Ecological Modelling 230 (2012) 50–62

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Metrics for evaluating performance and uncertainty of Bayesian network models

Bruce G. Marcot*

U.S. Forest Service, Pacific Northwest Research Station, 620 S.W. Main Street, Portland, OR 97205, United States

ARTICLE INFO

Article history:
Received 12 September 2011
Received in revised form 10 January 2012
Accepted 11 January 2012

Keywords:
Bayesian network model
Uncertainty
Model performance
Model validation
Sensitivity analysis
Error rates
Probability analysis

ABSTRACT

This paper presents a selected set of existing and new metrics for gauging Bayesian network model performance and uncertainty. Selected existing and new metrics are discussed for conducting model sensitivity analysis (variance reduction, entropy reduction, case file simulation); evaluating scenarios (influence analysis); depicting model complexity (numbers of model variables, links, node states, conditional probabilities, and node cliques); assessing prediction performance (confusion tables, covariate- and conditional probability-weighted confusion error rates, area under receiver operating characteristic curves, *k*-fold cross-validation, spherical payoff, Schwarz' Bayesian information criterion, true skill statistic, Cohen's kappa); and evaluating uncertainty of model posterior probability distributions (Bayesian credible interval, posterior probability certainty index, certainty envelope, Gini coefficient). Examples are presented of applying the metrics to 3 real-world models of wildlife population analysis and management. Using such metrics can vitally bolster model credibility, acceptance, and appropriate application, particularly when informing management decisions.

Published by Elsevier B.V.

1. Introduction

Bayesian networks (BNs) are models that link variables with probabilities and that use Bayes' theorem and associated Bayesian learning algorithms to calculate posterior probabilities of outcome states (Jensen and Nielsen, 2007). BN models are used in many ecological and environmental analyses (Aalders, 2008; McCann et al., 2006; Pourret et al., 2008), in part spurred by the availability of user-friendly computer modeling shells such as Hugin (www.hugin.com), Netica (www.norsys.com), and others, and use of the WinBugs open-source modeling platform (www.mrc-bsu.cam.ac.uk/bugs). As their popularity increases, it becomes more important to ensure rigor in their application to real-world problems (Uusitalo, 2007). Two such areas addressed here are methods for evaluating performance and uncertainty of BN model results. Performance pertains to how well a BN model predicts or diagnoses some outcome, that is, the accuracy of model results. Uncertainty pertains to the dispersion of posterior probability values among different outcome states, that is, the spread of alternative predictions. Ideally, the best model would have high performance and low uncertainty, but to date their measures are either lacking or have not been well summarized.

The purpose of this paper is to present a selected set of existing and new metrics for gauging BN model performance and uncertainty, including: assessment of model sensitivity and influence of input variables; various measures of model complexity, prediction performance, error rates, model selection, and model validation; and various metrics for depicting uncertainty of model output. I demonstrate application of the metrics to published, real-world BN models, and their degree of correlation and performance characteristics. I then summarize the utility and caveats of the metrics and conclude with the need for considering such metrics to bolster model credibility, acceptance, and appropriate application, particularly when informing management decisions.

2. Methods

2.1. Background on Bayesian network models

BN models can vary in their construction but most consist of variables represented as nodes with discrete, mutually exclusive states (Cain et al., 1999). Each state is represented with a probability. Types of variables ("nodes") in a BN model include: inputs (covariates, prediction variables) with states comprised of unconditional, marginal, prior probabilities; outputs (response variables) with states calculated as posterior probabilities; and, in many models, intermediate summary nodes (latent variables), with states

Abbreviations: AIC, Akaike information criterion; AUC, area under the (receiver operating) curve; BIC, Bayesian information criterion; BN, Bayesian network; CPT, conditional probability table; GCM, global circulation model; GHG, greenhouse gas; PPCI, posterior probability certainty index; $PPPCI_{MAX}$, maximum PPCI value given one or more state probability values; $PPCI_{MIN}$, minimum PPCI value given one or more state probability values; PPD, posterior probability distribution; ROC, receiver operating characteristic (curve); SP, spherical payoff; TSS, true skill statistic; VR, variance reduction.

* Tel.: +1 11 503 808 2010.
E-mail addresses: bmarcot@fs.fed.us, brucem@SpiritOne.com

0304-3800/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.ecolmodel.2012.01.013

comprised of conditional probabilities (Marcot et al., 2006). Variables also can constitute scalars and continuous equations.

Nodes are linked according to direct causal or correlative relations between variables. BN model structure – including selection and linkage of variables and their states, and their underlying probability values – can be defined by expert judgment, use of empirical data, or a combination. BN decision models include decision nodes and utility nodes. Before calculations can be made of posterior outcome probabilities, most nodes in a BN model must be "discretized" whereby continuous values are represented as discrete states or value ranges. Running a BN model typically consists of specifying input values and calculating the posterior probability distribution (PPD) of the outcome variable(s). For a given application, the set of expected or normal values of the input variables constitute the "normative" model scenario (e.g., as used by Jay et al., 2011).

2.2. Metrics of model sensitivity and influence

2.2.1. Sensitivity analysis

Sensitivity analysis in BN modeling pertains to determining the degree to which variation in PPDs is explained by other variables, and essentially depicts the underlying probability structure of a model given prior probability distributions. Model sensitivity can be calculated as variance reduction with continuous variables or entropy reduction with ordinal-scale or categorical variables.

As used in the modeling shell Netica (B. Boerlage, pers. comm.), variance reduction (VR) is calculated as the expected reduction in the variation of the expected real value of an output variable Q that has q states, due to the value of an input variable F that has f states. The calculation is $VR = V(Q) - V(Q|F)$, where $V(Q) = \sum_q P(q)[X_q - E(Q)]^2$, $V(Q|F) = \sum_q P(q|f)[X_q - E(Q|f)]^2$, $E(Q) = \sum_q P(q)X_q$, where X_q is the numeric real value of state q, $E(Q)$ is the expected real value of Q before applying new findings, $E(Q|f)$ is the expected real value of Q after applying new findings f for variable F, and $V(Q)$ is the variance of the real value of Q before an new findings. Entropy reduction, I, is calculated as the expected reduction in mutual information of Q from a finding for variable F, calculated as

$$I = H(Q) - H(Q|F) = \sum_q \sum_f \frac{P(q,f)\log_2[P(q,f)]}{P(q)P(f)}$$

where $H(Q)$ is the entropy of Q before any new findings, $H(Q|F)$ is the entropy of Q after new findings from variable F, and Q is measured in information bits (Marcot et al., 2006). Alternatively, sensitivity structure can be determined through simulation (e.g., Thogmartin, 2010), such as by generating a large number of simulated data sets and analyzing the covariation between values of input variables and PPDs.

From the results of a sensitivity analysis, input variables can be rank-ordered or compared quantitatively as to the degree to which each reduces variance or uncertainty (entropy) in a specified outcome variable. Typically, sensitivity is calculated with input variables set to their default prior probability distributions because specifying the value of an input variable sets its sensitivity value to zero, which can also affect sensitivity of the remaining variables; however, this may be a useful method for determining residual sensitivity behavior if one or more inputs are known. More generally, BN models can be used to evaluate sensitivity of a response variable to the probability distributions of other variables.

2.2.2. Influence analysis

In contrast to sensitivity analysis is what I term influence analysis, which refers to evaluating effects on PPDs from selected input variables set to best- or worst-case scenario values. Resulting PPDs are then compared with those generated from the normative model and from other scenario settings. The difference between influence analysis and sensitivity analysis is that specifying the value of an input variable forces that variable's sensitivity value (variance or entropy reduction) to zero whereas it still may have a high influence on the PPD outcome. Conducting influence runs can help reveal the degree to which individual or sets of input variables could affect outcome probabilities. This is helpful in a decision setting, where management might prioritize activities to best effect desirable, or to avoid undesirable, outcomes.

2.3. Metrics of model complexity

Much of ecological statistical modeling strives to balance accuracy with parsimony in explanation of some outcome (Burnham and Anderson, 2010), because overly complex models can perform poorly (Adkison, 2009). The parsimony criterion refers to identifying the simplest model that still provides acceptable results, and can be depicted by several metrics of BN model complexity.

Two simple metrics of BN model complexity are number of variables (nodes) and number of links. More involved metrics include total numbers of: node states (of categorical, ordinal, and discretized continuous states of all variables), conditional probabilities (excluding marginal prior probabilities), and node cliques (subsets of fully interconnected nodes). Total number of conditional probabilities is

$$\sum_{i=1}^{V} \left[S \prod_{j=1}^{n} P_j \right]$$

where S = no. states of the child node, P_j = no. of states of the jth parent node, for n parent nodes, among all V nodes in the model.

Further, any of these metrics of model complexity could be partitioned by type of node (nature, decision, utility, or constant) involved. Overall, metrics of model complexity are not necessarily correlated. For example, a model with n nodes could be structured (linked) in many different ways, with nodes bearing few to many states. Thus, using >1 metric of model complexity can help to represent the fuller array of model architectures when addressing questions of parsimony.

2.4. Metrics of BN model prediction performance

Several metrics can be used to evaluate the performance of BN models when cases are available for which outcomes are known. Existing metrics useful to BN modeling include use of confusion tables, receiver operating characteristic curves, k-fold cross-validation, and performance indices such as spherical payoff, Schwarz' Bayesian information criterion, and true skill statistic. New metrics offered here also include covariate-weighted and conditional probability-weighted confusion error.

2.4.1. Error rates and confusion tables

Evaluating the performance accuracy of BN model predictions typically entails comparing highest-probability predictions to known case outcomes. Error rates are then calculated for false positives (Type I error, rejecting a true hypothesis), false negatives (Type II error, failing to reject a false hypothesis), and their sum, and are depicted in so-called confusion tables (Kohavi and Provost, 1998).

A new variation on this approach is to consider acceptable thresholds of posterior probability outcomes of predictions that might be less than the dominant probability prediction, or, as used by Gutierrez et al. (2011) where predictions match known outcomes within ±1 "bin" (discrete outcome category). In decision analysis, the risk attitude of the decision-maker determines the

degree of error they might accept. For example, it may be acceptable to consider any outcome of some population density level at, say, >40% probability, or predictions that range ±1 bin of actual outcomes, as acceptable predictions, even if they may not all be the dominant predicted outcome. In this case, it is possible that >1 population density state might qualify as acceptable, so that overall model error rates could be lower than if only the highest-probability prediction was used to calculate model error rate. In this case, the modeler could define a minimum population density state that is required, such as under a species recovery plan, so that all densities above that threshold would be deemed acceptable. Yet another variation in confusion tables may be simply to weight errors by their prediction probabilities.

2.4.2. Weighted confusion error rates

One new way to address model parsimony and prediction accuracy is to weight confusion error rates by the number of covariates. Lower values then denote the more parsimonious models with low error rates, where parsimony refers to the number of variables in the model. Variations could include using subsets of the overall model error rate, that is, error rates only for particular state outcomes, for example if it was more important for the model to correctly predict a particular habitat condition or stage class of a population than for others, or more important to avoid Type I or Type II errors.

A similar new measure is to weight confusion error rates by the number of conditional probabilities. As with covariate-weighted confusion error, lower values denote better-performing and more parsimonious models but, when weighting by number of conditional probabilities, parsimony refers to complexity of the underlying probability structure.

2.4.3. ROC curves and AUC

A different, commonly used means of depicting model prediction performance is the receiver operating characteristic (ROC) curve (Dlamini, 2010; Hand, 1997). ROC curves plot percent true positives ("sensitivity") as a function of their complement, percent false positives ("1-specificity"). Further, the area under the ROC curve (AUC) is a metric commonly used to judge overall performance of classification models (Hand, 1997). AUC values range [0,1], where 1 denotes no error, 0.5 denotes totally random models, and <0.5 denotes models that more often provide wrong predictions. Different models can be compared by plotting outcomes on the same ROC diagram and comparing AUC values. Further, Cortes and Mohri (2005) provided a useful method for calculating AUC confidence intervals based on confusion error rates.

2.4.4. k-Fold cross-validation

One can also subdivide an empirical data set ("case file") and conduct cross-validation testing by parameterizing the model with one subset of cases and then testing it against the other set. In *k*-fold cross-validation (Boyce et al., 2002), one randomizes the case file set; sequentially numbers the resulting cases; extracts the first $1/k$th of the cases in sequence; parameterizes the model with the remaining $[1 - 1/k]$ cases; and then tests that model against the first $1/k$th cases left out, recording confusion error rates of model predication. Next, the second $1/k$th set of cases are extracted from the full case file set, and the procedure is repeated until all k case subsets have been used. The resulting *k* confusion tables are then averaged for overall model performance.

This approach often uses $k = 10$, although there is no rigorous rule for this. *k*-fold testing is more reliable with large data sets, such that for *c* number of cases, you want to select *k* such that c/k provides a large enough subset of cases to represent replicates of all possible combinations of covariate input values. Specific sample sizes will depend on model complexity, but typically one would want hundreds or even several thousand cases.

2.4.5. Spherical payoff

Another metric to evaluate classification success of BN models is spherical payoff (Hand, 1997), an index that ranges [0,1] with higher values denoting better model performance. Spherical payoff SP is calculated as:

$$\mathrm{SP} = \mathrm{MOAC} \cdot \frac{P_c}{\sqrt{\sum_{j=1}^{n} P_j^2}}$$

where MOAC = mean probability value of a given state averaged over all cases, P_c = the predicted probability of the correct state, P_j = the predicted probability of state *j*, and *n* = total number of states (B. Boerlage, pers. comm.). Spherical payoff is a better metric than the standard AUC when nuances of probability values are an important consideration.

2.4.6. Schwarz' Bayesian information criterion

Schwarz' Bayesian information criterion (BIC) is useful as an index for selecting among alternative model structures when comparing model results to known outcomes (Schwarz, 1978). Training BN models with known case outcomes entails testing alternative CPT values to find the maximum likelihood Bayes network, that is, the network that is most likely given the case data (Neapolitan, 2003). $\mathrm{BIC} = -2 \cdot \ln(\mathrm{ML}) + k \cdot \ln(n)$, where ML = maximum likelihood value, *k* = number of parameters in the model, and *n* = number of observations. BIC is similar to Akaike information criterion (AIC, Akaike, 1973) but the former penalizes more for potential errors in overfitting models to data when increasing the number of model parameters to produce lower classification error rates (e.g., see Huang et al., 2007). As with AIC, one subtracts the lowest BIC value among all models being compared from the BIC value of each alternative model. The smallest differences (ΔBIC) denote the best-performing and most parsimonious model, that is, the model that best balances model error and dimension (Burnham and Anderson, 2010). However, covariate- or conditional probability-weighted confusion error (Section 2.4.2) may have an advantage over BIC by more explicitly incorporating prediction error rates into the performance metric.

2.4.7. True skill statistic and Cohen's kappa

The true skill statistic (TSS) – also called the Hanssen–Kuiper discriminant or skill score – is an index of model performance combining frequencies from a 2×2 confusion table (Allouche et al., 2006; Mouton et al., 2010). TSS is calculated from rates of true positives, true negatives, and Type I and II errors. TSS values range $[-1,1]$; analogous to interpretation of AUC scores, 1 represents a perfectly performing model with no error, 0 a model with totally random error, and −1 a model with total error. Similar to TSS is Cohen's kappa (Boyce et al., 2002), commonly used to test classification success in geographic information systems (e.g., Gutzwiller and Flather, 2011; Zarnetske et al., 2007). Kappa is calculated as the difference between correct observations and expected outcomes, divided by the complement of expected outcomes. Kappa values range [0,1], with 1 being perfect classification.

2.5. Metrics of uncertainty in posterior probability distributions

Several existing and new metrics offered here depict the degree of uncertainty in BN model outcomes, that is, the dispersion of PPD values. Such metrics can be used to help inform decisions based on BN model results in a risk management framework where the level of certainty of predictions are weighted with the risk attitude of the decision-maker. These metrics of uncertainty include use of

Bayesian credible intervals, a posterior probability certainty index and certainty envelope, and a new adaptation of the Gini coefficient and Lorenz curve to depict inequality of PPDs.

2.5.1. Bayesian credible intervals

One existing approach to denoting uncertainty in PPDs is use of Bayesian credible intervals (Bolstad, 2007; Curran, 2005), which are very loosely an analogue to confidence intervals in frequentist statistics (in some literature, they are confusingly referred to as Bayesian confidence intervals). An *X*% Bayesian credible interval of some PPD of an ordinal or continuous scale variable (but not a categorical variable) refers to state-wise probabilities when *X*/2% is excluded from the lowest and highest outcome states. Put another way, it is the interval determined for the expected value over replicate calculations based on uncertainty distributions of the input variables, not for the PPD of a given instance of input values. A Bayesian credible interval represents the PPD at a specified level of acceptability, and in this way differs from a probability density function (and from a frequentist confidence interval).

One must decide what credible interval value *X* to use. In practice, it should be < twice the probability of the most extreme state of interest. For example, if a manager would be concerned over some extreme outcome, such as a species' extinction, if there was a 5% probability or greater of it occurring, then they should not use anything < a 90% credible interval, else it might exclude such an event.

2.5.2. Posterior probability certainty index

Another metric that can be used to evaluate uncertainty of BN model outcomes is the posterior probability certainty index (PPCI). This new metric, first presented here, is based on information theory and specifically is an adaptation of the classic evenness index from species diversity theory (Hill, 1973). Evenness has long been used to measure the relative distribution of species' abundances in a community. Here, I extend the concept to PPDs which consist of p_i probability values among N number of states, where p_i ranges [0,1] and

$$\sum_{i=1}^{N} p_i = 1.0.$$

PPCI is calculated as $(1 - J')$, where $J' = H'/H'_{\max}$,

$$H' = -\sum_{i=1}^{N} p_i L$$

where

$$L = \begin{cases} \ln(p_i), & p_i > 0 \\ 0, & p_i = 0 \end{cases},$$

and $H'_{\max} = \ln(N)$. J' normalizes the metric proportional to N, so that the degree of certainty of PPDs can be compared among outcomes with different numbers of states N.

In information theory, J' is a measure of entropy or uncertainty. In the context of risk management, however, one would hope for the most *uneven* PPD, that is, an outcome that most clearly suggests a particular state with the highest probability. Thus, PPCI ranges [0,1] with higher values denoting greater certainty (greater loading of outcome probabilities into fewer outcome states). Models with higher PPCI values of their PPDs denote greater certainty in outcome predictions. Since PPCI is normalized, it can be compared among different models with different numbers of outcome states.

2.5.3. Certainty envelope

An associated new metric is the certainty envelope, which is the range of possible PPCI values given the probability of one or more states (up to $N-2$ states; solutions for > $N-2$ states are trivial). A specific PPCI value calculated from a known PPD for a given scenario can then be compared to the certainty envelope to determine the relative degree of certainty of that outcome to the range of possible values. The certainty envelope is variable because the possible range of PPCI values varies as a function of the probability values of given states, and only achieves values of 0 and 1 under special circumstances of uniform posterior probability distributions and when a single state achieves a probability of 1.0, respectively. Otherwise, PPCI can be scaled to the more constrained range of values calculated as the certainty envelope.

The certainty envelope has utility in some applications where the probability of only a subset of outcome states is fixed or known, where others might be more fluid or unknown, and where the manager would want to know the certainty of the PPD given just the known outcome states. For example, if a BN model is structured with five different population levels of a wildlife species (possible outcome states), and a particular model run results in predicting probabilities of the two lowest levels, the manager may wish to know the overall degree of certainty of the PPD given that particular outcome. That is, how clumped are probability values as distributed among the outcome states? The less they are evenly distributed, given a particular result for specific outcome states, the more certain can the manager be of overall model results.

The PPCI certainty envelope for a PPD with N states is calculated as follows. For a given state or set of j states, $1 \le j \le (N-2)$, and their known, marginal (summed) posterior probability value(s)

$$m = \sum_{i=1}^{j} P_i$$

there exists a specific range of possible values [PPCI_{MIN}, PPCI_{MAX}], where PPCI_{MIN} is calculated by setting all other $N-j$ states to uniform probabilities, and PPCI_{MAX} is calculated by setting only one other state to the remaining probability and all other states to zero. That is, PPCI_{MIN} is calculated from

$$H' = -\left\{ \sum_{i=1}^{j} P_i L + \sum_{i=j+1}^{N} \left[\left(\frac{1-m}{N-j} \right)_i \ln \left(\frac{1-m}{N-j} \right)_i \right] \right\}$$

and PPCI_{MAX} is calculated from

$$H' = -\left\{ \sum_{i=1}^{j} P_i L + (1-m)\ln(1-m) \right\}$$

where L is defined above. Thus, for a given PPD with a specified probability of a given state or set of j states, $\text{PPCI}_{\text{MIN}} \le [\text{PPCI}|P(j)] \le \text{PPCI}_{\text{MAX}}$. For instance, with $N = 5$ states where the probability of one state j is known, the raw (non-normalized) certainty envelope results in concave upward distributions of PPCI_{MIN} and PPCI_{MAX} with convergence at 1.0 (Fig. 1). The PPCI_{MIN} curve increasingly skews to the right with greater number of states.

To best compare PPCI values among competing models particularly with different total numbers of states N or different numbers of specified state values j, the range [PPCI_{MIN}, PPCI_{MAX}] can itself be normalized to [0,1], and the relative position of a given value of [PPCI|$P(j)$] within this range can be calculated by simple linear interpolation. Thus, the interpolated value of [PPCI|$P(j)$] represents the proportion (or percentage) of total possible certainty for a given outcome state(s) j. This could be valuable information for a manager faced with only one piece of information, such as the probability of

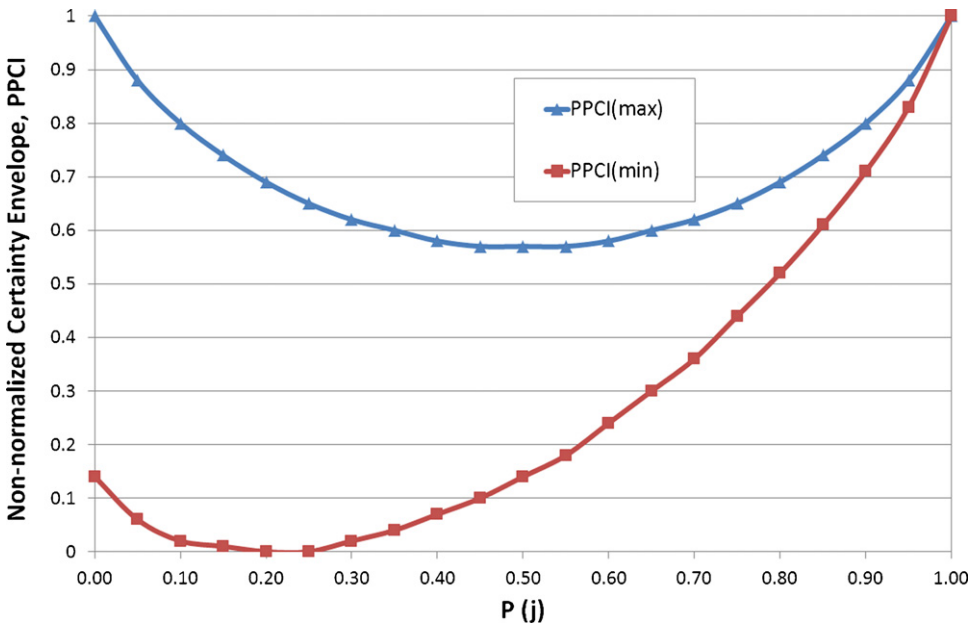

Fig. 1. Non-normalized certainty envelope of a probability distribution with $N = 5$ states across the full range of probability values of a single state, $P(N_S)$.

extinction of a species, who might also wish to know the level of certainty of the rest of the distribution, i.e., the general dispersion of probabilities among the remaining states.

2.5.4. Inequality of posterior probability distributions

Another new measure of the dispersion of PPD values can make use of what is known as the Gini coefficient used in econometrics to depict geographic and social distribution of wealth and resources (Atkinson, 1970; Sadras and Bongiovanni, 2004). The Gini coefficient varies in the range [0,1], and if applied to PPDs in BN models, a value of 0 represents a uniform probability distribution (complete uncertainty) and 1 represents a distribution with one state at 100% probability and all other states at 0% (complete certainty).

The Gini coefficient is calculated as the area under the Lorenz curve, which, applied to BN modeling, is the cumulative probability among outcome states rank-ordered by decreasing values of their individual probabilities. Lorenz curves have been used in ecology to represent the distribution of species abundance proportions in ecological communities (Ricotta and Avena, 2002). Applied here to BN modeling, the Lorenz curve plots cumulative proportion of posterior states as a function of cumulative probability of posterior states, anchored to the (0,0) plot origin.

For example, say there is a PPD among 4 outcome states with the respective probability vector [0.10, 0.85, 0, 0.05]. The vector is first reordered in decreasing values, viz., [0.85, 0.10, 0.05, 0], and then the respective cumulative frequency distribution is calculated as [0.85, 0.95, 1, 1]. To account for the origin anchor, an initial zero is inserted as [0, 0.85, 0.95, 1, 1], constituting *x*-axis values. Corresponding *y*-axis values are merely the increasing proportion of the number of states beginning with 0; for our 4-state example here, those values become [0, 0.25, 0.50, 0.75, 1]. Because PPDs in BN models sum to 1, associated Lorenz curves always span the domain [0,1].

The Lorenz curve plot also includes a positive diagonal line, here spanning (0,0) to (1,1); this is called the line of perfect evenness in econometric literature, and here represents the line of total uncertainty or highest entropy, that is, the line formed from a uniform probability distribution. The Gini coefficient then is calculated as the area subtended between the Lorenz curve and the line of total uncertainty (perfect evenness). One great advantage of using the Gini coefficient as a measure of the dispersion of posterior probabilities is that values derived from different models with different numbers of outcome states can be directly compared because Lorenz curves derived from PPDs always span [0,1].

However, one final correction needs to be applied for use in BN models. Because BN variables are usually discretized into a finite number of mutually exclusive states, the maximum value of the area within the Lorenz curve – doubled, so that the resulting Gini coefficient theoretically falls in [0,1] – for n states is $1 - (1/n)$, and asymptotically, $\lim_{n\to\infty}(2 * \text{area}) = 1.0$ (Fig. 2). E.g., if an outcome node has $N = 4$ discrete states, the maximum value of 2*area = 0.75. This maximum area value can be used as a normalizing constant, that is, by dividing the observed area by this correction factor. Then, the resulting adjusted Gini coefficient values can be directly compared among models with different numbers of outcome states. Because of the asymptotic nature of the correction curve, the actual Gini values will range [0,1). Exact and approximate calculations of the Gini coefficient with discrete BN model outcome states are presented in Appendix A.

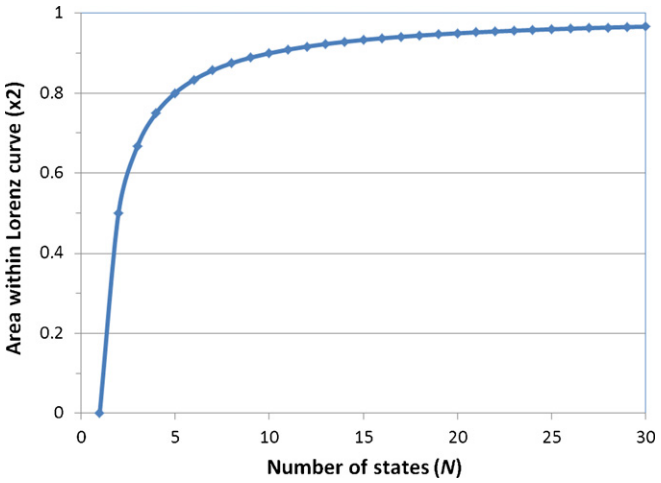

Fig. 2. Maximum value of the area within the Lorenz curve (=Gini coefficient, here doubled) increases asymptotically to 1.0 as a function of the number of states, as applied to posterior probabilities from Bayesian network models.

2.6. Metrics for comparing alternative posterior probability distributions

Another way to compare results of running BN models among alternative scenarios is to directly compare PPDs with existing statistical tests. Several such tests may be appropriate. For example, one could use Mantel's r correlation test (Mantel, 1967) which determines the correlation between two matrices. In BN modeling, it could determine the correlation between confusion error matrices resulting from competing models. For this test, ideally the matrices must be of the same order and rank, so either the BN models must have identical states structures denoted in the output node or results from one model must be collapsed to match the number and state structure of the other model.

PPCI metrics also can be statistically compared among model results. PPCI metrics, like species diversity indices from information theory, likely tend to be non-normally distributed. Thus, it may be more appropriate to use non-parametric tests such as the Kolmogorov–Smirnov test which can be used to compare two PPDs presented as cumulative frequency distributions.

Another approach for comparing PPDs is to generate simulated case files as "replicates" for different model scenarios, and determine the distribution of PPD values among the replicates for each scenario and compare PPCI values among the scenarios. Scenarios could represent competing models, or different data sets could be simulated with different variables missing to determine which missing variable produces the greatest uncertainty in PPDs as a way to prioritize research or monitoring activities. Such an analysis would complement strict sensitivity and influence analyses.

3. Results and discussion

To demonstrate application of the performance and uncertainty metrics discussed here, I use 3 case examples selected from my published Bayesian network models pertaining to forecasting population viability of polar bears (*Ursus maritimus*; Amstrup et al., 2008) and Pacific walrus (*Odobenus rosmarus divergens*; Jay et al., 2011), and using tissue samples to predict age of martens (*Martes americana* and *M. caurina*; Pauli et al., 2011). I also statistically compare metric values through correlation analysis to determine their degree of redundancy or complementarity, and provide a summary of the purpose, assumptions, strengths, and weaknesses of each metric as a guideline to help select appropriate metrics.

3.1. Case example 1: Polar bear populations

This case example illustrates sensitivity analysis metrics, the comparison of normative to influence analysis model runs, use of PPCI values in depicting confidence of PPDs, and calculations of Gini coefficient values under the Lorenz curve, for identifying the major factors influencing polar bear populations and determining the degree of confidence and certainty of those projections.

An assessment of global polar bear populations entailed developing BN models to evaluate the species' future viability as influenced by a broad suite of stressors in each of 4 ecoregions of the world. Stressors included potential effects of climate change on sea ice habitat and prey availability, changes in disease and predation, direct human disturbance, pollution, and influence of habitat threats on population demography. The model was parameterized from ensemble mean outcomes of a suite of 18 global circulation (climate change) models (GCMs) for the sea ice variables, and from review of the literature, professional judgment, and peer review for the other stressor variables. The normative models were run for recent historic, current, and 4 time periods into the 21st century, and under 4 greenhouse gas (GHG) concentration scenarios. Polar bear outcomes were denoted on a 5-category ordinal scale

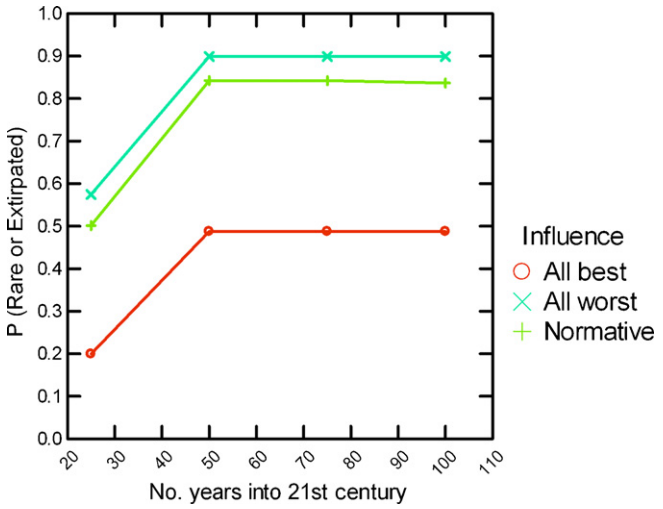

Fig. 3. Example of influence analysis of polar bear populations, under the IPCC B1 greenhouse gas scenario, plotting the probability of polar bear populations in the Seasonal Ice Ecoregion becoming rare or extirpated outcomes at 4 future time periods (Amstrup et al., 2010). "Normative" results refer to setting all stressors to their expected values for each time period, whereas the "all best" and "all worst" results refer to setting them to their best and worst case influences.

comparing population distribution and abundance in each ecoregion to current conditions, viz., larger than now, same as now, smaller than now, rare, and extirpated. Here, we combine probabilities of rare and extirpated states, $P(r \cup e)$, for clarity of presentation.

The model consisted of 36 nodes, 44 links, and 1658 conditional probability values. Of the 36 nodes, 17 were inputs, representing the array of population stressor variables. Sensitivity analysis of the polar bear outcome node to these 17 nodes – with all inputs set to their uniform probability distributions – suggested that 6 of the top 7 nodes with greatest sensitivity influence, constituting a cumulative 92% of total sensitivity in the model, pertained to the potential influence of climate change on sea ice habitat for the species (Table 1).

Two future GHG scenarios (B1 and CCSP450) were further run under 17 influence analyses in which various sets of stressors were set to best- and worst-case conditions, and compared to normative results, totaling 575 model scenario conditions. One set of influence runs demonstrated that $P(r \cup e)$ could be drastically lowered by eliminating population stressors, and that effects of normative (expected) future conditions already are at nearly the worst case levels (Fig. 3). Additional analysis (Amstrup et al., 2008, 2010) partitioned out the influence of each type of stressor and to identified that reducing climate change impact on future reduction of sea ice habitat could have the greatest positive benefit on the species' viability. Further influence analyses revealed that reducing harvest, adverse effects of pollution and sea traffic, and other controllable stressors would help reduce adverse population outcomes, but to a far lesser extent than would ameliorating climate change impacts. Results of these influence analyses have been used to inform recovery planning for the species and to convey to managers realistic expectations of implementing such control guidelines.

As for the degree of certainty of PPD outcomes, under normative results from expected future stressor levels, PPCI values plotted against posterior probability outcomes suggest greater certainty in BN probability outcomes when $P(r \cup e)$ is either very low (e.g., <0.2, only at near-future time periods) or very high (e.g., >0.8, only at mid-century or later time periods) (Fig. 4). That is, the PPCI plot clearly reveals that greater uncertainty of outcomes (more even spreads of outcome probabilities of various potential population conditions) occurs at intermediate values of $P(r \cup e)$, which occurs with various combinations of ecoregions and time

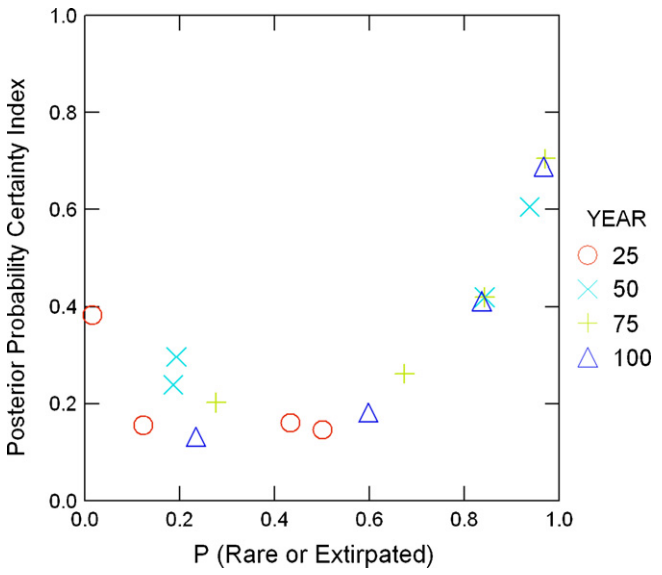

Fig. 4. Example of the posterior probability certainty index (PPCI; 1 = full certainty, 0 = full uncertainty) as a function of the posterior probability of polar bear populations becoming rare or extirpated in each of four global ecoregions, by future time period, under the IPCC B1 greenhouse gas scenario.

periods. This result suggests that those particular combinations might warrant greater scrutiny by which to better forecast future population outcomes with higher confidence and certainty. In this way, PPCI values can contribute information to help prioritize further research and population model refinement.

In addition, Gini coefficient values were calculated for 16 normative scenario conditions, representing two GHG scenarios (B1 and CCSP450) in two ecoregions (Seasonal Ice and Archipelago), each over 4 time periods into the 21st century (Amstrup et al., 2008). For this example set, $P(r \cup e)$ ranged 0.001–0.842, Gini values ranged 0.38–0.78 and PPCI values ranged 0.13–0.54. Correlations between Gini and PPCI values are explored further below.

3.2. Case example 2: Pacific walrus population

This case example illustrates another use of influence analysis model runs, and demonstrates use of PPCI and certainty envelope values, as well as Bayesian credible intervals in depicting confidence of model results, for identifying which major stressors most influence the future walrus population and the degree of certainty of those projections.

The Pacific walrus population was modeled under conditions of 1 past, 1 near-present, and 4 future time periods. Future conditions derived from 2 alternative IPCC GHG scenarios (A1B and A2) and associated anthropogenic and other environmental stressors projected over the next century throughout the species' range in the Chukchi and Bering Seas. Walrus outcomes were depicted as an ordinal-scale set of 5 possible population conditions, viz., robust, persistent, vulnerable, rare, and extirpated. Three seasonal submodels addressed the walrus' important life history events with somewhat different arrays of stressors in each season. For example, variables depicting sea ice over the continental shelf were included in the winter submodel as breeding substrates, and in the spring submodel as birthing platforms. As with the polar bear model, sea ice inputs and CPTs were established using a distribution of two sets of projections from 18 GCMs for the sea ice variables, and using literature review, professional judgment, and peer review for the other stressor variables.

A total of 120 BN model runs were conducted including both normative and influence analysis scenarios, and PPCI indices were calculated for each outcome. Influence runs were conducted to compare the potential effects of best- and worst-case conditions among specific sets of stressors with the normative (expected) outcomes. For example, influence runs analyzed the probability $P(v \cup r \cup e)$ of walrus populations becoming vulnerable, rare, or extirpated from possible positive and negative effects of climate change on ocean benthic food productivity, frequency of ship and air traffic, extent and duration of sea ice, and direct harvest by people (Fig. 5). In this influence-run analysis, two levels of harvest greater (worse) than normative (expected) conditions were modeled. Results of the influence analyses were quite helpful in identifying that $P(v \cup r \cup e)$ could be lowered the most by retaining or restoring sea ice, and could be raised the most with the higher level of harvest. However, the small influence noted in varying climate change on benthic production and ship and air traffic likely was due to lack of scientific understanding of those stressor effects. In this way, influence analysis helped identify the key factors, and key unknowns, affecting the degree to which populations could be sustained should conditions or stressors be fully regulated or become most deleterious.

PPCI values were calculated for each of the 120 normative and influence analysis model runs, providing information on the level of certainty of model results. As examples, PPCI outcomes for normative and a subset of best-case influence analysis runs strongly suggested that conditions of best-case future sea ice habitat provides a substantially greater certainty of model results (Fig. 6). That is, the manager could be most confident that the model provided a dominant outcome (in this case, of persistent or robust walrus populations, the complement of probabilities shown in Fig. 5) if future sea ice were to be restored to highest levels observed historically. Otherwise, certainty levels of model results are less increased over those of normative conditions when provided with a combination of best habitat conditions other than sea ice and minimizing walrus harvests (Fig. 6). However, it is vital to remember that, without independent validation, greater dominance of a predicted future outcome state is not confirmation that the model is accurate.

An example of calculating the PPCI certainty envelope is based on the posterior probability of walrus populations becoming extirpated under specific influence conditions, in this case that result in $P(\text{extirpation}) = 0.02$. Minimum certainty ($\text{PPCI}_{\text{MIN}}$) is calculated by retaining the model outcome value of $P(\text{extirpation})$ and then setting all remaining states to uniform values so that all states sum to 1, and maximum certainty (PPCI_{MAX}) is calculated by setting only one remaining state to $1 - P(\text{extirpation})$ and the rest to 0. In this example, the PPCI value from the model outcome (PPCI = 0.23) is closer to the minimum possible value ($\text{PPCI}_{\text{MIN}} = 0.09$) than the maximum possible value ($\text{PPCI}_{\text{MAX}} = 0.94$) – and the interpolated PPCI value normalized to a range [0,1] is 0.16 – suggesting relatively high uncertainty of population outcome under these specific modeled conditions.

Again, this may be valuable information for managers concerned with future rarity or loss of this threatened species. In this example, although results suggest relatively high uncertainty in outcome conditions (normalized PPCI = 0.16), the outcome of $P(\text{extirpation}) = 0.02$ might itself be low enough to not necessarily warrant more detailed investigation and analysis, should this particular influence scenario be of potential management interest and the certainty of other outcome states of less concern. Further, this model outcome can be viewed in the context of a broader example of 10 BN model influence runs which vary by level of certainty and range of the certainty envelope; this particular influence run provided slightly higher certainty levels than did most of the other runs (Fig. 7). In this broader context, other model runs with far higher $P(\text{extirpation})$ and far lower levels of certainty (PPCI values) might warrant greater scrutiny.

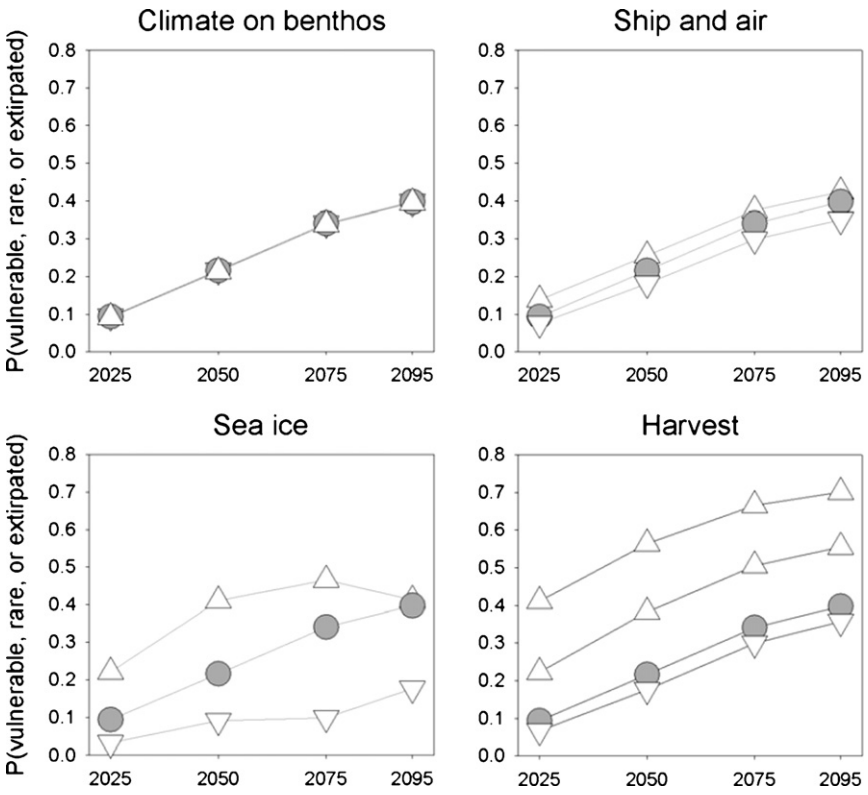

Fig. 5. Results of BN model runs of future Pacific walrus populations under normative (gray circles) and best-case and worst-case influence analyses (open triangles pointing downward and upward, respectively), plotting probability of future populations becoming vulnerable, rare, or extirpated as a function of future time period, under the IPCC A1B greenhouse gas scenario.

The influence analysis model run considered above also can illustrate use of Bayesian credible intervals. Posterior probability values of outcome states robust, persistent, vulnerable, rare, and extirpated were [0.41, 0.33, 0.21, 0.03, 0.02], respectively. With such ordinal-scale outcome states, one simple use of an *X*% Bayesian credible interval is to strip *X*/2% from the upper and lower ends of this PPD, thereby preserving equal probability of being above and below the interval. E.g., the 90% Bayesian credible interval for this particular distribution is [0.36, 0.33, 0.21, 0, 0]. In other words, after applying this particular influence analysis on walrus populations, a 90% credible interval means that the posterior probability of the population outcome state being robust, persistent, or vulnerable, is 0.9. Other ways to form credible intervals include centering the trimming of probability values on the mean, and ensuring that the median or most-probable outcome state are included.

3.3. *Case example 3: Marten telomere–age relationships*

This case example illustrates comparative metrics of model complexity and prediction performance. The models in this case entail testing whether age class or age in years of martens (members of the weasel family) can be determined from telomere length

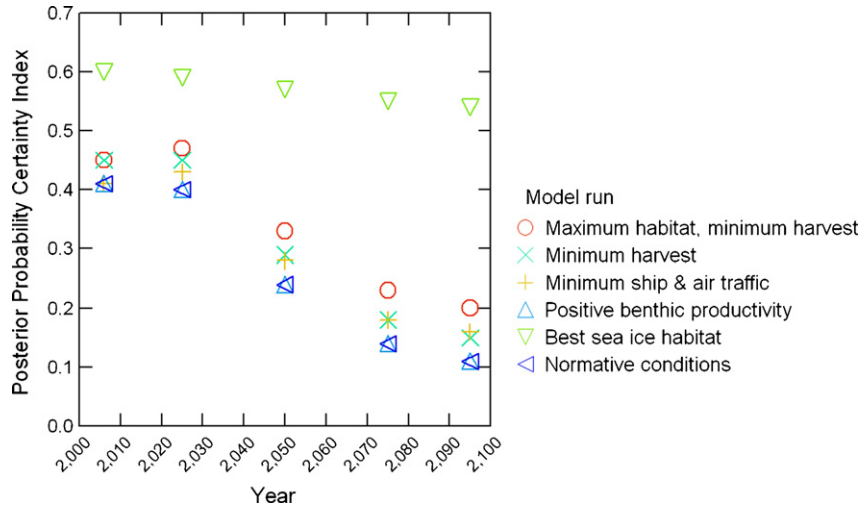

Fig. 6. Example of the posterior probability certainty index (PPCI; 1 = full certainty, 0 = full uncertainty) as a function of future time period, from posterior probability distributions of Pacific walrus populations, under normative and selected best-case influence analysis scenarios, under the IPCC A1B greenhouse gas scenario.

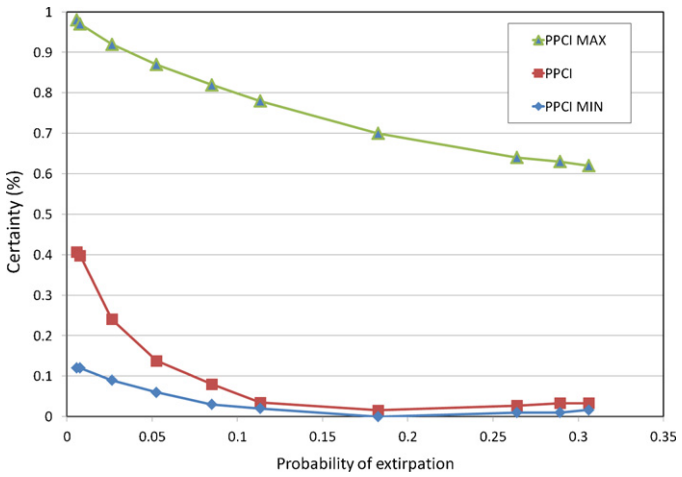

Fig. 7. Example of 10 influence analysis BN model results on Pacific walrus population outcomes, illustrating how the PPCI certainty envelope and run-specific PPCI outcomes vary by probability of extirpation.

Table 1
Example of results of sensitivity analysis conducted on a Bayesian network model of global polar bear population response to environmental contexts and stressors. Stressor parameters are listed here in decreasing sensitivity value (entropy reduction; see text for explanation).

Stressor parameter	Entropy reduction	Cumulative percent of entropy reduction
Foraging habitat quantity change	0.1297	42
Foraging habitat absence change	0.0488	58
Ecoregion	0.0417	71
Alternate regions available	0.0259	79
Intentional takes	0.0161	84
Shelf distance change	0.0139	89
Foraging habitat character	0.0104	92
Bear–human interactions	0.0082	95
Parasites and disease	0.0051	96
Hydrocarbons and oil spills	0.0027	97
Oil and gas activity	0.0025	98
Shipping	0.0020	99
Predation	0.0009	99
Contaminants	0.0007	99
Alternate prey availability	0.0007	99
Relative ringed seal availability	0.0007	100
Tourism	0.0004	100

Source: Amstrup et al. (2008).

and other variables. A set of 399 marten specimens of two species were measured in the lab for various body metrics, and their key features of their habitats were recorded from their field locations. Telomeres are the ends of chromosomes that generally progressively shorten with each cell division and thus over the life of an individual, but not necessarily linearly so. Telomere length itself can be affected by health of the individual, which in turn is affected by disease, population crowding and stress, availability and productivity of food, and perhaps gender and species; these covariates or their proxies were recorded for each marten specimen, along with actual age measured by standard methods of counting cementum annuli (rings) in cross-sections of selected teeth. The aim of the study was to develop models to predict age of martens from non-invasive samples such as hair and faeces, so that the age structure of local populations could be determined without needing to capture individuals of these secretive and elusive species.

Thirty-six alternative models were developed using different sets of covariates and response variables. CPT values of each model were established from all or a subset of the 399 marten cases by using the expectation maximization algorithm (see Marcot, 2006

Table 2
Example of metrics of Bayesian network model complexity and prediction performance, using models predicting age of martens (*Martes* spp.) from genetic, somatic, and environmental covariates (Pauli et al., 2011). Shown here are results of 2 models (selected as final versions based on criteria for low to moderate complexity and high prediction performance) and the range of outcomes of the full set of all 36 models.

Metric	Live capture model[a]	Non-invasive model[b]	Range of values among all 36 models
Model complexity			
Number of covariates (=number of node links[c])	6	4	4–7
Number of conditional probabilities[d]	56,266	1051	1051–1,663,299
Number of node states[d]	45	28	28–93
Model prediction performance			
Confusion error (%)			
Juveniles only	0.5	14.0	0.0–48.6
Adults only	19.4	29.7	10.2–80.6
All ages	9.5	21.5	6.5–38.6
10-Fold cross-validation error (%)			
Juveniles only	0.4%	24.9%	N/C
Adults only	71.9%	40.2%	N/C
All ages	34.9%	32.3%	N/C
Covariate-weighted confusion error			
Juveniles only	2.9	56.0	0.0–243.2
Adults only	116.2	118.8	40.7–403.1
All ages	57.1	86.0	38.8–193.0
Conditional probability-weighted confusion error			
Juveniles only	272	226	0–63,926
Adults only	10,900	312	107–517,245
All ages	5359	226	130–242,342
Spherical payoff	0.94	0.84	0.70–0.96
Schwarz' Bayesian information criterion	643.5	513.6	70.8–1350.2

N/C, not calculated for all models.

[a] Models intended for use with captured animals for which age is unknown. Covariates for these models include telomere length, marten population density, zygomatic (jaw) width, species, sex, age class (juvenile or adult); the models predict age class (age in years 0, 1, 2, 3, and 4+). Zygomatic width data were unavailable for continent-only specimens, so that model was not developed.

[b] Models intended for use with animals sampled non-invasively (i.e., without capture). Covariates include telomere length, marten population density, species, and sex; the models identify juveniles (<1 year old) from adults (≥1 year old).

[c] All models were structured with simple links from each covariate to one response variable.

[d] Including covariate and response variables.

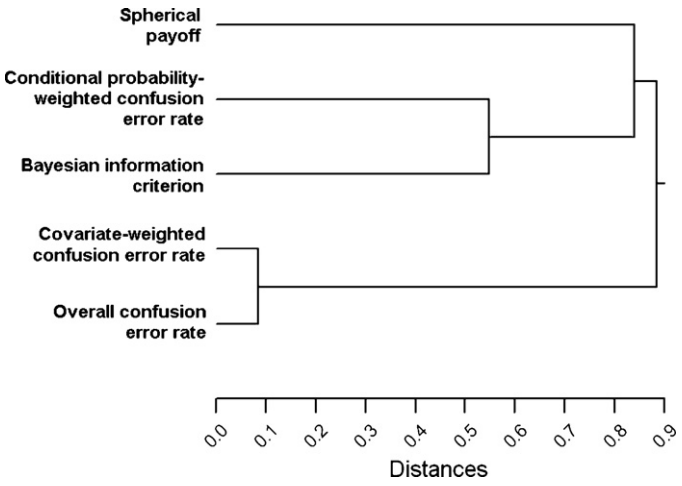

Fig. 8. Similarity (single linkage using Euclidean distance) of 5 metrics of Bayesian network model prediction performance based on the case example of modeling marten (*Martes* spp.) age as a function of genetic, somatic, and environmental variables (see text).

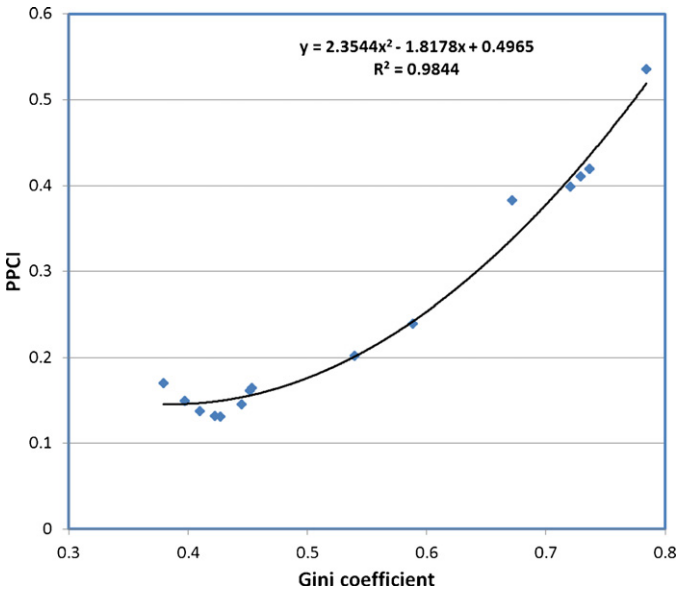

Fig. 9. Nonlinear relationship (best-fit polynomial) between two metrics of Bayesian network model certainty (dispersion of posterior probability values) – the posterior probability certainty index (PPCI) with the Gini coefficient – based on 16 influence analysis runs of the polar bear population outcome model (Fig. 4).

for formula details). Each model was then characterized with metrics of model complexity and tested against part or all of the 399 marten cases for accuracy of classification success (Pauli et al., 2011). Because some models resulted in prediction of >2 outcome categories, the performance metrics AUC and TSS were not calculated.

Results of evaluating and testing all models were used to identify the best models having low to moderate complexity and highest prediction performance (Table 2). The two final models considered here were parsimonious in having only 4 or 6 covariates, although the number of conditional probabilities escalated multiplicatively with the total number of node states in the models. Testing the models suggested far lower classification error when predicting juvenile than adult age classes. Overall model performance ranged good to excellent, however, as judged by spherical payoff values (Table 2).

3.4. Comparison of metrics

The various measures of sensitivity, influence, complexity, prediction performance, and uncertainty discussed here provide complementary insights into the probability structure, relative importance of input variables, parsimony, accuracy, and reliability of BN models. The literature and the case examples explored here also provide some comparison of selected metrics.

3.4.1. Complexity metrics

Beyond their use of summarizing basic model architecture, complexity metrics are quite useful for comparing alternative model structures, such as used with the marten case example. Metrics of numbers of variables, links, node states, conditional probabilities,

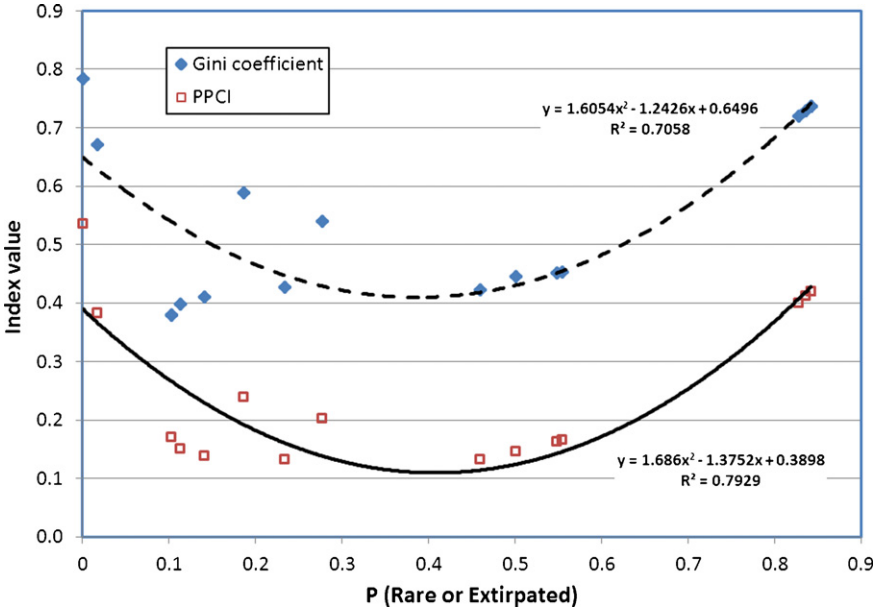

Fig. 10. Nonlinear and nonmonotonic relationships of two Bayesian network model certainty metrics, with posterior probability values of polar bear populations being rare or extirpated, among 16 influence analysis runs (Figs. 4 and 9).

Table 3
Summary of metrics potentially useful for assessing performance and uncertainty of Bayesian network models.

Metric	Use	Caveats and assumptions
Model sensitivity and influence		
Variance reduction	Applied to continuous variables.	Input variables are set to their default priori probabilities unless specifically desired otherwise.
Entropy reduction	Applied to ordinal-scale or categorical variables.	(As above)
Case file simulation	Analysis of covariation between input variables and output variable probability distributions.	Simulated cases must cover all covariation conditions with adequate sample sizes.
Influence analysis	Determines incremental effects of selected inputs set to best, worst, or other specified values.	Best used for scenario analysis and not necessarily for prediction.
Model complexity		
Number of variables	Determines degrees of freedom.	Important to include latent variables.
Number of links	Depicts degree of connectivity of variables in the model.	Important to denote and include any links between correlated input variables.
Number of node states	Affects model precision and overall number of probability values in the model.	Count number of states after discretizing continuous functions to desired precision.
Number of conditional probabilities	Sensitive to model structure, including variable connections and precision.	Does not include prior (unconditional) probability tables; could include if desired.
Number of node cliques	Depicts degree of independence of effects from input node sets.	Node cliques resulting from use of latent variables (intermediate nodes) could bias independence.
Model prediction performance		
Confusion error[a]	Depicts rates of Type I and Type II errors in classification or prediction models.	Typically based on highest probability state, which may oversimplify the utility of the model if other results could be equally useful.
Area under curve (AUC)[a]	Depicts performance of classification success.	Conflates error types. Assumes 2-state outcomes.
k-Fold cross-validation[a]	Tests model predictions against subsets of empirical case data.	Error rates likely inflated with small data sets, as subsets do not represent full range of conditions with adequate replications.
Spherical payoff[a]	Indexes performance of classification models.	Influenced by number of states in the response variable.
Schwarz' Bayesian information criterion[a]	Overall index to model performance and complexity.	Conflates effects of prediction success, number of variables, and case data sample size. Conflates error types.
True skill statistic[a]	Indexes performance of classification models.	Conflates error types. Assumes 2-state outcomes.
Cohen's kappa[a]	Indexes performance of classification models.	Conflates error types. Assumes 2-state outcomes.
Covariate-weighted confusion error[a]	Indexes model prediction error rates weighted by number of covariates (input variables).	Conflates error types. Index values are open-ended.
Conditional probability-weighted confusion error[a]	Indexes model prediction error rates weighted by number of conditional probabilities in the model.	Conflates error types. Index values are open-ended.
Model outcome uncertainty		
Bayesian credible interval	Depicts range of model results for a specified percent level of acceptability.	Assumes continuous or ordinal-scale variables; not meaningful with categorical variables. Interval range should be selected prior to model development.
Posterior probability certainty index & certainty envelope	Indexes degree of dispersion of posterior probability values among outcome states, normalized for number of states and possible minimum and maximum values.	Useful if the degree of spread of probabilities among outcome states is of interest in a decision framework.
Gini coefficient (Lorenz curve area)	Indexes degree of dispersion of posterior probability values among outcome states.	Maximum value is constrained by number of states.

[a] Requires external, independent data.

and node cliques could be highly correlated when comparing alternative model structures based on a common data set. For instance, among the 36 alternative marten models explored, the number of conditional probabilities was significantly correlated with number of covariates (Pearson $r = 0.707$, Bonferroni $P = 0.001$). Nonetheless, especially when combined with measures of model prediction performance (confusion error rates), complexity metrics provided important insights into model parsimony as one criterion for model selection.

3.4.2. Prediction performance metrics

Although a number of prediction performance metrics are available (also see Section 3.4.3), some have more desirable properties than others. For example, Lobo et al. (2008) cautioned that the standard AUC metric may oversimplify error analysis because AUC is usually based on binary outcomes, ignores continuous probability results, and weights errors of omission and commission equally. Indeed, ROC curves based on different thresholds of acceptable error may yield different AUC values. Also concerned for the rigid nature of standard AUC calculations, Phillips and Elith (2010) provided a new tool, the presence-only calibration plot, for calibrating presence–absence models; their approach may be useful when dealing with bias from incomplete or incorrect absence data.

Allouche et al. (2006) compared the performance of TSS, AUC, and Cohen's kappa, and concluded that TSS and AUC should be favored over kappa because kappa is more limited by its undue sensitivity to prevalence of events. However, TSS and Cohen's kappa are calculated only from a 2×2 confusion table which greatly limits their applicability.

Among 5 main prediction performance metrics used with the 36 marten case example models, covariate-weighted confusion error was highly correlated with total confusion error (Pearson $r = 0.93$, Bonferroni $P < 0.001$) not surprisingly because the model variants included many of the same covariates, whereas the other performance metrics were uncorrelated, with spherical payoff being the most different of the set (Fig. 8). In the fuller context of confusion error rates also split by age class (Table 2), BIC also was uncorrelated with the other metrics. Although these outcomes are specific

to the marten models tested and could differ with other model sets, they suggest that several metrics can provide complementary information on model performance.

3.4.3. PPD uncertainty metrics

The metrics discussed here – Bayesian credible intervals, PPCI and the certainty envelope, and the Gini coefficient – provide complementary information for evaluating uncertainty of PPDs. Among the 16 example models from the polar bear case, PPCI and the Gini coefficient were highly correlated (Fig. 9), although not linearly so, with PPCI being more sensitive at lower to middle values. Each metric exhibited a similar response behavior to specified outcome state probabilities (Fig. 10). However, PPCI may be a better general choice over the Gini coefficient for evaluating BN model outcome uncertainty because PPCI may be a more sensitive indicator (Fig. 9) and it is far easier to calculate.

3.5. Utility of alternative metrics of BN model performance and uncertainty

The various metrics discussed here collectively provide a useful set of complementary insights into BN model performance and uncertainty. Depending on the modeling objective and the analysis or decision that the model is intended to inform, alternative metrics provide utility in various ways (Table 3). The modeler needs to be aware of assumptions and caveats when applying any such metric.

Most of the model performance metrics discussed here – particularly AUC, spherical payoff, TSS, Cohen's kappa, PPCI, and the Gini coefficient – can be rescaled to range [0,100] and thus viewed as percentages of best possible model performance, with lower values denoting greater uncertainty (error rates) of model predictions. In some cases, it may then be useful to denote such rescaled values with word descriptions such as used by the Intergovernmental Panel on Climate Change to denote certainty of predictions of future climate effects (Mastrandrea et al., 2010). E.g., normalized PPCI values, rescaled as 0–100 percentages, could denote certainty levels of PPDs as follows: 0–20 = highly uncertain, >20 to 40 = moderately uncertain, >40 to 60 = about as certain as not, >60 to 80 = moderately certain, and >80 = highly certain.

One final caution is to remember that composite indices, such as PPCI and the Gini coefficient, mask individual state outcomes, so those, too, should be presented.

4. Conclusions

This summary has by no means exhaustively surveyed all possible metrics and methods useful for evaluating performance and uncertainty of BN models (e.g., Ponciano et al., 2009; Speigelhalter et al., 2002; Van Allen et al., 2008). For example, BN models may best be developed stepwise using initial expert knowledge, rigorous peer review and model reconciliation, testing model performance with empirical data, updating the probability structure with data, and retesting (Marcot, 2006). Additional methods have been developed for evaluating model parameter uncertainty and structural uncertainty (e.g., Castillo et al., 1998; Williams, 2011). A rather vast literature exists on general statistical methods and algorithms for classification analysis and modeling uncertainty (e.g., Ayyub and Klir, 2006; Hand, 1997). Also, a number of other performance metrics not discussed here are available within BN modeling programs, such as the quadratic (Brier) score (Atger, 2004), logarithmic loss and log likelihood ratio scores (Dlamini, 2010), and surprise indexes found in Netica. However, most of these other approaches are redundant with the metrics I have described in this paper.

BN models can be used for a variety of purposes including prediction, scenario analysis, diagnosis, data mining, summarizing knowledge and identifying key data gaps, and aiding individual and collaborative decision-making. No one set of metrics of model performance and uncertainty pertain to all such uses but some are more universally applicable than others. Metrics of model sensitivity are of value when constructing models based largely or solely on expert knowledge, and when exploring the underlying probability structure of models induced from empirical data. Case file simulation and influence analysis are useful for evaluating scenarios and determining the bounds of how management control on selected inputs affect expected outcomes.

Metrics of model complexity reflect parsimony and are applicable to models of all purposes. One should not use complexity metrics alone, however, without also considering model performance, because overly simple and highly parsimonious models could suffer from undue structural uncertainty (Walters, 1986) where key variables or relationships are ignored. The aim should be to reasonably balance performance with parsimony, and to neither oversimplify nor overfit models to produce the best-performing outcome for a limited data set.

Metrics of prediction performance generally pertain to situations where independent empirical data are available by which to test model accuracy. In the absence of such test data, one could apply performance metrics with historic data in the sense of hindcasting, although adjusting a prediction model to fit historic patterns is more akin to calibration than validation. One could also generate a case data set based on polling domain experts not involved in the construction of a prediction, and test model performance against these expectations, although again this may serve to calibrate a model to expert judgment rather than to validate it against real-world outcomes. Generally, the modeler could select a small set of metrics of prediction performance that provide complementary measures of model accuracy, such as use of spherical payoff, BIC, and conditional probability-weighted confusion error rate.

As for metrics of model outcome uncertainty, I suggest use of PPCI and the associated certainty envelope over the Gini coefficient which is more complicated to calculate and may be less sensitive to variations in the distribution of posterior probabilities among outcome states. PPCI can used with BN models developed from expert knowledge, empirical data, or any combination, and does not require independent data sets.

In general, the prudent modeler would do well to identify early in the model-building process the sorts of metrics to be used for evaluating model performance and uncertainty, so as to avoid post hoc selection bias. Also, metrics of performance and uncertainty can be used to help select the best model among a set of competing models in a multi-model approach (Monte, 2009). The main lesson is that rigorous adherence to methods and metrics of evaluating BN model performance and uncertainty is critical to ensuring credibility, acceptance, and appropriate application of model results, especially when used to inform decision-makers in a risk management context (Ascough et al., 2008). Prediction error and uncertainty should be viewed as information too, for identifying key parameters and advising difficult decisions of resource management and allocation.

Acknowledgments

This work derives from collaboration with a number of research teams modeling wildlife species and habitats in a variety of international locations, and reviewing use of Bayesian network models across a wide spectrum of other applications. My thanks, for such collaborations and discussions, to Steve Amstrup, Brent Boerlage, David Douglas, Tim Haas, Chad Jay, Rob McCann, Scott McNay, Brian Nyberg, Olivier Pourret, Martin Raphael, Doug Steventon, Glenn Sutherland, Randy Sulyma (dec.), George Wilhere, and

others. I thank Bret Collier, Chad Jay, Jon Pauli, and two anonymous reviewers for comments on the manuscript. Mention of commercial products does not constitute endorsement by U.S. Forest Service.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2012.01.013.

References

Aalders, I., 2008. Modeling land-use decision behavior with Bayesian belief networks. Ecology and Society 13 (1), 16, http://www.ecologyandsociety.org/vol13/iss1/art16/ [online].

Adkison, M.D., 2009. Drawbacks of complex models in frequentist and Bayesian approaches to natural-resource management. Ecological Applications 19 (1), 198–205.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary, pp. 267–281.

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43 (6), 1223–1232.

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. In: DeWeaver, E.T., Bitz, C.M., Tremblay, L.-B. (Eds.), Arctic Sea Ice Decline: Observations, Projections, Mechanisms, and Implications. Geophysical Monograph 180. American Geophysical Union, Washington, DC, pp. 213–268.

Amstrup, S.C., DeWeaver, E.T., Douglas, D.C., Marcot, B.G., Dumer, G.M., Bitz, C.M., Bailey, D.A., 2010. Greenhouse gas mitigation can reduce sea-ice loss and increase polar bear persistence. Nature 468 (7326), 955–958.

Ascough II, J.C., Maier, H.R., Ravalico, J.K., Strudley, M.W., 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecological Modelling 219 (3–4), 383–399.

Atger, F., 2004. Relative impact of model quality and ensemble deficiencies on the performance of ensemble based probabilisitic forecasts evaluated through the Brier score. Nonlinear Processes in Geophysics 11 (3), 399–409.

Atkinson, A.B., 1970. On the measurement of inequality. Journal of Economic Theory 2, 244–263.

Ayyub, B.M., Klir, G.J., 2006. Uncertainty Modeling and Analysis in Engineering and the Sciences. Chapman & Hall/CRC, Boca Raton, FL, 378 pp.

Bolstad, W.M., 2007. Introduction to Bayesian Statistics, 2nd ed. Wiley-Interscience, New York, 464 pp.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecological Modelling 157, 281–300.

Burnham, K.P., Anderson, D., 2010. Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach. Springer Verlag, New York, 514 pp.

Cain, J.D., Batchelor, C.H., Waughray, D.K.N., 1999. Belief networks: a framework for the participatory development of natural resource management strategies. Environment, Development and Sustainability 1, 123–133.

Castillo, E., Gutierrez, J.M., Hadi, A.S., 1998. Modeling probabilistic networks of discrete and continuous variables. Journal of Multivariate Analysis 64, 48–65.

Cortes, C., Mohri, M., 2005. Confidence intervals for the area under the ROC curve. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 17. Proceedings of the 2004 Conference. The MIT Press, Cambridge, Massachusetts, pp. 305–312.

Curran, J.M., 2005. An introduction to Bayesian credible intervals for sampling error in DNA profiles. Law, Probability and Risk 4, 115–126.

Dlamini, W.M., 2010. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. Environmental Modelling & Software 25 (2), 199–208.

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 116, F02009, doi:10.1029/2010JF001891.

Gutzwiller, K.J., Flather, C.H., 2011. Wetland features and landscape context predict the risk of wetland habitat loss. Ecological Applications 21, 968–982.

Hand, D., 1997. Construction and Assessment of Classification Rules. John Wiley & Sons, New York, 232 pp.

Hill, M.O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54, 427–432.

Huang, H.-C., Martinez, F., Mateu, J., Montes, F., 2007. Model comparison and selection for stationary space–time models. Computational Statistics & Data Analysis 51 (9), 4577–4596.

Jay, C.V., Marcot, B.G., Douglas, D.C., 2011. Projected status of the Pacific walrus (*Odobenus rosmarus divergens*) in the 21st century. Polar Biology 34 (7), 1065–1084.

Jensen, F.V., Nielsen, T.D., 2007. Bayesian Networks and Decision Graphs, 2nd ed. Springer, 448 pp.

Kohavi, R., Provost, F., 1998. Glossary of terms. Journal of Machine Learning 30 (2–3), 271–274.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145–151.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27 (2), 209–220.

Marcot, B.G., 2006. Characterizing species at risk I: modeling rare species under the Northwest Forest Plan. Ecology and Society 11 (2), 10, www.ecologyandsociety.org/vol11/iss2/art10/ [online].

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.

Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R., Plattner, G.-K., Yohe, G.W., Zwiers, F. W., 2010. Guidance note for lead authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties, 6–7 July 2010. Available at http://www.ipcc.ch. Jasper Ridge, CA, 5 pp.

McCann, R., Marcot, B.G., Ellis, R., 2006. Bayesian belief networks: applications in natural resource management. Canadian Journal of Forest Research 36, 3053–3062.

Monte, L., 2009. Multi-model approach and evaluation of the uncertainty of model results. Rationale and applications to predict the behaviour of contaminants in the abiotic components of the fresh water environment. Ecological Modelling 220 (12), 1469–1480.

Mouton, A.M., De Baets, B., Goethals, P.L.M., 2010. Ecological relevance of performance criteria for species distribution models. Ecological Modelling 221 (16), 1995–2002.

Neapolitan, R.E., 2003. Learning Bayesian Networks. Prentice Hall, Upper Saddle River, NJ, 674 pp.

Pauli, J.N., Whiteman, J.P., Marcot, B.G., McClean, T.M., Ben-David, M., 2011. DNA-based approach to aging martens (*Martes americana* and *M. caurina*). Journal of Mammalogy 92 (3), 500–510.

Phillips, S.J., Elith, J., 2010. POC plots: calibrating species distribution models with presence-only data. Ecology 91 (8), 2476–2484.

Ponciano, J.M., Taper, M.L., Dennis, B., Lele, S.R., 2009. Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. Ecology 90 (2), 356–362.

Pourret, O., Naïm, P., Marcot, B.G. (Eds.), 2008. Bayesian Belief Networks: A Practical Guide to Applications. Wiley, West Sussex, England, 432 pp.

Ricotta, C., Avena, G.C., 2002. On the information-theoretical meaning of Hill's parametric evenness. Acta Biotheoretica 50, 63–71.

Sadras, V., Bongiovanni, R., 2004. Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks. Field Crops Research 90 (2–3), 303–310.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Speigelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64, 583–639.

Thogmartin, W.E., 2010. Sensitivity analysis of North American bird population estimates. Ecological Modelling 221 (2), 173–177.

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

Van Allen, T., Singh, A., Greiner, R., Hooper, P., 2008. Quantifying the uncertainty of a belief net response: Bayesian error-bars for belief net inference. Artificial Intelligence 172 (4–5), 483–513.

Walters, C., 1986. Adaptive Management of Renewable Resources. MacMillan Pub. Co, New York.

Williams, B.K., 2011. Resolving structural uncertainty in natural resources management using POMDP approaches. Ecological Modelling 222 (5), 1092–1102.

Zarnetske, P.L., Edwards Jr., T.C., Moisen, G.G., 2007. Habitat classification modeling with incomplete data: pushing the habitat envelope. Ecological Applications 17 (6), 1714–1726.