 10.1007/s10346-015-0614-1.

Zhou, S., Lafferty, J. & Wasserman, L. (2008) Time varying undirected graphs. *arXiv preprint arXiv:0802.2758*.

Zhu, S. & Wang, Y. (2015) Hidden Markov induced Dynamic Bayesian Network for recovering time evolving gene regulatory networks. *Scientific reports*, **5**, 17841.