Department of Computer Science
Series of Publications A
Report A-2022-13

Bayesian network modelling of complex systems with sparse data: Ecological case studies

Laura Uusitalo

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki in the main building of University of Helsinki, room U3032 on November 11th 2022 at 13 o'clock.

University of Helsinki
Finland

Supervisors

Allan Tucker, Brunel University London, UK
Laura Ruotsalainen, University of Helsinki, Finland

Pre-examiners

Ann Nicholson, Monash University, Melbourne, Victoria, Australia
Pedro Pereira Rodrigues, University of Porto, Porto, Portugal

Opponent

Simo Särkkä, Aalto University, Finland

Custos

Laura Ruotsalainen, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-8632-4 (paperback)
ISBN 978-951-51-8633-1 (PDF)
Helsinki 2022
Unigrafia

Bayesian network modelling of complex systems with sparse data: Ecological case studies

Laura Uusitalo

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
laura.uusitalo@iki.fi
http://laurau.iki.fi/

PhD Thesis, Series of Publications A, Report A-2022-13
Helsinki, October 2022, 70+84 pages
ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-8632-4 (paperback)
ISBN 978-951-51-8633-1 (PDF)

Abstract

This thesis discusses how Bayesian networks can be used to improve data analytics in the field of environmental assessment and management. The data-analytic challenge is that ecosystems are complex and potentially changing, while the available data are relatively sparse both in terms of the number of observations and in which ecosystem components they cover. This thesis takes steps towards better analysis of these sparse data through combining pre-existing, uncertain information such as modelling results and expert knowledge with modern, probabilistic data analysis. A better understanding of ecosystems' functioning and change is essential in order to manage the protection and sustainable use of ecosystems.

The thesis consists of five articles that focus on three aspects of ecosystem data analytics: prediction of the value of an ecosystem status indicator based on other observed variables, detection of systemic change in the ecosystem, and processing and presenting ecosystem projections and the linked uncertainties for decision making. For these purposes, the articles present a set of models for indicator prediction based on other ecosystem variables, a series of dynamic Bayesian networks for environmental change detection, and a Bayesian model emulator to bring to-

gether a set of different ecosystem models, assess the related uncertainties, and present their scenario simulation results in a unified framework.

Ecosystem indicators serve as simplified metrics indicating whether the specific ecosystem component is in acceptable ecological status, or whether management measures should be undertaken to improve it. However, the complexity of ecosystem interactions means that it may be difficult to assess which factors contribute to the good or poor state. This thesis explores how Bayesian network classifiers, particularly tree-augmented Naïve Bayes, can help understand which environmental factors affect the abundance of coastal fish. The approach allows maximal use of the available heterogeneous and sparse data. This work explores different discretizations of the class variable and discusses the implications of the differences between the resulting models and contributes to the still relatively quiet discussion about discretization schemes. The features are discretized to minimise information loss, and features for which no meaningful discretization is found are discarded, therefore constituting feature selection. It was found that the method can help find the relative importance of different features even with sparse data.

Detecting change in the ecological processes based on the sparse, often noisy data is challenging. This thesis builds dynamic Bayesian network models to detect change in the Central Baltic Sea ecosystem interactions, and explores the effect of different model structures and hidden variable set-ups in detecting the ecosystem change. It is shown that the hidden variables of the models are able to identify ecosystem change, and that this result does not depend on the exact model structure or hidden variable set-up.

Decision support models aim to integrate information regarding all of the interlinked aspects of the decision problem such as different parts of the ecosystems as well as economic and societal considerations. To be useful, decision support models need to be able to provide estimates of uncertainty of the different assessments and projections. This thesis reviews various methods that have been or could be applied to evaluate the uncertainty related to deterministic models' outputs, and evaluates when these methods are appropriate and what must be taken into account when applying them. Further, this thesis builds a large probabilistic meta-model to demonstrate how a Bayesian network based decision support model can be used to summarise a large body of research and model projections about potential management alternatives and climate scenarios. The data stem from multiple different models operating on different scales, and a crucial task

is to fit these results together in a manner that is faithful to the original results and maintains the consistency of the metamodel. The thesis evaluates critically the different sources of uncertainty and their interpretation in the environmental management context.

Bayesian networks are showing their strength for different tasks of environmental data analytics. Elegant handling of missing data, explicit and rigorous handling of uncertainty, and the possibility to use prior scientific knowledge and data together in analyses in a transparent way are strong advantages for Bayesian analysis for environmental data that often contain missing values and are scarce. In addition to being flexible and, thus, able to integrate different types of information and data, they are transparent, allowing critical assessment and discussion of the models. This is important as environmental data analytics are often used to support decision making on the use of ecosystems, affecting the lives of current and future generations.

Computing Reviews (2012) Categories and Subject Descriptors:

Applied computing → Physical sciences and engineering → Earth and atmospheric sciences → Environmental sciences
Computing methodologies → Artificial intelligence → Knowledge representation and reasoning → Probabilistic reasoning

General Terms:
Bayesian network, Hidden variable, Environmental modelling

Additional Key Words and Phrases:
Dynamic Bayesian network, Decision support

To Esra and Silja

Acknowledgements

The idea of this thesis (my second PhD, the first being in fisheries science in 2007) started rising its head after I had finished my computer science MSc in 2017. I was able to push the idea aside for a few years as impractical and useless, but finally had to give up. And indeed, I have been happy to learn more about data analysis and computational methods, even though my skills and knowledge always feel insufficient in the face of the vast ocean of everything I would like to learn.

Allan Tucker, my supervisor, introduced me to dynamic Bayesian networks and has been kind and generous with his time and expertise, as well as so much fun to work and hang around with. Thank you for everything! Laura Ruotsalainen kindly agreed to be another supervisor for this work, and has patiently provided constructive feedback to this summary in particular. Thank you! It's been a pleasure to work with you, and you've been nothing but supportive and solution-oriented all through the process. Maiju Lehtiniemi and Teemu Roos served in my advisory committee and offered support and encouragement. I really appreciate it, thank you!

Professors Ann Nicholson and Pedro Pereira Rodrigues kindly examined this work. I am thankful for such brilliant scientists to give their time to evaluate my work. Prof Nicholson provided excellent comments that helped improve the summary. Thank you!

My heartfelt thanks go out to the co-authors of the papers included in this thesis. This interdisciplinary work would not have been possible without you. I keep on being humbled by the depth and breadth of knowledge and skills of people I get to work with; it's truly an honor.

Arto Wikla kindly offered indispensable help multiple times during my CS MSc when I found it difficult to navigate the department bureaucracy. Pirjo Moen has been kind and supportive all through the PhD process and answered

all kinds of stupid and tricky questions with patience, as well as offered solutions even before I've had the mind to ask. Thank you both!

The articles in this thesis were written as part of collaborative projects BONUS BLUEWEBS and BONUS FUMARI, funded by the Academy of Finland and EU BONUS (Art 185) programme, and TALENTS project at the Finnish Environment Institute.

My closest and dearest have been there for me and supported me as I've chosen a path that has not always been the easiest (and perhaps also not the wisest). Edu, thank you for your unfaltering love and support, as well as for all the times you've helped me, listened to my musings, and offered a word of wisdom and balance. Jari, thank you for the love and warmth and shared geekery, and for believing in me. Esra and Silja, keep on being awesome. I hope that you too will find a path that nurtures you.

Helsinki, October 2022
Laura Uusitalo

Contents

Original papers **xiii**

1 Introduction **1**
1.1 Background and motivation 1
1.2 Thesis contributions 5

2 Overview of Bayesian networks **9**
2.1 Basics of Bayesian networks 9
2.2 Learning Bayesian networks 14
2.3 Hidden variables and dynamic BNs 16
2.4 Decision networks 19
2.5 Spatial and spatiotemporal BNs 20
2.6 When to use BNs? 21

3 Building the Bayesian networks in this thesis **23**
3.1 Identifying the variables 24
3.2 Identifying and discretizing the variable values 24
3.3 Finding the graph structure and parameters 28
3.4 Decisions and utilities 31

4 Probabilistic prediction of indicator value **33**

5 Dynamic Bayesian networks for detecting ecosystem change **37**

6 Environmental decision support **47**

7 Conclusions **53**

References **57**

Original papers

This dissertation consists of five published articles, printed at the end of the thesis:

I Lehikoinen, A., Olsson, J., Bergström, L., Bergström, U., Bryhn, A., Fredriksson, R., Uusitalo, L. 2019. Evaluating complex relationships between ecological indicators and environmental factors in the Baltic Sea: A machine learning approach. Ecological Indicators 101: 117–125. doi: 10.1016/j.ecolind.2018.12.053

II Uusitalo, L. Tomczak, M.T., Müller-Karulis, B., Putnis, I., Trifonova, N., Tucker, A. 2018. Hidden variables in a Dynamic Bayesian Network identify ecosystem level change. Ecological Informatics 45: 9–15. doi: 10.1016/j.ecoinf.2018.03.003

III Maldonado, A.D., Uusitalo, L., Tucker, A., Blenckner, T., Aguilera, P.A., Salmerón, A. 2019. Prediction of a complex system with few data: Evaluation of the effect of model structure and amount of data with dynamic bayesian network models. Environmental Modelling & Software 118: 281–297. doi: 10.1016/j.envsoft.2019.04 .011.

IV Uusitalo, L., A. Lehikoinen, I. Helle, and K. Myrberg. 2015. An overview of methods to evaluate uncertainty of deterministic models in decision support. Environmental Modelling & Software 63:24–31. doi: 10.1016/j.envsoft.2014.09.017

V Uusitalo, L., Blenckner, T., Puntila-Dodd, R., Skyttä, A., Jernberg, S., Voss, R., Müller-Karulis, B, Tomczak, M.T., Möllmann, C., Peltonen, H. 2022. Integrating diverse model results into decision support for good environmental status and blue growth. Science of the Total Environment 806 (Part 2): 150450. doi: 10.1016/j.scitotenv.2021.150450

- In **Paper I**, I had the methodological idea and supervised the modelling and data analysis together with Dr Olsson, as well as contributed to the writing.
- In **Paper II**, I had the idea, did the modelling and analysis work, and was responsible for the writing.
- In **Paper III**, I had the idea, supervised the work together with Prof. Salmerón, and contributed to the writing.
- In **Paper IV**, I had the idea and the main responsibility for the review and writing.
- In **Paper V**, I had the idea, contributed to the modelling decisions, analysed the results, and was responsible for the writing.

Chapter 1

Introduction

This thesis falls into the category of data science, in the intersection of statistics, computer science, and domain knowledge, in this case, environmental science. Data science is, by nature, interdisciplinary, but environmental data science has not been intensely developed so far (Blair et al. 2019). This thesis consists of five articles that explore the use of Bayesian networks as a method to better understand the marine ecosystem. The data-analytic challenge is that the systems are complex and potentially changing, while the available data are relatively sparse. This thesis takes steps towards better analysis of these sparse data through combining pre-existing, uncertain information such as modelling results and expert knowledge with modern, probabilistic data analysis.

1.1 Background and motivation

As the human footprint on the world increases, protection of ecosystems is crucial for the continued existence and wellbeing of humanity. In order to protect the ecosystems while still being able to enjoy the benefits they give us, such as food, raw materials, and recreation, we need to understand their functioning. Ecosystems are, however, complex systems that include a high number of potentially interacting components such as habitats, species, and different age and size groups behaving in different ways. Temporal and spatial scales that are relevant for these components vary from days to decades and from centimetres to hundreds of kilometres, and interactions between these components also take place on multiple spatial and temporal scales. This complexity is a challenge to data science (Blair et al. 2019).

Anthropogenic pressures drive ecosystems towards new states, such as different species abundances, distributions, and predation patterns (Nelson et al. 2006). This change may be driven by changes in biogeochemical factors such as temperature, salinity, pH changes, or nutrient availability (Capuzzo et al. 2018; Jüssi et al. 2008; Mäkinen et al. 2017; Suikkanen et al. 2013), changes in species biomass due to fishing and hunting (Daskalov et al. 2007; Eero et al. 2011; Harding and Härkönen 1999), introductions of new species that may change the system dynamics (David et al. 2017; Jormalainen et al. 2016; Katsanevakis et al. 2014; Norkko et al. 2012), or any combination of these factors (Blenckner et al. 2015; Griffith et al. 2011; Halpern et al. 2015; Möllmann et al. 2008). These changes may be gradual (Duarte et al. 2009; Hillebrand et al. 2020) or relatively abrupt (Alheit et al. 2005; Beaugrand 2004; Conversi et al. 2010), and they may be difficult to discern from the natural variability of the system. The possibility of these changes adds a layer of challenge on top of the already challenging analysis of the complex system, as they may hinder the knowledge discovery and generalisation from past data (Sáez et al. 2015). The data analyst must be vigilant of potential changes in the system dynamics while not overreacting to data variability and outliers.

Environmental data comes from multiple sources, such as autonomous and semi-autonomous measurement instruments, satellite images, and field sampling campaigns, as well as citizen science observations, historical records, and e.g. social media data mining (e.g. Blair et al. 2019; Lehtiniemi et al. 2020; Mack et al. 2020). While modern technologies offer increasing opportunities for spatially and temporally intensive monitoring (Mack et al. 2020), understanding many ecosystem processes still requires data obtained through field sampling. In temperate climates, the seasonal cycle has a strong effect on the functioning of the ecosystems, imposing a yearly cycle of growth and reproduction on many species, and making the year a natural time step for many ecological observations. This means that even impressive long-term datasets may have only 30–50 observations, and even with more frequent observations, the yearly cycle needs to be taken into account when analysing the data.

Complications in the study of longitudinal data arise also from the fact that the data collection process has only rarely stayed consistent across all variables all through the time series, but there are changes in the timing, location, or gear of the sampling, and not all variables are available all through the time series; some perhaps only for one year or every 3rd, 5th, or 10th year. Observations may be missing due to equipment failures, weather conditions, etc. There may be differences in the timing, location, and accuracy of the data of different ecosys-

tem components, making it challenging to combine them in data analysis in a straightforward manner.

Therefore, even high-quality datasets may be small in terms of number of observations, and the data may be sparse due to missing observations. Additionally, any environmental monitoring process only includes a part of the ecosystem components due to historical, economic, epistemic, and other reasons. For example, in the central Baltic Sea, fish stock data exists for the key commercially exploited species herring (*Clupea harengus membras*), sprat (*Sprattus sprattus*), and cod (*Gadus morhua*), but a central species in the food web, the three-spined stickleback (*Gasterosteus aculeatus*), remains largely unobserved. Therefore, the sparseness of data is also of a conceptual nature: some variables that are known or suspected to be important for the system are missing from the data altogether. These features make ecological datasets challenging for many modern data analysis methods such as many machine learning applications that are often relatively data-hungry.

Ecological data are traditionally analysed through multivariate statistical analyses and quantitative ecosystem models (Hilborn and Mangel 1997; Zuur et al. 2007). Multivariate statistical analyses often require full datasets and cannot cope with missing data, leading to the need to either impute the missing values and/or leave out variables with multiple missing values (James et al. 2013). Leaving out data that domain experts consider relevant, in a situation where data are sparse, is regrettable. Imputation is often preferable, but not without problems.

Ecosystem models are based on a conceptual model of interactions between ecosystem elements and mathematical expressions that describe the relationships between them (Jackson et al. 2000). These models are usually built and parameterized through a combination of theoretical considerations and data analysis. Often the forms of the functions - i.e. whether the response is linear, sigmoid, or something else - are decided based on theory and/or modelling restrictions, and the parameters of the functions - i.e. the slope of the line etc. - are computed from the data. These models, if suitably set up, can capture ecosystem parameters such as relative biomasses, spatial distributions, consumption patterns, etc. Using ecological knowledge to guide the data analysis by restricting the interactions between the variables (as model structures customarily do) helps fight the curse of dimensionality that might cripple a simple data-driven analysis.

When models are parameterised using data, it is usually assumed that the underlying relationships between the ecosystem components, described as mathematical functions between model variables, are unchanging through the time

series, i.e. arise from a stationary distribution (Gama et al. (2004) note the same with machine learning models). These functions are estimated using the whole time series, assuming that all the data points are samples from the same function. This assumption might be untrue, as ecosystems are known to sometimes undergo fast structural changes that have a major effect on the ecosystem dynamics (Alheit et al. 2005; Beaugrand 2004; Conversi et al. 2010), and pressures such as climate change, introductions of alien species, and exploitation and land use changes may also change the shape of the ecological interactions (Bulleri et al. 2020). From the modelling and data analysis perspective, these changes pose a challenge, since the same functional forms may not describe the relationships between the variables before and after the change (Blenckner et al. 2015).

Ecological models are often fraught with uncertainty stemming from both epistemic (lack of knowledge) and aleatory (natural variation and randomness) sources (Regan et al. 2002), whether this uncertainty is included into the model or not. Assessment of uncertainty is crucial for any management problem that requires balancing of different goals, such as nature conservation versus serving the needs of humanity. Therefore, a methodology that provides uncertainty estimates is needed. Bayesian methods, which are capable of providing estimates of uncertainty and incorporating both aleatory and, to some degree, epistemic uncertainty, are increasing their popularity in ecology (Anderson et al. 2021). The parameter estimates of Bayesian models are probability distributions, allowing not only the comparison of the expected (mean), "average" (median) or most likely (mode) values, but also for example the probability that the value is smaller or larger than some critical value. As an example, a fisheries manager choosing the best fishing strategy is interested in the expected catches, but also wants to avoid population collapse, i.e. minimise the probability that the population size will shrink below a threshold value. These estimates can be obtained directly from Bayesian model results, making these models particularly useful for aiding practical ecosystem management decisions.

Identifying changes in ecosystem interactions is a major challenge even for large data sets, as the natural variation in many observed variables is often high, making it difficult to separate signal from noise (Fulton et al. 2003). Further, it is possible that the changes are driven by variables that are not being observed and therefore are missing from the data. Dynamic Bayesian networks (DBN) (Dean and Kanazawa 1989; Kjærulff 1992; Murphy 2002; Nicholson 1992), including those with hidden (latent) variables, avoid the implicit assumption of system stability, and explicitly capture the systemic change of the modelled systems (Tucker and Liu 2004). Hidden variables in these models represent quantities of

interest that cannot be observed directly, or are unobserved for some reason. In a dynamic model, hidden variables enable modelling of non-stationary dynamics, and the detection of changes in the functional forms of the model. However, this possibility is only emerging in ecological analyses (Trifonova et al. 2019; Trifonova et al. 2015; Trifonova et al. 2017; Trifonova et al. 2021, this thesis). The methodology enables the detection of ecosystem changes even when there are no data on the components causing the change, such as new species invasions, as the latent variables can be set up to detect changes in different parts of the system. Furthermore, Bayesian methods are perceived as highly useful in ecological research due to their ability to rigorously integrate prior knowledge with data, reflected in the recent strong rise of Bayesian methods in ecology (Anderson et al. 2021).

The combination of modern data analytics and knowledge accumulated through decades of research offers a promising avenue to analyse complex ecosystem processes through the small and sparse data sets that are available. While BNs are increasingly applied to ecological studies, DBNs particularly are relatively few (Chee et al. 2016; Rachid et al. 2021), and these applications only rarely include hidden variables. Therefore, their capabilities and limits to help understand ecological change are still underexplored (however, see e.g. Harwood 2020; Jiang et al. 2021; Wu et al. 2018) and poorly understood.

1.2 Thesis contributions

This thesis contributes to the field of environmental data science by exploring opportunities and limitations on Bayesian networks in environmental modelling and prediction tasks.

Paper I explores Bayesian network classifiers, particularly the use of tree-augmented Naïve Bayes, in understanding which environmental factors affect the abundance of coastal fish. This approach allows maximal use of the heterogeneous and sparse data that were available. The paper explores different discretizations of the class variable and discusses the implications of the differences between the resulting models and contributes to the still relatively quiet discussion about discretization schemes. The features are discretized to minimise information loss, and features for which no meaningful discretization is found are discarded, thereby constituting feature selection. It was found that the method can help find the relative importance of different features even with sparse data.

Paper II builds dynamic Bayesian network models to detect change in the Central Baltic Sea ecosystem interactions, and explores the effect of different hidden variable set-ups in detecting the ecosystem change. Change detection is difficult as the data are sparse both in terms of ecosystem components and spatial and temporal resolution, and there is noise in the data. It is shown that the hidden variables of the models are able to identify ecosystem change, and that this result does not depend on the exact hidden variable set-up.

Paper III expands the analyses of Paper II and builds and evaluates additional models to evaluate the effect of the model structure for the detection of ecosystem change. Paper III evaluates the effect of the amount of data and the model structure on the predictive accuracy of the models through using only parts of the data and evaluating the prediction results. Fully Bayesian and Maximum likelihood estimations were compared for the different models. More data improves the predictions, while the different setups of hidden variables did not make a critical difference.

Paper IV addresses the problem that while environmental management needs uncertainty estimates and probabilistic assessments, many ecological models are not able to provide them. Decision support models, aiming to integrate information regarding all of the interlinked aspects of the decision problem, need to be able to provide uncertainty estimates. Paper IV reviews various methods that have been or could be applied to evaluate the uncertainty related to deterministic models' outputs, and evaluates when these methods are appropriate and what must be taken into account when applying them. A combination of domain-related considerations and aspects related to the source models and their data are central in this consideration.

Paper V builds a large probabilistic meta-model based on different modelling results to demonstrate how a Bayesian network based decision support model can be used to summarise a large body of research and model projections about potential management alternatives and climate scenarios. As the data stem from multiple different models operating on different scales, a crucial task is to fit these results together in a manner that is faithful to the original results and maintains the consistency of the metamodel. Paper V evaluates critically the different sources of uncertainty and their interpretation in the environmental management context.

The rest of this summary is organised as follows: Chapter 2 gives background to Bayesian networks in general and particularly in relation to this work. Chapter 3 discusses the process of building Bayesian networks and situates the models of this thesis in that process. Chapter 4 discusses probabilistic prediction of indica-

tor value (topic of Paper I), Chapter 5 discusses dynamic Bayesian network (topic of Papers II and III), and Chapter 6 discusses environmental decision support through BNs (topic of Papers IV and V). Chapter 7 draws final conclusions.

Chapter 2

Overview of Bayesian networks

This chapter gives an overview of Bayesian networks both generally and particularly in relation to the papers that form this thesis. Key references in this section include textbooks by Jensen (2001), Korb and Nicholson (2011) and Murphy (2012).

2.1 Basics of Bayesian networks

Bayesian networks (BN) are a type of graphical models that efficiently encode the joint probability distribution over a domain. They are fully probabilistic models that consist of a qualitative part, the model structure, and a quantitative part, the quantitative connections between the variables (Pearl 1986). The qualitative part is a directed acyclic graph (dag), i.e. the dependencies between the variables are encoded through directed arcs (Figure 2.1).

The absence of an arc implies the lack of direct connection, however, the variables can be informative about each other through the other variables in the network. The arcs may not form a directed loop, i.e. such structures are not allowed where you would return to the starting node by following the arcs in their direction. If there is an arc pointing from node (variable) A to node B, A is called the parent of B, and B the child of A; similarly, children of children are sometimes called descendants. There are no theoretical restrictions regarding the number of parents or children a variable can have.

The quantitative part is encoded as conditional probability distributions of the variables, given their parent variables. If the variable has no parents, it has a single marginal probability distribution. The probability distributions of BNs

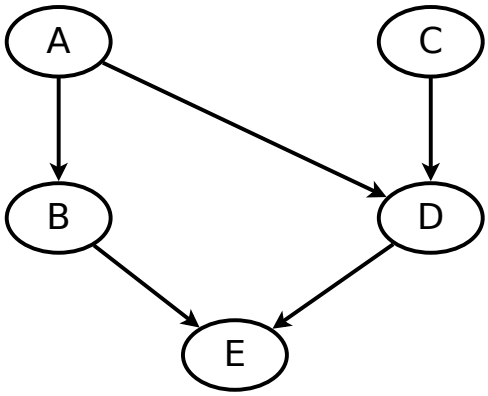


Figure 2.1: A simple Bayesian network. A and C are both parents of D; E is a descendant of all of the other variables.

can be either continuous or discrete. However, solving the networks, computing the posterior distributions given some evidence, requires different algorithms depending on whether the distributions are continuous or discrete. The state of the art of solving continuous-distribution models (e.g. Lye et al. 2022; Margossian et al. 2020) is to run Markov chain Monte Carlo simulations and sample the Markov chains for approximations of the posterior distribution (Gilks et al. 1995); discrete models can be solved analytically using Bayes rule (Equation 2.1):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

where
A, B = events,
$P(A), P(B)$ = probabilities of events A and B,
$P(A|B)$ = probability of A given that B is true,
$P(B|A)$ = probability of B given that A is true.

The Bayes rule can also be used to calculate marginal distributions, i.e. the probability distributions of variables without a reference to other variables. Marginal distribution is the distribution of the variable, when everything that is known of the other parts of the model is taken into account.

In the example of Figure 2.2, the probability table of the *Rain* variable already gives the marginal distribution, $P(Yes) = 0.2$, $P(No) = 0.8$. However, to get

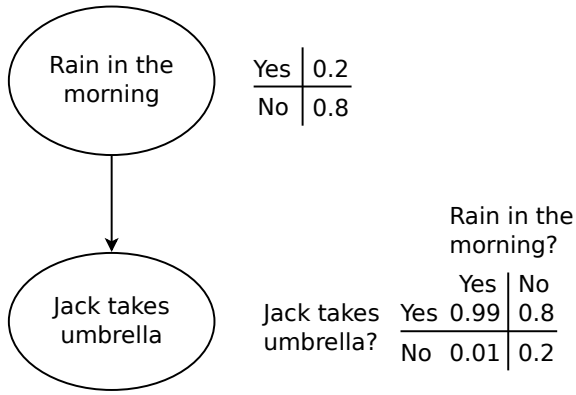


Figure 2.2: A simple example of conditional probability distribution. The variable *"Jack takes an umbrella"* is conditionally dependent on the variable *"Rain in the morning"*, meaning that its probability distribution varies as a function of the *Rain* variable.

the marginal distribution of the *Umbrella* variable, we need to marginalise the *Rain* variable out by summing over the conditional probabilities of each state multiplied by the probabilities of the conditions:

$$\begin{aligned} P(Umbrella = Yes) &= P(Umbrella = Yes|Rain = Yes) * P(Rain = Yes) + \\ &\quad P(Umbrella = Yes|Rain = No) * P(Rain = No) \\ &= 0.99 * 0.2 + 0.8 * 0.8 = 0.838 \end{aligned}$$

$$\begin{aligned} P(Umbrella = No) &= P(Umbrella = No|Rain = Yes) * P(Rain = Yes) + \\ &\quad P(Umbrella = No|Rain = No) * P(Rain = No) \\ &= 0.01 * 0.2 + 0.2 * 0.8 = 0.162 \end{aligned}$$

The joint distribution of the variables in the model is represented in a factorized form as (Scutari and Strimmer 2011):

$$P(X) = \prod_{i=1}^{n} P(X_i|parents(X_i)), \text{ for discrete variables,} \tag{2.2}$$

$$f(X) = \prod_{i=1}^{n} f(X_i|parents(X_i)), \text{ for continuous variables.} \tag{2.3}$$

For the example of Figure 2.2, this would be

$$P(Rain, Umbrella) = P(Rain)P(Umbrella|Rain)$$

and for the model in Figure 2.1

$$P(A, B, C, D, E) = P(A)P(C)P(B|A)P(D|A, C)P(E|B, D)$$

Bayesian networks can be used for encoding expert understanding over a domain (Pearl 1986; Uusitalo et al. 2005), encoding high-dimensional data into a more compact form (Barber 2012), as well as combining pre-existing knowledge (such as expert knowledge) and new data in a mathematically transparent way (Korb and Nicholson 2011). Bayesian networks are not, by definition, causal models; they quite simply encode the joint probability distribution over the problem domain. However, if the BN model structure is built to reflect the causalities in the modelled domain, this has some benefits, leading many authors such as Korb and Nicholson (2011) to strongly favour the causal interpretation of the models. It can also be shown that among Markov-equivalent models, i.e. those that encode the probability distribution similarly, the causal model is the simplest, i.e. the one including the least arcs (Korb and Nicholson 2011).

The model structure defines which variables can interact with each other in the network, given different observations (evidence). The parents, children, and the children's other parents of variable A form the so-called Markov's blanket of variable A; if the states of all of these variables are known without uncertainty, no changes elsewhere in the model can give any further information about the state of variable A (and vice versa; any improved information about variable A's status will not update the information about any of the other variables in the model). The circumstances under which information about variable X can affect the state of variable Y can be inferred through d-separation (directed separation) properties of the model structure. Therefore, the model structure plays an important role in the model inference, and the modeller needs to check carefully that the d-separation properties of the model make sense in the substance interpretation of the model.

D-separation can be analysed through understanding the three possible types of connections between three variables: serial, diverging, and converging (Figure 2.3, Jensen 2001). Serial connection (Figure 2.3a), is straightforward: if the value of A is updated, this will update the distribution of B, which, in turn, will update the distribution of C. Similarly, new information on C will update the distribution of B, and that, again, the distribution of A (following Bayes rule,

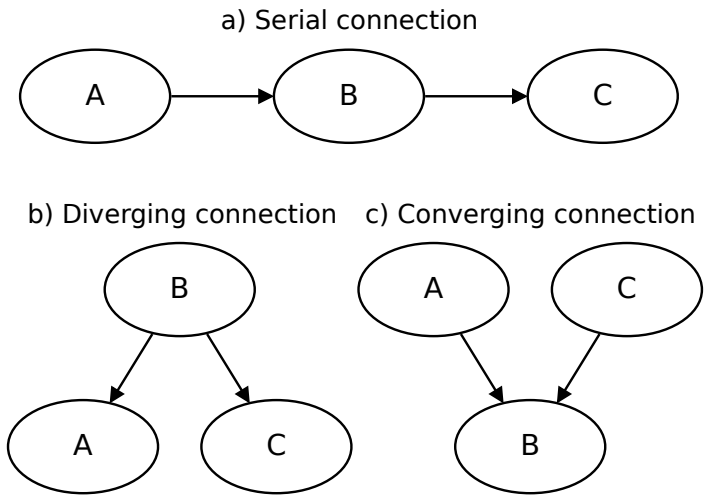


Figure 2.3: Serial, diverging, and converging connection types.

Equation 2.1). However, if the value of B is known exactly (it is *instantiated*), any updated information about A will not change it. Therefore, as B does not change, the distribution of C will not change either. In this case, *A and C are d-separated given B.*

In a diverging connection (Figure 2.3b), information flows through the parent variable (B) between the children (A and C), if the parent variable is not instantiated. However, if the parent variable is instantiated, this blocks the information flow between the children. In a diverging connection, the children are d-separated given the parent.

In a converging connection (Figure 2.3c), if there is no information about the common child B, the parents are d-separated; updated knowledge about one parent does not give any additional information about the other parents. However, if we get updated information about the connection node B, this opens the communication channel between the parents. The channel is opened both if B is instantiated, or if its distribution is updated because one of its descendants is instantiated and its distribution is therefore updated. This is easiest understood intuitively if the BN is viewed as describing causal connections. The opening of the link is called the *explaining away* effect if: we have information about a consequence, it might just as well be caused by any of the alternative causes (whose probability distributions will be updated accordingly). However, if we

know that one of the causes has taken place, the probabilities of the other reasons will decrease; or if we get information that one of the causes has not happened, this will increase the probability of the other causes. However, if we do not have any information about the common consequence (or its consequences), there is no reason that information about the occurrence of one parent variable should affect the probabilities of the others.

Note that blocking the information flow, in the case of serial and diverging connection, requires that the value of variable B is known without uncertainty (sometimes called *hard evidence*); however, just updating the probability of variable B (sometimes called *soft evidence*) is enough to open the communication channel in converging connection (Jensen 2001).

2.2 Learning Bayesian networks

Machine learning with BNs includes both structure learning and parameter learning from data. Both of these problem domains are NP hard, and practical algorithms include approximations and assumptions. For the structural learning problem, multiple algorithms exist with somewhat different starting points. The number of dags increases exponentially as the number of variables increases, making the evaluation of all possible dags an intractable problem. Some of the algorithms, therefore, base the structure search on the evaluation of statistical correlations between the variables, while others rely on user supervision to restrict the search space. For parameter learning, the iterative expectation-maximisation (EM) algorithm (Dempster et al. 1977; Lauritzen 1995) is the industry standard (Algorithm 1) when there are missing data. The algorithm iterates the expectation E-step and the maximisation M-step: starting from some initial values for the parameters, the E-step computes the probability distribution or expected sufficient statistics (i.e. the figures needed to describe the distributions) for a complete data, taking into account the parameter values and the observed data. M-step uses these values to compute new parameter values, maximising the log-likelihood of the parameters using either maximum likelihood or maximum a posteriori probability (Barber 2012; Kjaerulff and Madsen 2008; Korb and Nicholson 2011). These steps are iterated until a stopping criterion is satisfied, e .g. until the difference between consecutive log-likelihoods is small enough.

However, the expectation step is usually implemented by computing a sufficient statistic for the missing data e^* instead of the full probability distribution (Korb and Nicholson 2011). In that case, Step 1 of Algorithm 1 becomes compu-

Algorithm 1 The EM algorithm's general form (Korb and Nicholson 2011).

0: Initialization step: Set $\hat{\theta}$ to an arbitrary, legal value. Select the desired precision ε of $\hat{\theta}$. Set value $\hat{\theta}'$ to an illegally large value.
while $|\hat{\theta}' - \hat{\theta}| > \varepsilon$: **do**
$\hat{\theta} \leftarrow \hat{\theta}'$, (except on first iteration)
1: Expectation step:
Compute the probability distribution over missing values

$$P(e^*|e, \hat{\theta}) = \frac{P(e|e^*, \hat{\theta})P(e^*|\hat{\theta})}{\sum_{e^*} P(e|e^*, \hat{\theta})P(e^*|\hat{\theta})}$$

2: Maximisation step:
Compute the new ML or MAP estimate $\hat{\theta}'$ given $P(e^*|e, \hat{\theta})$
end while

tation of the expected counts of joint instantiations of X_i and $Parents(X_i)$ for all variables X_i and their states. In Step 2, the new ML or MAP estimate of the parameters is computed.

The EM algorithm can deal with missing data and usually converges quickly, but can get stuck to local maxima (Barber 2012). Variants of the EM algorithm include penalised EM algorithms for combining expert knowledge with data in the parameter estimation (Kjaerulff and Madsen 2008) and extensions to increase the efficiency of the algorithm through simplifying assumptions (Barber 2012).

Bayesian networks can also be constructed based on expert knowledge only, in a process called expert elicitation. This is useful particularly if data are sparse or extensive knowledge on the research topic exists beyond the available data set, for example if ecological interactions can be generalized from studies from other scales or areas. While expert elicitation of both model structure and parameters is a common practise in environmental modelling (Barton et al. 2012; Chen and Pollino 2012; Kuikka and Varis 1997; Uusitalo et al. 2005), there is little literature on best practises of structure elicitation. Kjaerulff and Madsen (2008) propose identifying the set of relevant variables and their types, then proceeding to find the causal ordering of these variables. However Korb and Nicholson (2011) note that deciding on the model structure is always a tradeoff between model fidelity (often increasing arcs and nodes) and the ease of building and using the model. They note that causation is an important starting point for the

structure building, as causal models are simpler and more compact. They also propose asking direct questions about causation, prevention and interference to determine the d-separation structures of the model. Boneh et al. (2006) present a visual model that helps experts evaluate whether their model structure represents their knowledge correctly.

Chen and Pollino (2012) propose starting the model building with a conceptual model which can then be modified to fit the purpose of the modelling. They also note that all included variables must be relevant in the scale of the model and affect of be affected by the interest variable, and that attention should be paid to how the length of the impact path in the model may affect the sensitivity of the output to the inputs.

More literature exists on expert elicitation of parameters (conditional probabilities), discussing different methods for eliciting probabilistic assessments, both individually and in groups (Hanea et al. 2018; Hanea et al. 2017; Hemming et al. 2018; Morgan and Henrion 1990; O'Hagan et al. 2006). The methods aim to arrive at an unbiased estimate of the expert's subjective probabilities, or the expert group's joint or averaged probabilities.

2.3 Hidden variables and dynamic BNs

The EM algorithm can perform its task even if some variables are not observed at all, i.e. variables that may be important for the model structure but for which there are no observations. Such variables are called hidden or latent variables. If there is a hypothesis about which real-life, unobserved variable or process the hidden variable stands in for, the hidden variable can be linked to the relevant observed variables (Barber 2012; Trifonova et al. 2015), or, to ensure that they are relevant and not redundant, we can set them as parents to all of the observed variables (Friedman 1997). This allows us to see whether they pick up patterns in temporal data (Trifonova et al. 2015; Trifonova et al. 2021). Inclusion of hidden variables may allow building models that are closer to the "true" natural process e.g. through making the d-separation properties of the model more realistic (Kwoh and Gillies 1996; Pearl 2000). Hidden variables can also serve to simplify the model e.g. through introducing a measurement error factor that will naturally account for outliers, thus simplifying the relationship between the observed variables (Murphy 2002).

A famous example of hidden variables are the hidden Markov models (HMM, Figure 2.4) which assume that there is an unobserved (hidden) Markov process,

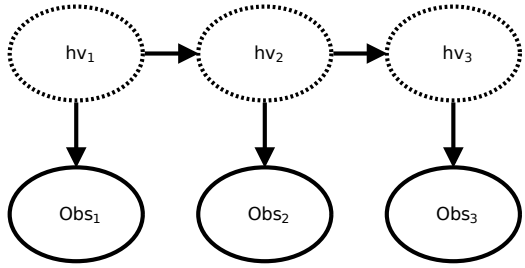


Figure 2.4: Hidden Markov model rolled out for three time slices. The dashed *hv* variables represent the hidden Markov process; the Obs variables with the solid border are the observed variables.

and on each step, we can observe a variable that depends on the value of the hidden variable at that time step. The simplest first-order HMM needs just two conditional probability tables (CPT): the transition probability between the states of the hidden variable (the Markov process) and the conditional probability of the observed variable given the value of the hidden variable.

Dynamic Bayesian networks (DBN) are BNs that are rolled out to represent multiple time slices of the same system, with autocorrelative and/or other links between the time slices (Figure 2.5). Their parameters may also be learnt from data so that they can be used to predict e.g. missing data on some time slices or the future. A combination of expert judgement based model structure and data-based parameters (with or without expert judgement based priors) allows the use of expert knowledge along with data in the analyses. This approach enables drawing from the vast body of experimental and theoretical work on ecological interactions, while allowing the local data to dictate the specifics of the model and adapt to the local, specific conditions. The model variables are the species abundances and environmental variables such as temperature, and the arcs in the models represent assumed causal connections, such as predation, between the variables.

For DBNs, the joint probability becomes

$$P(X^t|X^{t-1}) = \prod_{i=1}^{n} P(X_i^t|Parents(X_i^t))) \tag{2.4}$$

where $Parents(X_i^t)$ can be in the current or one of the previous time slices. For example, in Figure 2.5, $Parents(B_2)$ are B_1 and A_2.

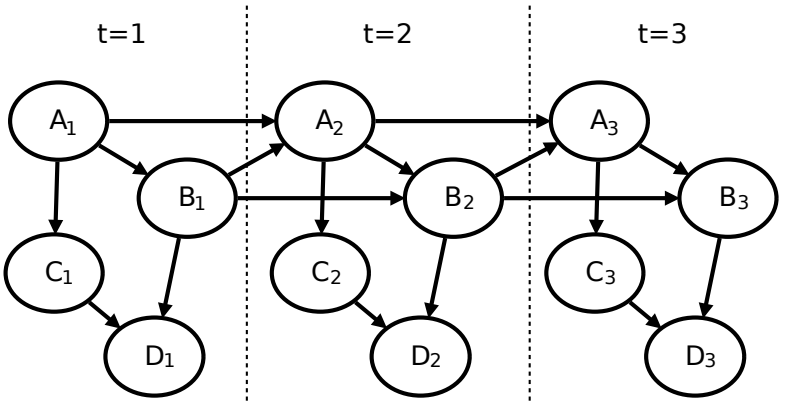


Figure 2.5: A simple DBN model rolled out for three time steps.

Dynamic Bayesian networks with hidden variables can be thought of as extensions of HMM and Kalman Filter models (KFM) (Murphy 2002). They are discrete-time state-space models with the possibility of having both continuous and discrete variables. Unlike an HMM, DBN can present the state space in a factored form, using multiple variables, instead of a single random variable, and unlike KFM, they can represent arbitrary probability distribution shapes. In other words, DBNs relate variables to each other over time steps, usually by developing a model for one time slice that includes all the relevant variables, and then defining the transition probabilities between consecutive time steps much like in HMM models. However, in DBNs there can be multiple, and more complex, links also across time steps.

In ecological modelling, DBNs are particularly useful because they provide an escape from a major restriction in stationary BNs, namely their lack of ability to include direct feedback loops due to the acyclicity requirement (Uusitalo 2007). Many ecological processes, e.g. population dynamics of plants and animals, include temporal dynamics that would, in stationary models, emerge as directed cycles. Dynamic models allow more realistic modelling of these processes.

When the hidden variable, which has no data, is linked to a model that is taught with data, the hidden variable value accommodates to the observed data in the model. If it were linked to only one variable of the model, which had also other parents, the hidden variable would be parameterized to fit to the variability or noise in the CPT. If the hidden variable is set as a parent to all or a selected set of the variables in the model, it would also try to explain the residual in all

of the CPTs. This means that we would expect random variation, "white noise", in the value of the hidden variable. However, if there is a simultaneous shift in multiple parameters of the model, this will be reflected in the hidden variable as an observable change beyond the error margins.

This behaviour of the hidden variable means that we can fit the model with the observed data, and detect possible changes in the ecological interactions in time through the hidden variable. Multiple hidden variables can be set up to observe different parts of the model to pinpoint the changes more accurately.

Dynamic Bayesian networks with hidden variables have been proposed as a potential method for early detection of structural changes in the ecosystem. Trifonova et al. (2015) fitted different models to the food web data of the North Sea and concluded that DBNs with hidden variables were able to reflect the food web dynamics, while the hidden variables captured the ecosystem dynamics. These models include hidden (latent) variables that enable modelling of unobserved variables, as well as non-stationary dynamics (Ceccon et al. 2011; Robinson and Hartemink 2009; Tucker and Liu 2004), which will allow the detection of previously unseen events and changes in the underlying relationships. While these approaches have been used in data-rich fields such as medicine and finance (Chandola et al. 2009), they are novel in ecological analyses.

2.4 Decision networks

Bayesian networks can be used to aid decision-making in situations involving uncertainty and different decision alternatives through augmenting the BN with an explicit representation of decisions and utilities (Howard 1965; Howard and Matheson 1981; Jensen 2001; Kjaerulff and Madsen 2008; Korb and Nicholson 2011; Madsen et al. 2005). These models are called influence diagrams or decision networks. They make it possible to determine the expected utility related to each decision option, and in consequence, to determine the optimal set of decisions to maximise the expected utility.

Decision variables encode the set of decision options available to or considered by the decision-maker. They are not stochastic, but the decision-maker has full control over which decision will be taken. Therefore, it is not usually relevant to define a probability distribution over the decisions or to evaluate the model under the assumption that any of the decisions can be taken under certain probabilities. The decision options also do not need to cover all possible decisions that could be taken (for example, all possible levels of nutrient load reduction or fishing effort),

only those that are considered in the modelled case. Because of this, decision variables are also useful in encoding scenarios, for example, different climate change projections or societal pathways. Use of decision variables in these cases implies that the variable does not exhaustively cover all possible states of the variable, and that the model results should not be averaged over them.

Utility nodes include a function that maps the values of their parent nodes to a real value that can be either positive or negative. Negative values usually denote incurred costs or undesirable states of the parent variable (for example, the extinction of a species could be linked to a high negative utility), while positive values denote desirable states (profits, improved environmental status, etc.). The expected utilities of the decision alternatives can be computed, given any available evidence, based on the probability of all possible outcomes given the actions and evidence. Kaikkonen et al. (2020) note that sometimes the highest expected utility is associated with the highest uncertainty regarding the outcome, or high probability of failing to meet management objectives. This, however, can be argued to be a failure on the model specification part, as the failure to meet objectives could be penalized using the utility functions.

2.5 Spatial and spatiotemporal BNs

Spatially explicit BNs are highly useful in ecological and socio-ecological research when the aim is to understand spatial changes or spatially varying ecological risks. While spatial BNs are not the topic of this thesis, they are covered here briefly for completeness. The references in this sub-chapter are not exhaustive, but only examples of relevant work.

A common approach is to combine BNs with ecological or both ecological an societal data from a spatial grid to analyze each grid cell separately (Balbi et al. 2016; Grêt-Regamey and Straub 2006; Guo et al. 2020; Ropero et al. 2018; Sahin et al. 2019; Stelzenmüller et al. 2010). Das et al. (2017) used spatial and temporal information in meteorological predictions by computing spatial and temporal weights for each location and year in respect to the predicted year to be used in their BN model.

Some authors also combine spatial and dynamic Bayesian modelling. Trifonova et al. (2015) created a dynamic BN model with hidden variables that also explicitly includes ecological interactions between the modelled areas. Chee et al. (2016) extend dynamic BNs to spatial domain through using GIS data and explicitly modelling spatial processes. Stritih et al. (2020) presented a sys-

tem based on geospatial grid data that accounts for spatial interactions such as neighborhood effects, and which can be run iteratively over multiple time steps.

2.6 When to use BNs?

In general, BNs are the most useful in machine learning when:

- The probability distribution of the prediction is required.
- There is background information such as domain knowledge and theory that the modellers want to incorporate into the analysis.
- There is a need to be able to analyse the solution, i.e. a black box solution is not desirable.
- There is a need to evaluate the whole system, not only one response variable given a number of features.

They may not be the best machine learning solution if the problem is essentially to predict the outcome (without its uncertainty) based on data; in that case a standard ML solution such as an artificial neural network, random forest, support vector machine etc. may be simpler to construct and serve equally well or better.

Bayesian networks have been used to aid environmental management (Aguilera et al. 2011; Barton et al. 2012; Kaikkonen et al. 2020). Their strengths include their transparency, i.e. the fact that the graphical causal models are relatively easy to grasp intuitively, their flexibility in incorporating different types of data, and the explicit treatment of uncertainty that may be crucial in decision making. Models that aim to aid decision making by evaluating the current status of the environment and outlining the probable outcomes of different management options are sometimes called decision support tools or decision support systems (Beest et al. 2020). It is possible to augment BNs with decision variables which outline the possible decision options or considered scenarios, and utility variables that include information of the utility, i.e. the benefit or harm, associated with the potential states of nature (Madsen et al. 2005). For example, the utility functions of a decision support tool dealing with alternative fisheries management models could be linked to the fish catches, the ecological status of the fish stock, and the costs and perceived fairness of the different management options. This makes it possible to compare the management options quantitatively through the utility associated with each decision option.

Chapter 3

Building the Bayesian networks in this thesis

The key stages of building a Bayesian network model are (Korb and Nicholson 2011):

1. Identifying the variables and their possible values / states.
2. Finding the graph structure.
3. Finding the parameters (i.e. probabilities).

Additional steps when creating a decision support model are:

4. Identifying the available decisions and their effects.
5. Identifying the utility nodes and their dependencies.
6. Finding the utilities.

The authors note that expert elicitation is a central tool for all of these steps. In this thesis, Steps 1–3 are taken in all the created models, and Steps 4–6 in the model of Paper V. In this chapter, I discuss how these steps were carried out in the models of this thesis. Korb and Nicholson (2011) emphasize that these steps are usually not taken once and in this order, but BN models are rather built iteratively and incrementally. Also the models in this paper were built iteratively to some degree: Especially in the models of Papers II, III, and V, the variable selection and graph structure finding was a simultaneous process in which the models' scopes were discussed and the model structure and variable selection were

somewhat entangled. In the case of model of Paper V, the discussions regarding the utilities were in focus since the beginning, and this shaped the process of building the entire model.

3.1 Identifying the variables

Feature/variable selection is an essential step before or during building the model structure. Chen and Pollino (2012) stipulate that all variables in the model, with the exception of latent variables, should either affect or be affected by the final outcome variable(s), and should be either manageable, observable, or predictable, as well as relevant at the scale of the model (Borsuk et al. 2004). They encourage making the model as simple as possible, pointing out that variables that are left out contribute to the model error or variance unexplained by the model, and are included into the probability distributions of the included variables, and hence not completely ignored.

Feature selection in this thesis is largely guided by the availability of data, as well as the substance experts' ecological understanding. In Paper V, the features that were selected into the model were the ones that are societally interesting and subjects of environmental management.

In Paper I, additional feature selection is a byproduct of the discretization based on the minimum description length principle (Fayyad and Irani 1993), and so happening in parallel with finding the values of the variables. This discretization algorithm finds break points that minimise the entropy of the feature given the class, maximising the predictive power. In Paper I, the candidate feature was excluded from the model if the discretization resulted in zero cut-points, as this implies that there was no discretization that would have reduced the entropy, meaning that the mutual information of the feature and class was very low. Therefore, the feature is not a useful predictor for the class. The discretization is further discussed in the next sub-chapter.

3.2 Identifying and discretizing the variable values

Most ecological measurement data are continuous by nature, but discretization can bring benefits such as easier understandability and interpretability of the models and more accurate learning (Liu et al. 2002). Whether it is better to discretize the data or to create a BN with continuous variables depends on the type of data as well as the assumed interactions between the variables. Discretization

of data allows for complex conditional probability tables that do not need to follow any parametric distribution. This allows for large flexibility in describing the system interactions, such as complex nonlinear dependencies between the variables. These are often called "nonparametric" models because statistical distributions' parameters do not restrict the interactions, but in reality the models naturally have a large number of parameters to capture the CPTs of the multinomial distributions, and therefore require large amounts of data to estimate them. In this thesis these are called unrestricted distributions. This may become a problem if the model is based on relatively small data. In this thesis, Paper I solves this problem by reducing the parameter space through fitting tree-augmented Naïve Bayes (TAN) models, and Paper V by using simulated data sets that are much larger than any data that could be gathered from nature.

Nojavan et al. (2017) note that discretization of continuous variables affects the modelling results, and the consequences of discretization to the inference have not been properly discussed. Therefore, they recommend avoiding discretization if possible. In Paper I we too note that the discretization scheme may have a strong effect on the results, but argue that in contrast to seeing the different results under different discretization schemes as a weakness, it should rather be seen as an informative result and insight into understanding the data better. In a world of nonlinear responses, so often the case in ecological data (e.g. Hsieh et al. 2008), data analysis must be sensitive to different thresholds and also to the possibility of different factors being important at different ranges of the response. For example, there may be a certain environmental factor that strongly prevents or encourages extremely high values of the class variable from occurring, but otherwise doesn't correlate with the class. This type of cases can be detected effectively using the scheme proposed in Paper I, given that the class discretization is selected suitably to follow domain specific, e.g. ecological or judicial, considerations.

Chen et al. (2017) discuss different data discretization approaches proposed for BNs. They note that methods such as equal interval based on the marginal distribution of the variable lead to suboptimal performance. They propose a new Bayesian discretization method that takes into account the network structure. This is a very promising development for data analysis through Bayesian networks. Ropero et al. (2018) propose the Chi-Merge discretization method for environmental data based on a case study with a binary class variable prediction. However, the modeller must be sensitive to the tradeoff of understandability and model performance - if the model is to be used by or analysed with stakeholders, discretization breakpoints that are easily understandable, such as equal intervals,

may be preferable to the computationally more optimal, but less transparent discretizations.

If the data are to be discretized, the modeller must choose between algorithmic solutions (Liu et al. 2002) and expert-judgement-based discretization. Which one of these is preferable depends on the aims of the modelling exercise. If the aim is to predict the value of the interest variable as accurately as possible, algorithm-based discretization that maximises the predictive power of the model is preferable (Paper I). However, if the aim is for the model to be understandable by stakeholders such as decision-makers, binning that is intuitive to humans, such as equal-width discretization (Paper V) is preferable.

Data in Paper I are discretized using an algorithm based on minimum description length principle (Fayyad and Irani 1993). The class variables of the prediction models are discretized using expert judgement (and different discretization principles are explored), and the features are discretized to minimise the entropy of the class given the feature. The conditional entropy of the class given an observed variable is

$$H(C|V) = -\sum_{v \in V} P(v) \sum_{c \in C} P(c|v) log(P(c|v)) \tag{3.1}$$

where $P(v)$ is the probability of each value of the feature V, and $P(c|v)$ is the conditional probability of each value of class C given feature V.

This discretization approach maximises the predictive power of the models, and in Paper I, helps identify the variables that may have the strongest effect on the variable of interest. It has to be noted, however, that feature selection like this that looks at the mutual information between just two variables can fail if there are interaction effects (Murphy 2012). This is a potential problem with the models of Paper I as well, as it is possible that a feature is only informative given another, confounding, feature.

Paper V uses causal ecosystem models, which are in turn calibrated with empirical data, to simulate a large amount of data under different future scenarios. These simulated data were used to learn the parameters of the BN model, and they allow a robust assessment of even rather complex CPTs. It has to be noted that as the data originate from models and not nature, it is already a simplification of the real system. The simulated data are discretized into intervals of equal width for easy communication to decision-makers and stakeholders.

When discretizing the variables using expert judgement such as in Paper V, the amount of available data must be taken into account, as unrestricted distri-

butions have a high number of free parameters, and therefore require a lot of data to be estimated robustly.

As a comparison, the probability density function of a k-dimensional Gaussian distribution is

$$p(x|\mu\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{3.2}$$

where $\mu \in R^k$ is a k-dimensional location vector and $\Sigma \in R^{k*k}$ is a $k \cdot k$ -dimensional variance-covariance matrix, which is a symmetric positive definite.

As the variance-covariance matrix is symmetric, the number of free parameters in it is

$$df = \frac{k(k+1)}{2} \tag{3.3}$$

Therefore, fully defining a Gaussian joint probability distribution over k variables takes $\frac{k(k+1)}{2} + k$ parameters.

In an unrestricted, discrete joint distribution, the number of free parameters is

$$df = j^k - 1 \tag{3.4}$$

where k is the number of variables and j is the number of discrete bins in each variable. As the distribution is unrestricted, the probability of each combination of variables can freely take any value (subject to normalisation) except for the last one which is determined so that the total probability sums up to 1.

Therefore, the number of parameters to estimate increases fast as the number of bins per variable on one hand, and the number of parents of any given variable on the other hand increase. This means that a lot of data are needed to estimate these parameters reliably. In Paper V, the data originated from simulation runs of ecosystem models, and this large data set allowed the discretization to a relatively high number of bins.

When learning continuous conditional probability distributions from observational data as in Papers II–III, simplifying assumptions need to be made, typically that the variables are Gaussian and the associations between the variables linear. These assumptions allow for better generalisation from sparse data. If these assumptions hold, modelling the domain using continuous variables may be the best option (Nojavan et al. 2017). The price of this is that they are not able to fit nonlinear responses, which may be a problem in ecological research. Incorporating hidden variables into the models (as done in Papers II and III) introduces

Table 3.1: Properties of the BN models included in this thesis

Paper	Structure	Data
Paper I	ML several algorithms	Discretized (entropy minimization)
Paper II	Expert judgement	Continuous (Gaussian)
Paper III	Expert judgement	Continuous (Gaussian)
Paper V	Expert judgement	Discretized (equal interval)

more flexibility and therefore also more complexity to the model. However, if the data distributions are not well-behaved, or the associations between the variables are nonlinear, discretizing the variables may be advisable.

3.3 Finding the graph structure and parameters

Bayesian network graph structure can be either expert-judgement-based (usually causal) or derived from data through machine learning algorithms, which commonly are either constraint-based, using conditional independence tests to identify the dependence structures of the data; score-based, using goodness-of-fit scores; and hybrid algorithms that use both (Scutari et al. 2019). Models in this thesis have both machine learning and expert judgement based structures, and continuous and discretized data (Table 3.1).

Paper I explores different algorithms for finding model structures from data. The explored algorithms are the PC algorithm (Sprites et al. 1993), the score-based Greedy search-and-score algorithm, using both Akaike and Bayesian information criteria, and tree-augmented Naïve Bayes (Friedman 1997). Additionally, the Naïve Bayes structure was tested as well, but it does not require structural finding.

The PC algorithm begins with a fully connected model and removes arcs that can't be justified based on conditional independence measured through partial correlation. If an arc is removed in error early in the process, the error is likely to cascade. The PC algorithm works well with small models with large data, not so well with large models with moderate data (Korb and Nicholson 2011).

The greedy search algorithm for learning the structure of a Bayesian network (Chickering 2002; Scutari et al. 2019) searches through the possible network structure and returns the structure with the highest score. This implementation ignores the missing data, which is rather suboptimal for the present data set.

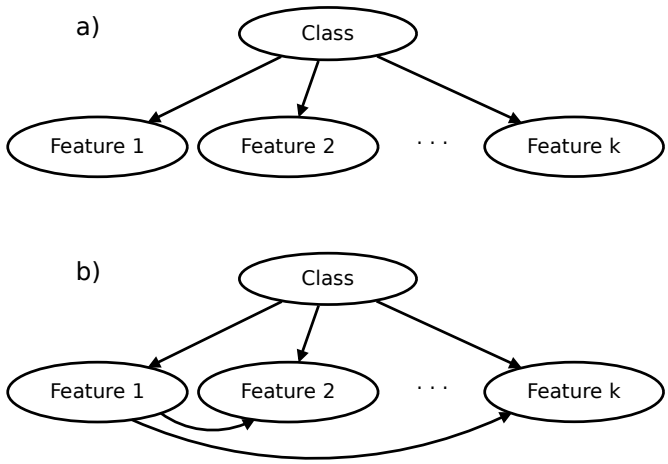


Figure 3.1: a) Naïve Bayes classifier. Class is the parent of all features, and the features are assumed to be d-separated given the class. b) Tree-augmented Naïve Bayes. A maximum of one additional parent is allowed for each feature.

Tree-augmented Naïve Bayes (Figure 3.1b, Paper I) is an extension of a Naïve Bayes (NB) (Figure 3.1a) model structure, usually used as a classifier. The Naïve Bayes classifier is structured so that the class variable is a parent to all the features, and there are no other arcs between the model components. This structure strongly reduces the parameter space and allows for robust assessments of all CPTs of the form $P(feature|class)$. The lack of arcs between the features imply that the features are independent (d-separated) given class, a simplifying assumption giving rise to the word naïve in the name. Despite their naivety, NBs perform well in many classification tasks. The probability distribution for the class variable is easily computed by applying the Bayes rule, for each possible state of the class:

$$P(c_i|x_1,\ldots,x_k) = \frac{P(x_1,\ldots,x_k|c_i)P(c_i)}{P(x_1,\ldots,x_k)} = \frac{P(x_1|c_i)\cdot\ldots\cdot P(x_k|c_i)P(c_i)}{P(x_1,\ldots,x_k)} \quad (3.5)$$

where c_i is the ith state of the class variable and $x1,\ldots,xk$ are the observed features.

TAN models relax the assumption of feature independence by allowing a maximum of one additional parent for each feature. This makes the classifier more

realistic through capturing the key correlations between the features while still keeping the parameter space small and therefore, the model robust. In Paper I, the TAN models were found using the Chow-Liu algorithm (Chow and Liu 1968). The Chow-Liu algorithm searches for a TAN model structure that maximises the likelihood of data given the model structure. This is done by computing the mutual information of all the possible edges given the class variable, and selecting the edges with highest mutual information.

There are also approaches that use Bayesian metrics, i.e. find a model that maximises the probability of the model given data. However, there are two difficult tasks in this: computing the scoring metric $P(model|data)$, and searching the exponential model space.

Cooper and Herskovits's K2 (Cooper and Herskovits 1992) was the first attempt at learning Bayesian networks structures over discrete variables. It is based on computing the metric $P(h_i|data)$ for individual hypotheses h_i by brute force through turning the problem into a combinatorial counting problem (which restricts it to discrete models. This method assumes that the samples are i.i.d., there are no missing values in the data, the causal factors don't interact with each other, and there's a uniform prior over the hypothesis space. Under these assumptions, counting $P(h_i|data)$ becomes polynomial, but there is still the issue of exponential space of hypotheses. Assuming that we know the temporal/causal ordering of the variables reduces the hypothesis space.

Structural learning may be useful for systems that are not well-known but for which ample data exist (Chen and Pollino 2012), but a hybrid approach in which expert judgement is used to restrict and/or guide the structural finding may still be better (Alameddine et al. 2011; Julia Flores et al. 2011). In ecology, there is often a wealth of information about the ecological relationships based on ecological theory, observations from different areas, etc., while the complexity and dynamic nature of the systems makes it difficult for algorithms to find an accurate structure (Chen and Pollino 2012). In these cases, expert judgement may be the best approach to model structure finding. Expert judgement based model structures are used in Papers II, III, and V of this thesis.

In Papers II–III, the model structure was based on ecological knowledge about the interactions between the ecosystem components. When discussing feeding interactions, ecologists talk about top-down and bottom-up effects: roughly speaking, top-down refers to the regulation of prey population by the predator, while bottom-up refers to the regulation of predators by the availability of prey. These effects are simultaneously in play in ecosystems, and it may be difficult to construct a model with sufficient temporal and spatial resolution to cover them, not

to mention that data are rarely available to estimate these effects. Therefore, Paper III explored replacing a local top-down model structure of Paper II with a bottom-up structure in the relationship between zooplankton and zooplanktivorous fish. This improved the predictions, but did not make a radical difference to the model interpretation. Model parameters, i.e. the conditional probability distributions, of all models are found using the EM algorithm described in Chapter 2, after the model structure has been decided.

3.4 Decisions and utilities

Bayesian network models can be complemented with decision options and scenarios as well as utilities (benefits or costs) related to different states of the system. Decision options differ from random or chance variables by the fact that they are assumed to be completely controlled by the person using the model; i.e. there is no uncertainty or randomness related to the value of the decision variable, but its value can be selected freely. Decision variables can be e.g. the treatment administered to a patient, a policy option that is chosen on a societal level, etc. Decision variables typically affect the chance variables in the model, e.g. the patient's health metrics or the ecosystem state.

In Paper V, climate scenarios were treated as decision variables. This choice was made to enable the evaluation of the future development in the first place. Models that describe the current status of a system are usually built to present the full joint probability distribution over the modelled domain, i.e. to show the full probability distribution of the variables over all possible combinations of the variables being modelled. In future scenarios this is not usually possible, however: there is an exponentially large number of potential future pathways, making it impossible to take them all into account. The climate projections and management options present only a few, selected scenarios among all the possible climates, nutrient loads and fishing mortalities. Therefore, the model must not be viewed as a representation of the joint probability distribution over the system's future.

Utility nodes encode the utility (benefit or harm) that results from the states of the system. If these utilities were all measured in the same units, e.g. in money, the model can be used to compute the decisions that maximise the overall expected utility, taking into account the probabilities of each outcome and their utilities, and if some variables are instantiated (their values are known), the effect of this knowledge on the expected utility can be taken into account.

The decisions, scenarios, and utilities of Paper V model were selected to be relevant for the management discussion of the Baltic Sea. The scenarios were selected to be in line with the climate change and nutrient scenarios that are often used in the Baltic Sea management research, to be able to contribute to this discussion. The utilities were based on the EU policy goals outlined in the Marine Strategy Framework Directive, and in the case of economic benefits, straightforward profits of the fishery. The utilities are not measured in the same units: while the fishery can be appraised based on its profits, there is no monetary value associated with reaching good environmental status. Therefore, the model is not used to rank decision options based on expected utilities.

Chapter 4

Probabilistic prediction of indicator value

Environmental indicators are increasingly used to reflect the status of the marine environment (Berg et al. 2015; Heiskanen et al. 2016; Shin and Shannon 2010; Teixeira et al. 2016; Uusitalo et al. 2016). Typically, they reduce the status of a selected ecosystem component, i.e. some value on a continuous scale, to two classes, acceptable and unacceptable; this gives the managers a quick view on which aspects of the ecosystem need improvement.

While the current indicator status indicates whether the ecosystem component is in an unacceptable state, it does not contain information on which issues have driven it to that state. Typically, a large variety of environmental factors can affect the state of any other ecosystem component, and studies are needed to understand the ecological interactions between the multiple components. However, the scarcity of data and the complex shapes of interactions can make this difficult. In Paper I, we evaluate different Bayesian classifiers to see how well they can predict two environmental indicators' status. The coastal fish abundance indicators are challenging, since there are a multitude of factors that are identified to potentially affect the coastal fish abundance, and data on them are sparse.

To do this, we fit a series of Bayesian network based classifiers (Friedman et al. 1997) to environmental indicator data in order to understand which environmental factors predict the indicator outcome (Paper I). The data are rather sparse, meaning that it is impossible to fit any statistical models that require full data sets as even imputation would have been difficult for the sparsest variables. The data consists of continuous variables such as water salinity and species abun-

dances. The data analytic challenge is that the data availability varies between the features: some are available all through the geographical and time span, some are available only for some years and/or areas. Therefore, an analysis framework that can cope with missing data is required in order to analyse these data together and get an idea of their relative importance. Robustness is required in order to avoid overfitting due to the sparsity of data (n=186).

Paper I evaluates the different structure finding algorithms mentioned in Chapter 3, namely the PC algorithm, greedy search, and tree-augmented Naïve Bayes (TAN), and additionally Naïve Bayes for which the structure is fixed. Model selection criteria are the predictive accuracy of the created models and expert validation of the model linkages. The TAN models were the best models in the light of these two criteria and were analysed in more detail. The superiority of TAN in this case is likely to stem from the fact that the TAN structure is restricted enough to guide the structure learning process to avoid overfitting to the sparse data, while still allowing for the correlations between the features and hence avoiding double-counting of strongly correlated features.

Bayesian classification methods such as Naïve Bayes and tree-augmented Naïve Bayes perform soft classification, i.e. they produce the probability that the class variable is in any given state rather than producing a *maximum a posteriori* estimate (Liu et al. 2011). Paper I uses three performance metrics to evaluate the behaviour and goodness of the models. Error rate and area under curve (AUC) (James et al. 2013; Murphy 2012) are based on the MAP interpretation of the posterior distribution, while entropy reduction assessment is based on the probability distribution of the prediction. AUC and error rate evaluate how well the models predict the indicator value overall, while entropy reduction can be used to analyse which features have the highest predictive power in relation to the class.

Analysing the entropy reduction and the breakpoints found by the entropy minimisation discretization can provide insight to the ecological interactions of the studied system. Spurious correlations, missing confounders, and other issues related to the causal interpretation of the models do not affect the predictive accuracy, but the location of the breakpoints and the relative strength of the interactions (as measured through entropy reduction) may, however, bring additional insight to the system and help create hypotheses that can be tested in other areas and with new data sets.

An additional analysis that could provide insight to the robustness of these analyses and potentially help ecosystem managers would be to beach the models with data originating from some of the geographical areas, and use these models

to predict the indicator values for the other areas. This would give insight to whether the models are generalisable within the northern Baltic Sea. If they generalised over different sites across this area, that would be highly useful information to managers who could, to some extent, extrapolate the fish indicator values based on the other, more easily monitored environmental variables. Further, experimental fishing effort could be targeted to areas where the indicator value prediction is the most uncertain, and therefore the entropy reduction gained from the new data the largest.

Chapter 5

Dynamic Bayesian networks for detecting ecosystem change

As discussed in Chapter 1, environmental data of temperate ecosystems are often composed of yearly observations, and therefore the number of data points can be just tens of observations. These data are used to fit parameters of models that describe the relationships between ecosystem components, such as predator-prey dynamics. Usually the whole time series is used to fit the curves, implicitly assuming that the relationship has stayed the same along the whole time series. However, anthropogenic stressors change ecosystems, and this assumption may not hold. The change in the functional responses may, however, be difficult to identify from the relatively short and noisy time series.

Dynamic Bayesian networks (DBN) with hidden variables (HV) provide one avenue to detect concurrent changes among multiple such functional responses. Hidden variables are set up as parents of all or a subset of variables in a dynamic BN model. The latent, non-observed HV fits to the noise in the relationships, and if the fit of the functions changes so that the residuals have a pattern, this is reflected in the HV, causing it to have a pattern instead of just a steady noisy signal. Therefore, DBNs with HVs enable the modelling of non-stationary processes (Murphy 2012). Different set-ups of HVs enable the detection of changes in different parts of the modelled system (Papers II–III).

In Papers II–III, Dynamic Bayesian network models with hidden variables are fitted to the food web data of the Gotland Basin in the central Baltic Sea. No such models have been explored previously in the Baltic Sea, despite the fact that the Baltic Sea ecosystem features relatively good data on its comparatively simple food web. Paper II evaluates how this model type performs on the Baltic

Sea data and whether such models could be used for detection of ecosystem shifts, while Paper III continues this work to evaluate the effect of the amount of data as well as different structures of the core model.

Paper II builds the model structure for the central Baltic Sea food web based on expert knowledge (Figure 5.1). The model structure reflects the scientific understanding of the key components of the food web and their interactions. The model structure was created using expert judgement instead of structure learning from data to leverage the knowledge accumulated in marine science in general and in Baltic Sea research in particular, and also since current methods for learning hidden variables require experts to choose a fixed network structure or a small set of possible structures (Friedman 1997). The model structure includes the predator-prey relationships between the fish species and zooplankton as well as the dependence of cod recruitment on reproductive volume, i.e. salinity and oxygen conditions in the Central Baltic Sea. Fishing-induced mortality of the three fish species is included. The model includes the juvenile stages of the fish species, i.e. the 0- and 1-year-old individuals of all three species, and additionally the 2- and 3-year-old individuals of cod. After these ages, the juvenile fish are assumed to mature and join the spawning stock biomass (SSB). The juvenile stages for which there are no data are marked with dashed bubbles in Figure 5.1.

The dynamic model is defined in one-year time steps. The spawning stock biomasses of the three fish species are autoregressive, since they consist of individual fish that live many years. Each young age class of fish are modelled separately until they reach maturity and join the spawning stock; they exhibit temporal dependency so that the k year old fish are $k+1$ years old in the following time slice, until they reach maturity and join the spawning stock at age 2 (herring, sprat) and 4 (cod). The rest of the variables are assumed not to directly depend on variables in the preceding year, although they may exhibit temporal autocorrelation due to the fact that variables affecting them are temporally autocorrelated.

It is clear that also other variables and processes can affect the ecosystem dynamics, but have not been either identified or on which no data exist. To examine this possibility, three versions of the model were built with different sets of generic hidden variables (Friedman 1997; Kwoh and Gillies 1996): A model with one generic hidden variable, linked to all other variables in the model ("generic HV"); a model with two semi-generic hidden variables: one that is linked to all cod variables (cod fishing, all cod life stages) (the "cod HV"), and another similarly linked to all sprat and herring variables (the "clupeid HV"); and a model with both the generic HV and the clupeid and cod HVs.

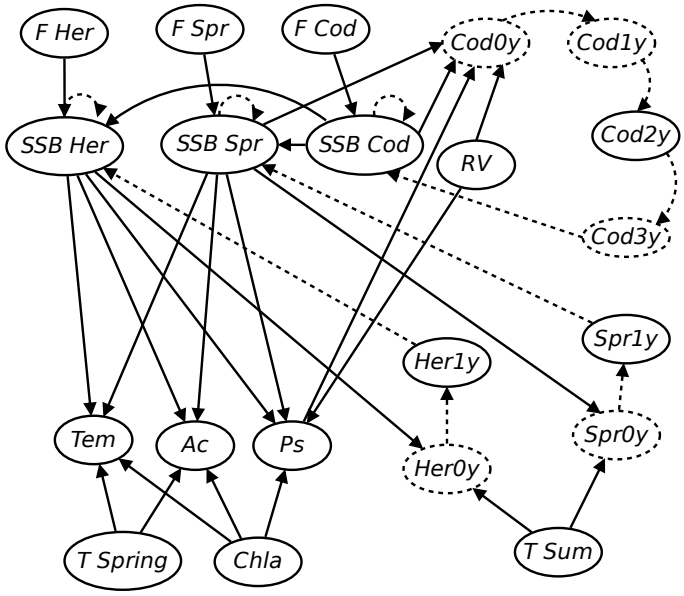


Figure 5.1: One time-slice of the expert knowledge based model for the Central Baltic Sea food web. The hidden variables are not shown. Figure reproduced from Paper II.

Hidden variables in these models relax the inherent assumption of model fitting that the data arises from a stable distribution - the hidden variables, serving as parents for some or all of the observed variables, effectively allow for the conditional probability distributions of the children given the parents to change, fitting to the residuals of these data fits. However, as the HVs are linked to, and therefore must fit to the residuals of multiple variables, this restricts their values and prevents overfitting of the hidden variables to explain all the variance or noise of the parent-child relationship. Therefore, if a pattern is observed in a hidden variable, it implies a pattern of model fit residuals in those observed variables it is linked to. This way, we can detect synchronous changes in the functional forms of relationships between the observed variables.

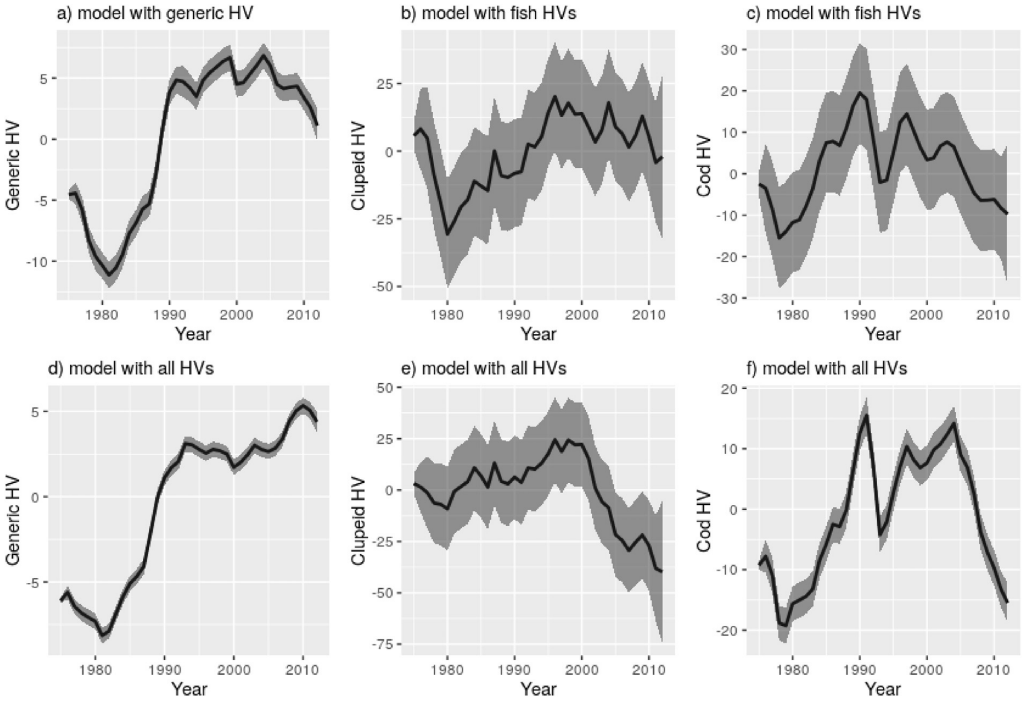


Figure 5.2: The temporal patterns (mean + sd) of the hidden variables in the three model versions of Paper II. Figure from Paper II.

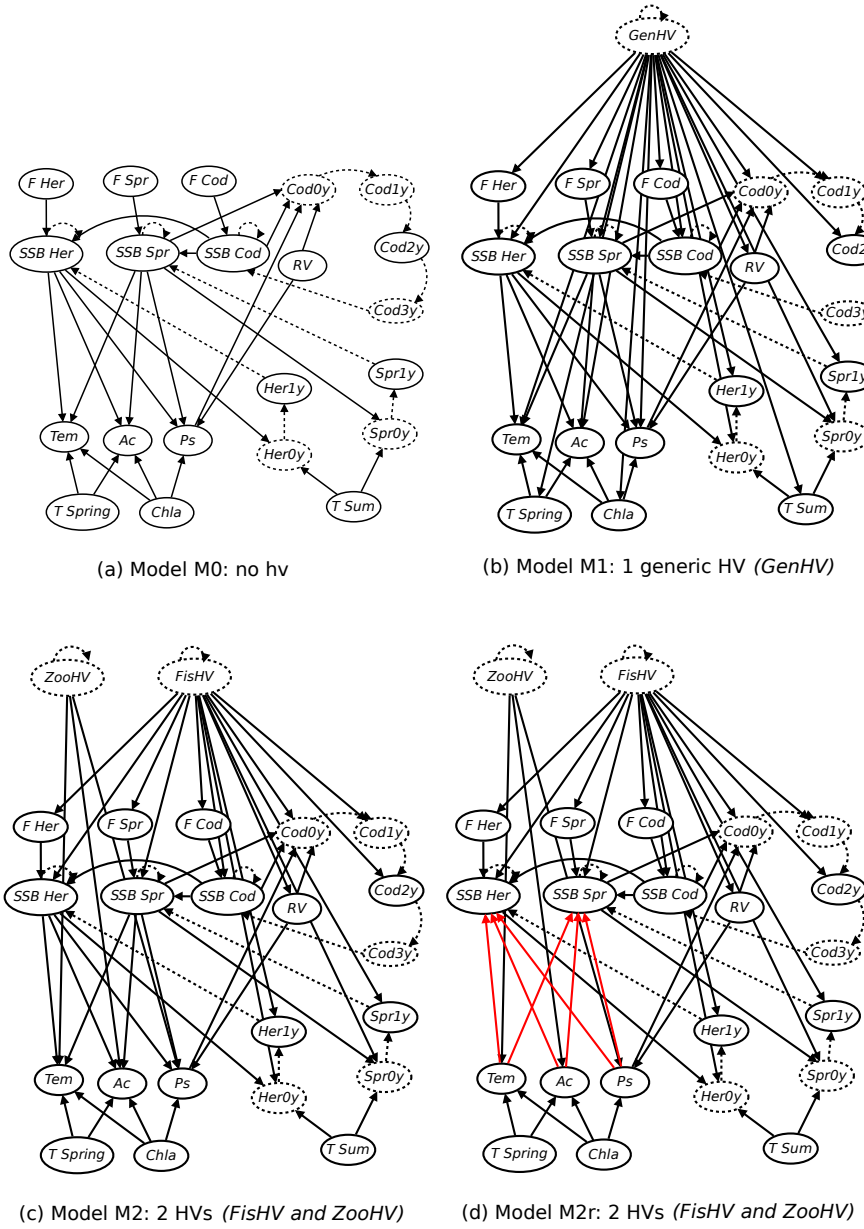
The three models were parameterized using the available data. Since EM learning is prone to getting stuck in local optima, the parameterization was run 100 times for each model, selecting the model with the lowest log-likelihood. The models were evaluated to see whether the hidden variables can detect the Baltic Sea regime shift (Alheit et al. 2005), and whether there are differences between the models. The analysis shows that hidden variables are indeed able to detect the regime shift based on these relatively few data. This is in line with the observations of Trifonova et al. (2019), Trifonova et al. (2015), Trifonova et al. (2017), and Trifonova et al. (2021) that DBNs with hidden variables can be useful in detecting patterns in changing ecosystems. The patterns of the HVs (Figure 5.2) were consistent across the different model versions, although the variance was much larger in some model versions.

Setting multiple hidden variables to monitor different parts of the ecosystem can help pinpoint the change to specific parts of the ecosystem. For example, the HV pattern in the model with only one generic HV (Figure 5.2a) exhibits a steady rise in the 1980s and a downward turn since 2000, whereas in the model with the generic HV, the cod HV and the clupeid HV (Figure 5.2d-f) the generic HV exhibits the increase in the 1980s, but not the decrease in the 2000s, while the clupeid and cod HVs exhibit a clear decrease in the 2000s. This implies that the decrease of the 2000s is focussed on the cod. However, including multiple HVs carries an elevated risk of overfitting the hidden variable, so care must be taken. The DBN models are transparent, allowing the examination and critical assessment of all the model components and model fits. The causal nature of the model structure together with this transparency makes the models useful in better understanding the ecological interactions and their changes.

Paper III continues exploring and developing the models of Paper II, evaluating the amount of data needed for making reliable assessments, and the future prediction abilities of these models. Further, alternative model structures, including a Naïve Bayes model with the hidden variable as the class variable, are explored (Figure 5.3). If there are multiple plausible expert-judgement-based model structures, it becomes crucial to understand whether the model structure has a strong effect on the inference. Therefore in Paper III we fit a model with reversed links between these fish and zooplankton and examined a zooplankton-specific HV. In addition, we fit a Naïve Bayes model in which the hidden variable is the class, and other variables features.

In Paper III, we examine whether different model structures affect the modelling result. While the DBN model allows the explicit representation of feedback loops between the variables, the one-year time step is still so long that some of the ecosystem components interact during that time, and the question remains how those interactions should be modelled. One such interaction is between plankton-feeding fish and zooplankton, which can be argued to affect each other within one time step, and ecosystem experts could argue to have the parent-child relationship between these variables either way: either so that the fish affect (reduce) the zooplankton biomass through eating zooplankton, or that the zooplankton biomass affects the growth and survival of the fish through food availability (Figure 5.3d). How this is modelled changes the d-separation properties of the model, so the change has an effect on how the model parameters are evaluated.

We fit all of these models both with an increasing amount of data, and with a moving window of the same amount of data to examine how the amount of data affects the predictive power and learning the HV patterns. We show that with



(a) Model M0: no hv

(b) Model M1: 1 generic HV *(GenHV)*

(c) Model M2: 2 HVs *(FisHV and ZooHV)*

(d) Model M2r: 2 HVs *(FisHV and ZooHV)*

Figure 5.3: Alternative model and HV structures evaluated in Paper III. Figure reproduced from Paper III.

just a small amount of data, the HV estimation is uncertain, but with more data, they all stabilise to more or less the same pattern (Figure 5.4). This indicates that the exact model set-up may not be critical to the results. This is good news for modellers, as identifying the "correct" model structure may be impossible. It is found that the different model structures don't make a major difference for the predictive accuracy of the models. The model M2r (Figure 5.3d) HVs have a lower variance than those of the model M2 (Figure 5.3c). The patterns of the fish HVs were similar in both models, while the ZooHV shows a clearer stepwise change in the model M2r while its patterns in model M2 is more zig-zagging with higher variance (Figure 5.4b-c).

When the models are used for future prediction (Paper III), the uncertainties increase rapidly and the models don't therefore offer useful predictions beyond the insight that prediction of a complex system without using any "forcing" variables, such as setting scenarios for the physical features such as temperature, is highly uncertain. The different models are compared in terms of their predictive accuracy, and the Naïve Bayes model predicted the future slightly better than the more complex, and arguably more realistic, models. In the context of feature selection, it's been shown that prediction-optimal feature selection doesn't result in model selection consistency, i.e. the feature set that performs best in cross-validation is not the true feature set (Meinshausen and Bühlmann 2006; Murphy 2012). Therefore, best prediction ability must not be interpreted as being the sign of the model being closer to the true model than others.

Multiple improvements could be made to the family of Central Baltic Sea DBN models presented in Papers II–III. Higher-order time lags, spanning over more than one time step, could be experimented with to see if they help with the predictions due to the multiannual cycles of the fish individuals, which have now been carried through the time slices in the model structure. Reliable parameter estimation would, however, need more and more data as the model complexity grows (Murphy 2002), which is likely to be a practical hindrance to this avenue.

Sáez et al. (2015) present probabilistic methodology to detect and visualise changes in data distribution, based on incrementally estimated posterior Beta distribution of the Jensen-Shannon distance, a symmetrized and smoothed version of the Kullback-Leibler divergence. They demonstrate the method with pdfs of individual variables, while the shift detection in ecological time series might require monitoring multiple variables at the same time, due to the often large error and variability in the data in comparison to the shift amplitude. It would be interesting to compare their approach to the DBN-HV approach.

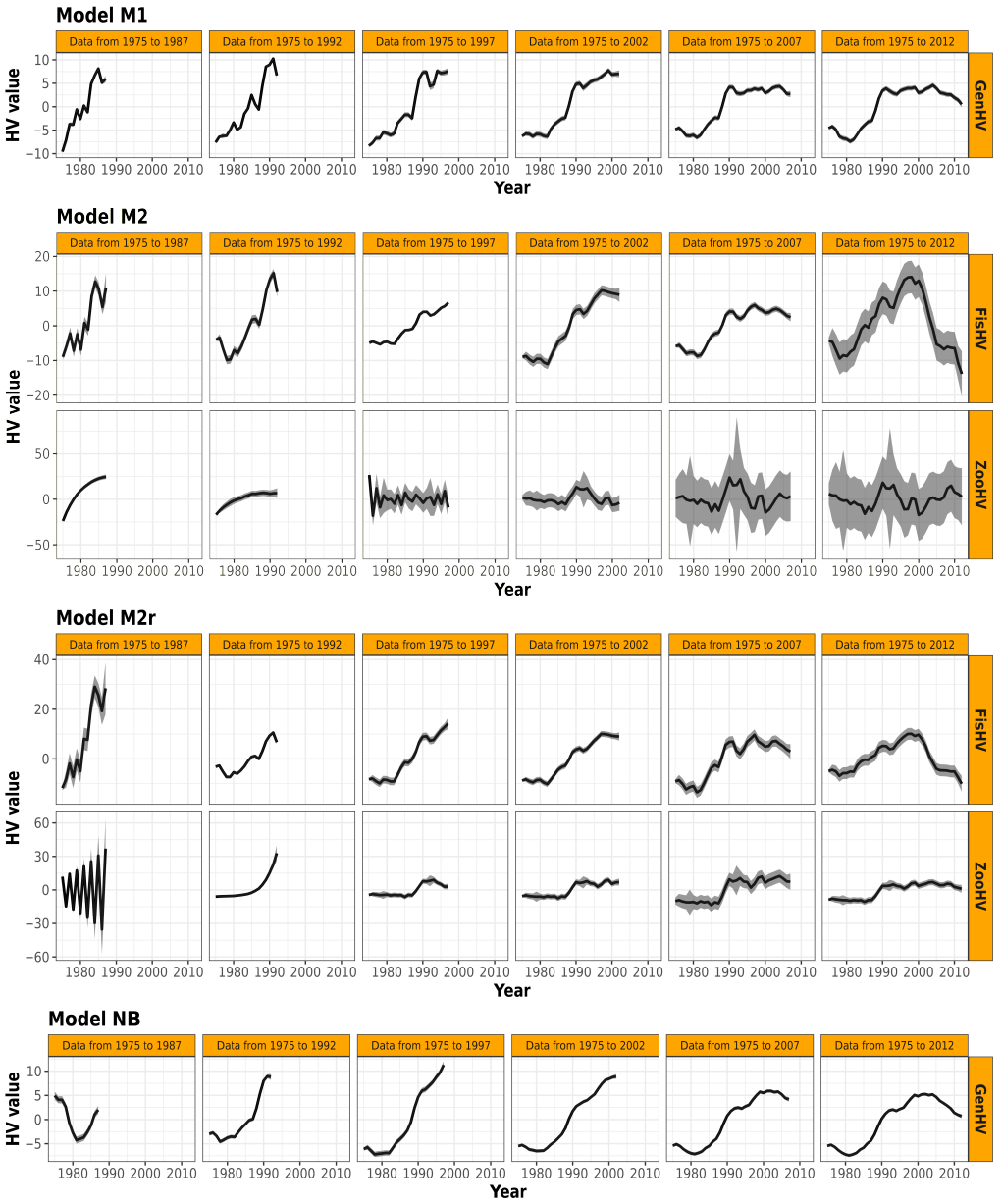


Figure 5.4: The HV patterns in the different model versions of Paper III, when an increasing amount of data is used. Figure from Paper III.

In this work, the HVs detecting the ecosystems shift were modelled as continuous variables. Binary or trinary HVs could be explored to see if the system seems to have just two or three alternative ecosystem states, as done by Trifonova et al. (2021).

Currently the DBN model is just one spatial unit, representing the Central Baltic Sea. Additional areas could be added to a spatially explicit DBN model (Trifonova et al. 2015; Trifonova et al. 2017; Trifonova et al. 2021).

Currently the DBN models assume that the observations are error-free, i.e. what is in the data is the true situation in nature. This is not likely to be true. The models could be improved by implementing an observation model, in which the observed value is dependent on, but not the same as, the modelled value in the system. Adding this observation model would increase the number of latent variables considerably, increasing the risk of overfitting (Murphy 2012). This could potentially be remedied by careful selection of priors, and as the observation uncertainty would explain some of the random variation, it might improve the model fit.

Chapter 6

Environmental decision support

A central aim of data science is to process data into a form in which it provides true and useful information for the purposes of various types of decision making. In environmental management, the evaluation of different management options is challenging due to the holistic nature of the decisions, i.e. that any management decision is likely to affect multiple parts of the ecosystem, and any ecosystem component is likely to be affected by multiple decisions. As ecosystems are complex and include processes on very different spatial and temporal scales, different parts of the socio-ecological system are usually studied using different models and analyses. Further, uncertainty estimates are often not available for these results (Schuwirth et al. 2019; Yates et al. 2018).

Paper IV discusses the nature and practical intricacies of assessing uncertainty especially when building metamodels that use other models' results as their input, as a review of these methods was lacking despite the need for them in environmental management related modelling. The review covers the use of expert judgement, model emulation, sensitivity analysis, temporal and spatial variability in the model outputs, the use of multiple models, and statistical approaches for assessing uncertainty related to the output of deterministic models. There is no single method that is the best for all purposes of uncertainty estimation, but the best way depends on the source models and the amount and quality of information available to the modeller.

Expert judgement is by definition subjective to some degree, but it can be the best approach if there are only small amounts of data from the studied system, but scientific understanding about the study topic (e.g. species, habitat) exists. There are also numerous methods for helping to elicit the expert judgement (see e.g. Morgan and Henrion 1990; O'Hagan et al. 2006; O'Hagan 2012).

Model uncertainty or sensitivity analyses are based on evaluating what the model results would be if some parameters or starting values of the model would be different but within a reasonable range (Chu-Agor et al. 2011; Saltelli et al. 2010; Tomassini et al. 2007). If the model's outputs vary widely due to these changes, it can be concluded that there is higher uncertainty about the model results. The uncertainty can be roughly quantified through experiments with the key parameters, but a thorough assessment requires a large number of model runs. Sensitivity analyses can be combined with expert knowledge to get the benefits of both approaches. Model emulators, low-order statistical approximations of the original models, can be used to reduce the number of model runs required in sensitivity analysis. They are often based on gaussian processes, and have been used particularly with very complex models such as climate models.

If the model gives predictions in space or time, this variability has sometimes been used as a proxy for uncertainty of the model's results so that model-predicted spatial and/or temporal variability is used as a proxy for the uncertainty (Lehikoinen et al. 2014). This method does not require any additional model runs, and it is easy to understand for people not familiar with probabilistic modelling, which may be an advantage in decision support modelling. However, this approach requires that attention is paid to making sure that the spatial or temporal variability indeed corresponds to the quantity that is being modelled in the decision support model.

If multiple, independently developed models over the same system are available, comparing their results can give valuable insight to the uncertainty related not only to the model parameters as with sensitivity analysis, but also to the structure of the model, i.e. which variables are included into the model, how they are assumed to interact, etc. If all of these models are built by trained professionals who know their trade, yet the results differ largely, it can be concluded that there exists major uncertainty about the process.

Finally, statistical or data-based approaches can be used when enough data exists from the studied system or systems that can be assumed similar enough. Pollino et al. (2007) combined expert judgement with data so that they weighted the elicited estimates by the experts' confidence through equivalent sample size, when learning the conditional probability distributions from data.

Paper V demonstrates a real-life case of an international, interdisciplinary environmental management problem and presents a first iteration of a BN model emulator to provide unified decision support. Sectoral management that focuses on just one aspect of the environment has proven to be insufficient, and more holistic ecosystem-based management has been proposed as a solution (Curtin

and Prellezo 2010). This, however, requires modelling and data analysis approaches that are able to process and integrate data in different forms and different temporal and spatial resolutions. Paper V demonstrates that BNs are a good tool for integrating model results operating on different scales, presenting the related uncertainties, and hence supporting ecosystem managers in making holistic decisions. Causal BNs also allow for counterfactual, "what-if" types of scenario evaluations and diagnostic analysis (Kelly et al. 2013; Pearl 2010). Paper V presents a BN for the Central Baltic Sea that integrates results from multiple models, studies and disciplines, and accounts for uncertainty stemming from multiple sources, such as model selection, model projections, and uncertain future scenarios. The model focuses on a timely question in European marine management, namely, whether and under which circumstances and management decisions both sustainable economic growth and good environmental status can be attained. The study system is the Central Baltic Sea and its eutrophication and commercially exploited fish stocks. As eutrophication operates largely on chemical and physical levels, and fishery on higher trophic levels, these processes are difficult to model adequately in one system simulation model. The developed BN therefore integrates results from three climate models, two biogeochemical models, a biomass-dynamic food-web model, and a bioeconomic fishery model. The model brings together results from the best available models that predict the responses of the socio-ecological system to a set of climate scenarios and management alternatives (Figure 6.1). The model can be simultaneously seen as a model emulator for the set of models included, and a decision support system.

This BN harmonises and summarises the results of selected management options on environmental status (both eutrophication and fish abundance) and society (fish catches and fisheries economics). The model consists of model emulators for two different biogeochemical models, which predict the nutrient and chlorophyll a concentrations in the sea based on climate and nutrient input scenarios, and being forced by different climate models, as well as a model emulator for a food web model that predicts the biomasses of various species (including the economically relevant fish species) based on scenarios of nutrient loading, climate, and fishing intensity. In Figure 6.1, the scenarios and other control variables appear in the top row, and the model predictions of the nutrient, chl-a, and fish abundances and catches, below them. The model predictions of the abundances are transformed into binary variables (further down in Figure 6.1) that encode whether these environmental states adhere to the definition of good environmental status or not. These variables give the probability that the ecological variable, e.g. the nutrient concentration, is in good environmental status

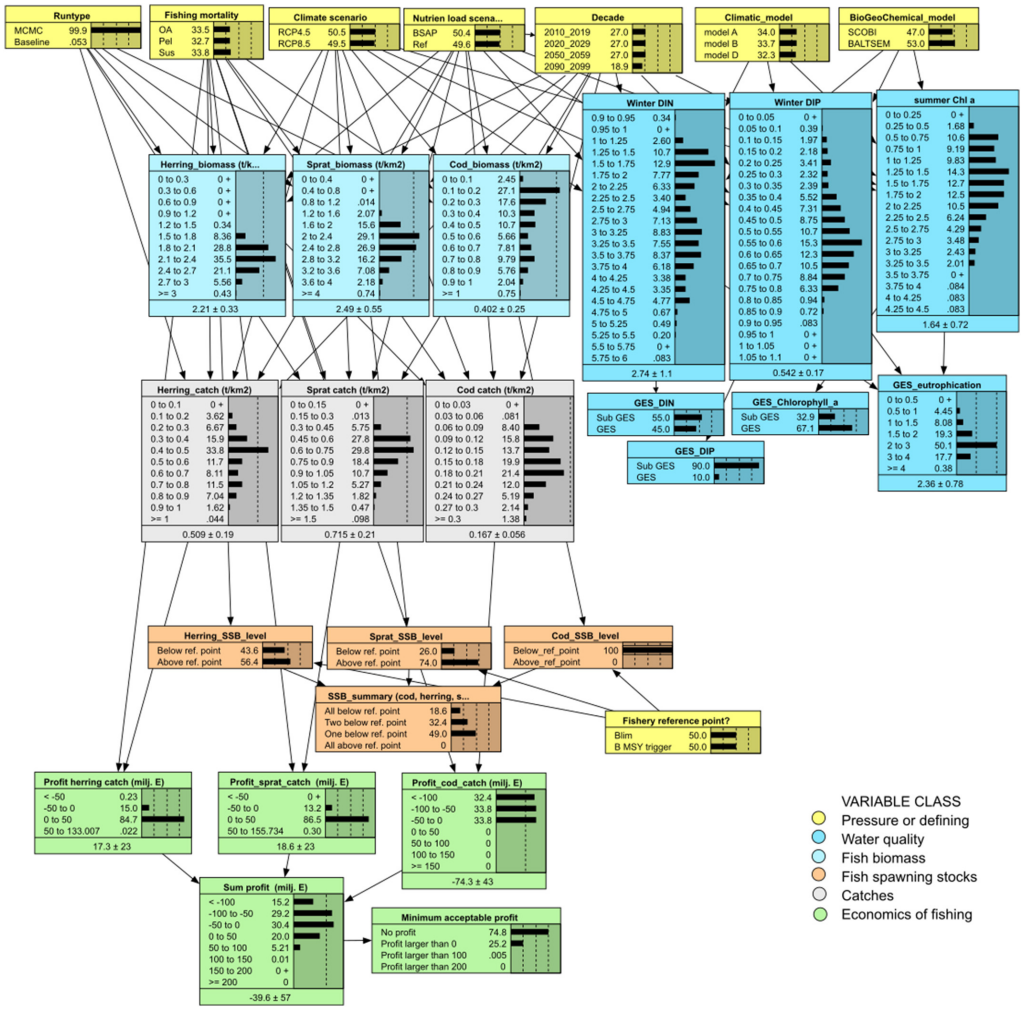


Figure 6.1: The model emulator / decision support BN model. Figure from Paper V.

as defined by European legislation. These probabilities are based on the best available uncertainty estimates, created through Monte Carlo simulations of the food web model, and different starting conditions of the biogeochemical models (forms of sensitivity analysis as discussed above).

Assessment of uncertainty of the model projections is an important, but unfortunately often overlooked, aspect of environmental modelling and particularly future projections and decision support (Beest et al. 2020). This issue has been recognized by the research community: Meier et al. (2019) reviewed and identified 14 different sources of uncertainty in scenario simulations of biogeochemical models, and Saraiva et al. (2019) found major uncertainties in the Baltic Sea biogeochemical model projections driven by different climate models. Beest et al. (2020) review how uncertainty is presented in Baltic Sea decision support models, and encourage better documentation of uncertainties and their communication to stakeholders. Bayesian networks offer a transparent and mathematically rigorous way to incorporate uncertainties into the decision support model, and a graphical tool to communicate these results, as we demonstrate in Paper V. In Paper IV, we acknowledge the difficulty of coming up with good uncertainty estimates, and propose a practical way of moving forward. I acknowledge that the uncertainty estimates generated in manners proposed in Paper IV are not likely to be perfect in the sense of capturing the true uncertainty of the phenomenon in question. It may also be hard to obtain uncertainty estimates that are comparable across the different models or domains that are evaluated together in a decision support model. If this is the case, the results need to be presented with sufficient care, communicating clearly that they are not fully quantitative. Even in this case, inclusion of the imperfect uncertainty estimates is better than the "ostrich strategy" of burying our heads in the sand and pretending the uncertainty does not exist. Again, Bayesian models offer a natural and elegant solution to the assessment and communication of (potentially cascading) uncertainties. While graphical presentations such as bar charts and interactive visualisations help communicate probabilities and uncertainties, communicating uncertainties stemming from lack of knowledge remains a challenge (Spiegelhalter et al. 2011).

Chen and Pollino (2012) note already in 2012 that BNs are increasingly being used to integrate multiple issues and system components and utilise information from different sources. While the basic concept of the BN model emulator (Paper V) is straightforward, there are multiple tricky issues that need to be solved in order to present the results in a meaningful and fair way. These issues include scaling the results of the individual models into the same scale (e.g. fish catches or biomasses can be represented per unit area or per the whole sea) and making sure that the shared variables (such as the management scenarios) are indeed defined in the same way in all parts of the model. If these prerequisites are not met, the results may be incomparable and therefore useless for management. This is naturally the case also if these models are presented in different scales and with

differently defined variables outside of a unified model emulator. To make sure that the results are meaningful, the data selection and treatment, as well as the results presentation, need to be done in close collaboration with domain experts.

In addition to these challenges, the assessment of uncertainty of the different model components can be very difficult (Papers IV–V). Blair et al. (2019) list a number of sources of uncertainty: the underlying data sources, the choice of model(s), model structure and parameterization, initial conditions and assumptions for model runs, model sensitivities to small changes, and scenarios. In Paper V, we did our best to come up with good uncertainty estimates for the model components. However, the uncertainty estimates are based on different types of evaluations: in the biogeochemical models, the uncertainty estimates stem from the ensemble of two different models (Blair et al. 2019), different starting conditions, and different climate models that are simulated, while on the food web model side, there is just one model, but it is, on the other hand, evaluated through making thousands of model runs and varying its parameters. Both of these approaches have their strengths, but they are not directly comparable. We could argue that the biogeochemical model uncertainty assessment scheme focuses on structural uncertainty, i.e. how much the different models differ, while the food web model uncertainty assessment focuses on parameter uncertainty, i.e. how much effect the assumptions on the model parameters have. Both of these are also affected by the differences in starting conditions. Perhaps due to these practical challenges, these kinds of models are still relatively rare.

A valuable aspect in BN models is that they can be improved part by part without an overhaul of the entire model. The model in Paper V could be improved by adding the estimates coming from another food web model, in order to evaluate the structural uncertainty on that part. This was not done as such models are very labour-intensive to make and therefore are not readily available, but the Atlantis model developed by Bossier et al. (2018) could be evaluated for its fitness to this purpose. The uncertainty estimates of the biogeochemical models could be improved to include parameter uncertainty - but on the other hand, as was shown in Paper V, there was a major discrepancy between the two models' projections so this issue might need some further elaboration before digging into parameter uncertainties of individual models. Additional utility nodes could be added to account for an even wider range of values the ecosystem state has, such as the coastal residents' experienced value of the fish stock abundance and eutrophication level.

Chapter 7

Conclusions

As noted in Chapter 1, ecological data are often relatively sparse, and describe a complex, adaptive, and potentially changing system. Even a long continuous observation will mean only tens or maybe a few hundred observations, a small number for many machine learning applications. Further, as ecosystem structures and the interactions within them are changing due to anthropogenic pressures, the data from decades or centuries ago may not be representative of the current, let alone the future, ecosystem. This is an issue that has received less attention than it deserves, perhaps due to the discouraging message that we may be even worse at predicting the future and the ecosystem responses to changes than we would expect. This situation calls for robust, uncertainty-explicit data analysis that allows the inclusion of different data sources and accuracies as well as the integration of knowledge accumulated through other means such as experiments and comparisons to other geographic areas.

The data scientific challenges of this thesis are focused on the analysis of sparse data on complex ecosystems. The sparseness comes in the forms of missing observations in a time series of observed variables, and as important variables missing altogether from the data. The ecosystem from which the data arises is rather complex compared to the available data, with potentially changing interactions on multiple spatial and temporal scales. Understanding such systems based on data requires collaboration between data analysts and domain experts in order to guarantee both the methodological soundness and domain-specific relevance. Data science has developed to respond particularly to the challenges of big data, meaning data with large size and/or high variety, velocity, value, and/or complexity (Kaisler et al. 2013). While the data sets in this thesis are not big in size or velocity (quite the opposite), the complexity of the systems

from which they arise, and hence the implicit complexity of the data sets, means that their analysis can benefit from data science methods.

When data are sparse, but scientific knowledge on the research topic is available, Bayesian methods are a good choice, as they allow transparent use of prior knowledge as well as simulation results from other types of models as part of the analyses (e.g. Korb and Nicholson 2011). This makes them a natural choice for environmental data analytics, reflected in their increasing popularity in the field (Aguilera et al. 2011; Anderson et al. 2021). On a more philosophical note, if and as the aim is to evaluate the uncertainties related to a unique event such as the ecosystem status, Bayesian interpretation of probability is the one to go for, as frequentist statistics don't allow probabilities related to unique events. This thesis evaluates different flavours of Bayesian networks for different data analysis tasks, namely classification (Paper I), modelling (Papers II–IV), and decision support (Paper V).

Machine learning is only ever as good as the data used in the analyses. Modelling uses augmenting information such as earlier scientific knowledge gained from other areas, knowledge about physical and biological processes, etc., to mitigate this shortcoming. Yet, it is often the case that there are relatively ample data on common or economically important species, but fewer data on more obscure elements of the ecosystem, which may nevertheless be crucial in respect to the ecosystem processes. Overcoming this challenge requires interdisciplinary work between the modellers and ecologists.

When complex computational methods are used on data from other domains, there is always a risk of misunderstanding and miscommunication due to different traditions and vocabularies of the disciplines. Scientists may fail to communicate relevant aspects of the work across the discipline gap, e.g. what is expected of the data and what is its reality. For example, the data may be biased in a way that is self-evident to the ecologist but not understood by the modeller, such as that data are collected only when organisms are expected to be present and the abundance distribution in the data doesn't therefore reflect the distribution in nature. These kinds of communication issues may lead to suboptimal or outright erroneous models. Interdisciplinary scientific work requires a learning process on the levels of individuals, disciplines, and types of knowledge, and it takes time (Haapasaari et al. 2012), but can lead to new advances in science (Grenci et al. 2019).

Ecological modelling often aims not only to increase knowledge, but also to ensure the sustainable management of the ecosystems. The management decisions may affect the income and opportunities of businesses and individuals.

Therefore, transparency and accountability are needed to build trust between the model providers and various stakeholder groups (Haataja et al. 2020). The systems need to be understandable, explainable, and verifiable (Fjeld et al. 2020; Haataja et al. 2020) on all steps of the process from data collection and data sets (Hutchinson et al. 2021) to algorithms and their application (Kim and Doshi-Velez 2021).

Bayesian networks are showing their strength for different tasks of environmental data analytics. Elegant handling of missing data, explicit and rigorous handling of uncertainty, and the possibility to use prior scientific knowledge and data together in analyses in a transparent way are strong advantages for Bayesian analysis for environmental data that often contain missing values and are scarce. The use of prior knowledge is a powerful addition to solving problems for this kind of data. This emphasises the point that the data analysts must be willing to work closely with domain experts and learn to understand the basics of the domain and the problem being solved, as the domain experts need to be willing to understand the fundamentals of the data analyses. This is the essence of interdisciplinary science. This thesis contains examples of different flavours of BN-based analyses that can help make sense of the relatively sparse data that are available for large portions of environmental science and management problems.

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., and Salmerón, A. (2011). "Bayesian networks in environmental modelling". In: *Environmental Modelling & Software* 26(12), pp. 1376–1388. DOI: 10.1016/j.envsoft.2011.06.004.

Alameddine, I., Cha, Y., and Reckhow, K. H. (2011). "An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics". In: *Environmental Modelling & Software* 26(2), pp. 163–172. DOI: 10.1016/j.envsoft.2010.08.007.

Alheit, J., Möllmann, C., Dutz, J., Kornilovs, G., Loewe, P., Mohrholz, V., and Wasmund, N. (2005). "Synchronous ecological regime shifts in the central Baltic and the North Sea in the late 1980s". In: *ICES Journal of Marine Science: Journal du Conseil* 62(7), pp. 1205–1215. DOI: 10.1016/j.icesjms.2005.04.024.

Anderson, S. C., Elsen, P. R., Hughes, B. B., Tonietto, R. K., Bletz, M. C., Gill, D. A., Holgerson, M. A., Kuebbing, S. E., MacKenzie, C. M., Meek, M. H., and Veríssimo, D. (2021). "Trends in ecology and conservation over eight decades". In: *Frontiers in Ecology and the Environment* 19(5), pp. 274–282. DOI: 10.1002/fee.2320.

Balbi, S., Villa, F., Mojtahed, V., Hegetschweiler, K. T., and Giupponi, C. (2016). "A spatial Bayesian network model to assess the benefits of early warning for urban flood risk to people". In: *Natural Hazards and Earth System Sciences* 16(6), pp. 1323–1337. DOI: 10.5194/nhess-16-1323-2016.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barton, D. N., Kuikka, S., Varis, O., Uusitalo, L., Henriksen, H. J., Borsuk, M., Hera, A. de la, Farmani, R., Johnson, S., and Linnell, J. D. (2012). "Bayesian networks in environmental and resource management". In: *Inte-*

grated Environmental Assessment and Management 8(3), pp. 418–429. DOI: 10.1002/ieam.1327.

Beaugrand, G. (2004). "The North Sea regime shift: Evidence, causes, mechanisms and consequences". In: *Progress in Oceanography*. Regime shifts in the ocean. Reconciling observations and theory 60(2), pp. 245–262. DOI: 10.1016/j.pocean.2004.02.018.

Beest, F. M. van, Nygård, H., Fleming, V., and Carstensen, J. (2020). "On the uncertainty and confidence in decision support tools (DSTs) with insights from the Baltic Sea ecosystem". In: *AMBIO* 50(2), pp. 393–399. DOI: 10.1007/s13280-020-01385-x.

Berg, T., Furhaupter, K., Teixeira, H., Uusitalo, L., and Zampoukas, N. (2015). "The Marine Strategy Framework Directive and the ecosystem-based approach – pitfalls and solutions". In: *Marine Pollution Bulletin* 96(1), pp. 18–28. DOI: 10.1016/j.marpolbul.2015.04.050.

Blair, G. S., Henrys, P., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S., and Young, P. J. (2019). "Data Science of the Natural Environment: A Research Roadmap". In: *Frontiers in Environmental Science* 7, p. 121. DOI: 10.3389/fenvs.2019.00121.

Blenckner, T., Llope, M., Möllmann, C., Voss, R., Quaas, M. F., Casini, M., Lindegren, M., Folke, C., and Chr. Stenseth, N. (2015). "Climate and fishing steer ecosystem regeneration to uncertain economic futures". In: *Proceedings of the Royal Society B: Biological Sciences* 282(1803), p. 20142809. DOI: 10.1098/rspb.2014.2809.

Boneh, T., Nicholson, A. E., and Sonenberg, E. A. (2006). "Matilda: A visual tool for modeling with Bayesian networks". In: *International Journal of Intelligent Systems* 21(11), pp. 1127–1150. DOI: 10.1002/int.20175.

Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2004). "A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis". In: *Ecological Modelling* 173(2), pp. 219–239. DOI: 10.1016/j.ecolmodel.2003.08.020.

Bossier, S., Palacz, A. P., Nielsen, J. R., Christensen, A., Hoff, A., Maar, M., Gislason, H., Bastardie, F., Gorton, R., and Fulton, E. A. (2018). "The Baltic Sea Atlantis: An integrated end-to-end modelling framework evaluating ecosystem-wide effects of human-induced pressures". In: *PLOS ONE* 13(7), e0199168. DOI: 10.1371/journal.pone.0199168.

Bulleri, F., Batten, S., Connell, S. D., Benedetti-Cecchi, L., Gibbons, M., Nugues, M. M., and Gribben, P. (2020). "Human pressures and the emergence of

novel marine ecosystems". In: *Oceanography and Marine Biology*. CRC Press, pp. 456–535.

Capuzzo, E., Lynam, C. P., Barry, J., Stephens, D., Forster, R. M., Greenwood, N., McQuatters-Gollop, A., Silva, T., Leeuwen, S. M. v., and Engelhard, G. H. (2018). "A decline in primary production in the North Sea over 25 years, associated with reductions in zooplankton abundance and fish stock recruitment". In: *Global Change Biology* 24(1), e352–e364. DOI: 10.1111/gcb.13916.

Ceccon, S., Garway-Heath, D., Crabb, D., and Tucker, A. (2011). "The dynamic stage bayesian network: identifying and modelling key stages in a temporal process". In: Proceedings of the International Symposium on Intelligent Data Analysis. Springer, pp. 101–112. DOI: 10.1007/978-3-642-24800-9_12.

Chandola, V., Banerjee, A., and Kumar, V. (2009). "Anomaly detection: A survey". In: *ACM Computing Surveys (CSUR)* 41(3), p. 15. DOI: 10.1145/1541880.1541882.

Chee, Y. E., Wilkinson, L., Nicholson, A. E., Quintana-Ascencio, P. F., Fauth, J. E., Hall, D., Ponzio, K. J., and Rumpff, L. (2016). "Modelling spatial and temporal changes with GIS and Spatial and Dynamic Bayesian Networks". In: *Environmental Modelling & Software* 82, pp. 108–120. DOI: 10.1016/j.envsoft.2016.04.012.

Chen, Y.-C., Wheeler, T. A., and Kochenderfer, M. J. (2017). "Learning Discrete Bayesian Networks from Continuous Data". In: *Journal of Artificial Intelligence Research* 59, pp. 103–132. DOI: 10.1613/jair.5371.

Chen, S. H. and Pollino, C. A. (2012). "Good practice in Bayesian network modelling". In: *Environmental Modelling & Software* 37, pp. 134–145. DOI: 10.1016/j.envsoft.2012.03.012.

Chickering, D. M. (2002). "Optimal Structure Identification With Greedy Search". In: *Journal of Machine Learning Research* 3 (Nov), pp. 507–554.

Chow, C. and Liu, C. (1968). "Approximating discrete probability distributions with dependence trees". In: *IEEE Transactions on Information Theory* 14(3), pp. 462–467. DOI: 10.1109/TIT.1968.1054142.

Chu-Agor, M. L., Muñoz-Carpena, R., Kiker, G., Emanuelsson, A., and Linkov, I. (2011). "Exploring vulnerability of coastal habitats to sea level rise through global sensitivity and uncertainty analyses". In: *Environmental Modelling & Software* 26(5), pp. 593–604. DOI: 10.1016/j.envsoft.2010.12.003.

Conversi, A., Umani, S. F., Peluso, T., Molinero, J. C., Santojanni, A., and Edwards, M. (2010). "The Mediterranean Sea regime shift at the end of the

1980s, and intriguing parallelisms with other European basins". In: *PLOS ONE* 5(5), e10633. DOI: 10.1371/journal.pone.0010633.

Cooper, G. F. and Herskovits, E. (1992). "A Bayesian method for the induction of probabilistic networks from data". In: *Machine Learning* 9(4), pp. 309–347. DOI: 10.1007/BF00994110.

Curtin, R. and Prellezo, R. (2010). "Understanding marine ecosystem based management: A literature review". In: *Marine Policy* 34(5), pp. 821–830. DOI: 10.1016/j.marpol.2010.01.003.

Das, M., Ghosh, S. K., Gupta, P., Chowdary, V. M., Nagaraja, R., and Dadhwal, V. K. (2017). "FORWARD: a model for forecasting reservoir water dynamics using spatial Bayesian network (SpaBN)". In: *IEEE Transactions on Knowledge and Data Engineering* 29(4), pp. 842–855.

Daskalov, G. M., Grishin, A. N., Rodionov, S., and Mihneva, V. (2007). "Trophic cascades triggered by overfishing reveal possible mechanisms of ecosystem regime shifts". In: *Proceedings of the National Academy of Sciences* 104(25), pp. 10518–10523. DOI: 10.1073/pnas.0701100104.

David, P., Thebault, E., Anneville, O., Duyck, P.-F., Chapuis, E., and Loeuille, N. (2017). "Impacts of invasive species on food webs: a review of empirical data". In: *Advances in Ecological Research*. Vol. 56. Elsevier, pp. 1–60.

Dean, T. and Kanazawa, K. (1989). "A model for reasoning about persistence and causation". In: *Computational Intelligence* 5(2), pp. 142–150. DOI: 10.1111/j.1467-8640.1989.tb00324.x.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1), pp. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Duarte, C. M., Conley, D. J., Carstensen, J., and Sánchez-Camacho, M. (2009). "Return to Neverland: Shifting Baselines Affect Eutrophication Restoration Targets". In: *Estuaries and Coasts* 32(1), pp. 29–36. DOI: 10.1007/s12237-008-9111-2.

Eero, M., MacKenzie, B. R., Köster, F. W., and Gislason, H. (2011). "Multi-decadal responses of a cod (Gadus morhua) population to human-induced trophic changes, fishing, and climate". In: *Ecological Applications* 21(1), pp. 214–226. DOI: 10.1890/09-1879.1.

Fayyad, U. and Irani, K. (1993). "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning". In: *JPL TRS 1992+*.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Ap-*

proaches to Principles for AI. SSRN Scholarly Paper ID 3518482. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3518482.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). "Bayesian Network Classifiers". In: *Machine Learning* 29, pp. 131–163. DOI: 10.1023/A:1007465528199.

Friedman, N. (1997). "Learning belief networks in the presence of missing values and hidden variables". In: Proceedings of the Fourteenth International Conference on Machine Learning. Vol. 97. July, pp. 125–133.

Fulton, E. A., Smith, A. D., and Johnson, C. R. (2003). "Effect of complexity on marine ecosystem models". In: *Marine Ecology Progress Series* 253, pp. 1–16. DOI: 10.3354/meps253001.

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). "Learning with Drift Detection". In: *Proceedings of the Advances in Artificial Intelligence – SBIA 2004*. Ed. by Bazzan, A. L. C. and Labidi, S. Berlin, Heidelberg: Springer, pp. 286–295. DOI: 10.1007/978-3-540-28645-5_29.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. CRC Press. 505 pp.

Grenci, G., Bertocchi, C., and Ravasio, A. (2019). "Integrating Microfabrication into Biological Investigations: the Benefits of Interdisciplinarity". In: *Micromachines* 10(4), p. 252. DOI: 10.3390/mi10040252.

Grêt-Regamey, A. and Straub, D. (2006). "Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS". In: *Natural Hazards and Earth System Sciences* 6(6), pp. 911–926. DOI: 10.5194/nhess-6-911-2006.

Griffith, G. P., Fulton, E. A., and Richardson, A. J. (2011). "Effects of fishing and acidification-related benthic mortality on the southeast Australian marine ecosystem". In: *Global Change Biology* 17(10), pp. 3058–3074. DOI: 10.1111/j.1365-2486.2011.02453.x.

Guo, K., Zhang, X., Kuai, X., Wu, Z., Chen, Y., and Liu, Y. (2020). "A spatial bayesian-network approach as a decision-making tool for ecological-risk prevention in land ecosystems". In: *Ecological Modelling* 419, p. 108929. DOI: 10.1016/j.ecolmodel.2019.108929.

Haapasaari, P., Kulmala, S., and Kuikka, S. (2012). "Growing into interdisciplinarity: how to converge biology, economics and social science in fisheries research?" In: *Ecology and Society* 17(1), p. 6.

Haataja, M., Fliert, L. van de, and Rautio, P. (2020). *Public AI Registers: Realising AI transparency and civic participation in government use of AI. Whitepaper Version 1.0*.

Halpern, B. S., Frazier, M., Potapenko, J., Casey, K. S., Koenig, K., Longo, C., Lowndes, J. S., Rockwood, R. C., Selig, E. R., Selkoe, K. A., and Walbridge, S. (2015). "Spatial and temporal changes in cumulative human impacts on the world's ocean". In: *Nature Communications* 6(1), p. 7615. DOI: 10.1038/ncomms8615.

Hanea, A., McBride, M., Burgman, M., and Wintle, B. (2018). "Classical meets modern in the IDEA protocol for structured expert judgement". In: *Journal of Risk Research* 21(4), pp. 417–433. DOI: 10.1080/13669877.2016.1215346.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., Twardy, C., Manning, B., and Mascaro, S. (2017). "I nvestigate D iscuss E stimate A ggregate for structured expert judgement". In: *International Journal of Forecasting* 33(1), pp. 267–279. DOI: 10.1016/j.ijforecast.2016.02.008.

Harding, K. C. and Härkönen, T. J. (1999). "Development in the Baltic Grey Seal (Halichoerus grypus) and Ringed Seal (Phoca hispida) Populations during the 20th Century". In: *AMBIO* 28(7), pp. 619–627.

Harwood, N. (2020). "Using Bayesian Networks to Investigate the Role of Arctic Variability in Midlatitude Circulation". PhD Thesis. Brunel University London.

Heiskanen, A.-S., Berg, T., Uusitalo, L., Teixeira, H., Bruhn, A., Krause-Jensen, D., Lynam, C. P., Rossberg, A. G., Korpinen, S., Uyarra, M. C., and Borja, A. (2016). "Biodiversity in Marine Ecosystems—European Developments toward Robust Assessments". In: *Frontiers in Marine Science* 3(184). DOI: 10.3389/fmars.2016.00184.

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., and Wintle, B. C. (2018). "A practical guide to structured expert elicitation using the IDEA protocol". In: *Methods in Ecology and Evolution* 9(1), pp. 169–180. DOI: 10.1111/2041-210X.12857.

Hilborn, R. and Mangel, M. (1997). *The Ecological Detective: confronting models with data*. Princeton University Press.

Hillebrand, H., Donohue, I., Harpole, W. S., Hodapp, D., Kucera, M., Lewandowska, A. M., Merder, J., Montoya, J. M., and Freund, J. A. (2020). "Thresholds for ecological responses to global change do not emerge from empirical data". In: *Nature Ecology & Evolution*, pp. 1–8. DOI: 10.1038/s41559-020-1256-9.

Howard, R. A. (1965). "Bayesian decision models for system engineering". In: *IEEE Transactions on Systems Science and Cybernetics* 1(1), pp. 36–40. DOI: 10.1109/TSSC.1965.300058.

Howard, R. A. and Matheson, J. E. (1981). "Influence diagrams". In: *Readings in Decision Analysis*. Vol. 2. 3, pp. 763–771.

Hsieh, C.-h., Anderson, C., and Sugihara, G. (2008). "Extending Nonlinear Analysis to Short Ecological Time Series." In: *The American Naturalist* 171(1), pp. 71–80. DOI: 10.1086/524202.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 560–575. DOI: 10.1145/3442188.3445918.

Jackson, L. J., Trebitz, A. S., and Cottingham, K. L. (2000). "An Introduction to the Practice of Ecological Modeling". In: *BioScience* 50(8), pp. 694–706. DOI: 10.1641/0006-3568(2000)050[0694:AITTPO]2.0.CO;2.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Series for Statistics for Engineering and Information Science. Springer.

Jiang, P., Liu, X., Zhang, J., Te, S. H., Gin, K. Y.-H., Fan, Y. V., Klemeš, J. J., and Shoemaker, C. A. (2021). "Cyanobacterial risk prevention under global warming using an extended Bayesian network". In: *Journal of Cleaner Production* 312, p. 127729. DOI: 10.1016/j.jclepro.2021.127729.

Jormalainen, V., Gagnon, K., Sjöroos, J., and Rothäusler, E. (2016). "The invasive mud crab enforces a major shift in a rocky littoral invertebrate community of the Baltic Sea". In: *Biological Invasions* 18(5), pp. 1409–1419. DOI: 10.1007/s10530-016-1090-9.

Julia Flores, M., Nicholson, A. E., Brunskill, A., Korb, K. B., and Mascaro, S. (2011). "Incorporating expert knowledge when learning Bayesian network structure: A medical case study". In: *Artificial Intelligence in Medicine* 53(3), pp. 181–204. DOI: 10.1016/j.artmed.2011.08.004.

Jüssi, M., Härkönen, T., Helle, E., and Jüssi, I. (2008). "Decreasing ice cover-age will reduce the breeding success of Baltic grey seal (Halichoerus grypus) females". In: *AMBIO* 37(2), pp. 80–86.

Kaikkonen, L., Parviainen, T., Rahikainen, M., Uusitalo, L., and Lehikoinen, A. (2020). "Bayesian Networks in Environmental Risk Assessment: A review". In: *Integrated Environmental Assessment and Management* 17 (1), pp. 62–78. DOI: 10.1002/ieam.4332.

Kaisler, S., Armour, F., Espinosa, J. A., and Money, W. (2013). "Big Data: Issues and Challenges Moving Forward". In: *Proceedings of the 2013 46th Hawaii International Conference on System Sciences*. ISSN: 1530-1605, pp. 995–1004. DOI: 10.1109/HICSS.2013.645.

Katsanevakis, S., Wallentinus, I., Zenetos, A., Leppäkoski, E., Çinar, M. E., Oztürk, B., Grabowski, M., Golani, D., and Cardoso, A. C. (2014). "Impacts of invasive alien marine species on ecosystem services and biodiversity: a pan-European review". In: *Aquatic Invasions* 9(4), pp. 391–423.

Kelly, R. A., Jakeman, A. J., Barreteau, O., Borsuk, M. E., ElSawah, S., Hamilton, S. H., Henriksen, H. J., Kuikka, S., Maier, H. R., Rizzoli, A. E., Delden, H., and Voinov, A. A. (2013). "Selecting among five common modelling approaches for integrated environmental assessment and management". In: *Environmental Modelling & Software* 47, pp. 159–181. DOI: 10.1016/j.envsoft.2013.05.005.

Kim, B. and Doshi-Velez, F. (2021). "Machine Learning Techniques for Accountability". In: *AI Magazine* 42(1), pp. 47–52.

Kjærulff, U. (1992). "A computational scheme for reasoning in dynamic probabilistic networks". In: *Uncertainty in Artificial Intelligence*. Ed. by Dubois, D., Wellman, M. P., D'Ambrosio, B., and Smets, P. Morgan Kaufmann, pp. 121–129. DOI: https://doi.org/10.1016/B978-1-4832-8287-9.50021-9.

Kjaerulff, U. and Madsen, A. (2008). *Bayesian networks and influence diagrams*. Science+ Business Media. Springer.

Korb, K. and Nicholson, A. (2011). *Bayesian Artificial Intelligence*. Second Edition. Computer Science and Data Analysis Series. Chapman & Hall.

Kuikka, S. and Varis, O. (1997). "Uncertainties of climatic change impacts in Finnish watersheds: a Bayesian network analysis of expert knowledge". In: *Boreal Environment Research* 2(1), pp. 109–128.

Kwoh, C.-K. and Gillies, D. F. (1996). "Using hidden nodes in Bayesian networks". In: *Artificial intelligence* 88(1), pp. 1–38. DOI: 10.1016/0004-3702(95)00119-0.

Lauritzen, S. L. (1995). "The EM algorithm for graphical association models with missing data". In: *Computational Statistics & Data Analysis* 19(2), pp. 191–201. DOI: 10.1016/0167-9473(93)E0056-A.

Lehikoinen, A., Helle, I., Klemola, E., Mäntyniemi, S., Kuikka, S., and Pitkänen, H. (2014). "Evaluating the impact of nutrient abatement measures on the ecological status of coastal waters: a Bayesian network for decision analysis". In: *International Journal of Multicriteria Decision Making* 4(2), pp. 114–134. DOI: 10.1504/IJMCDM.2014.060426.

Lehtiniemi, M., Outinen, O., and Puntila-Dodd, R. (2020). "Citizen science provides added value in the monitoring for coastal non-indigenous species". In: *Journal of Environmental Management* 267, p. 110608. DOI: 10.1016/j.jenvman.2020.110608.

Liu, H., Hussain, F., Tan, C. L., and Dash, M. (2002). "Discretization: An Enabling Technique". In: *Data Mining and Knowledge Discovery* 6(4), pp. 393–423. DOI: 10.1023/A:1016304305535.

Liu, Y., Zhang, H. H., and Wu, Y. (2011). "Hard or Soft Classification? Large-Margin Unified Machines". In: *Journal of the American Statistical Association* 106(493), pp. 166–177. DOI: 10.1198/jasa.2011.tm10319.

Lye, A., Cicirello, A., and Patelli, E. (2022). "An efficient and robust sampler for Bayesian inference: Transitional Ensemble Markov Chain Monte Carlo". In: *Mechanical Systems and Signal Processing* 167, p. 108471. DOI: 10.1016/j.ymssp.2021.108471.

Mack, L., Attila, J., Aylagas, E., Beermann, A., Borja, A., Hering, D., Kahlert, M., Leese, F., Lenz, R., Lehtiniemi, M., Liess, A., Lips, U., Mattila, O.-P., Meissner, K., Pyhälahti, T., Setälä, O., Strehse, J. S., Uusitalo, L., Willstrand Wranne, A., and Birk, S. (2020). "A Synthesis of Marine Monitoring Methods With the Potential to Enhance the Status Assessment of the Baltic Sea". In: *Frontiers in Marine Science* 7. DOI: 10.3389/fmars.2020.552047.

Madsen, A. L., Jensen, F., Kjærulff, U. B., and Lang, M. (2005). "The hugin tool for probabilistic graphical models". In: *International Journal on Artificial Intelligence Tools* 14(3), pp. 507–543. DOI: 10.1142/S0218213005002235.

Mäkinen, K., Vuorinen, I., and Hänninen, J. (2017). "Climate-induced hydrography change favours small-bodied zooplankton in a coastal ecosystem". In: *Hydrobiologia* 792(1), pp. 83–96. DOI: 10.1007/s10750-016-3046-6.

Margossian, C., Vehtari, A., Simpson, D., and Agrawal, R. (2020). "Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond". In: *Advances in Neural Information Processing Systems* 33, pp. 9086–9097.

Meier, H. E. M., Edman, M., Eilola, K., Placke, M., Neumann, T., Andersson, H. C., Brunnabend, S.-E., Dieterich, C., Frauen, C., Friedland, R., Gröger, M., Gustafsson, B. G., Gustafsson, E., Isaev, A., Kniebusch, M., Kuznetsov, I., Müller-Karulis, B., Naumann, M., Omstedt, A., Ryabchenko, V., Saraiva, S., and Savchuk, O. P. (2019). "Assessment of Uncertainties in Scenario Simulations of Biogeochemical Cycles in the Baltic Sea". In: *Frontiers in Marine Science* 6. DOI: 10.3389/fmars.2019.00046.

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34(3), pp. 1436–1462. DOI: 10.1214/009053606000000281.

Möllmann, C., Müller-Karulis, B., Kornilovs, G., and St John, M. A. (2008). "Effects of climate and overfishing on zooplankton dynamics and ecosystem structure: regime shifts, trophic cascade, and feedback loops in a simple ecosystem". In: *ICES Journal of Marine Science* 65(3), pp. 302–310. DOI: 10.1093/icesjms/fsm197.

Morgan, M. G. and Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press.

Murphy, K. (2012). *Machine learning: a probabilistic perspective*. Red. by Dietterich, T. Adaptive Computation and Machine Learning. The MIT Press.

Murphy, K. P. (2002). "Dynamic bayesian networks: representation, inference and learning". PhD thesis. University of California, Berkeley.

Nelson, G. C., Bennett, E., Berhe, A. A., Cassman, K., DeFries, R., Dietz, T., Dobermann, A., Dobson, A., Janetos, A., Levy, M., Marco, D., Nakicenovic, N., O'Neill, B., Norgaard, R., Petschel-Held, G., Ojima, D., Pingali, P., Watson, R., and Zurek, M. (2006). "Anthropogenic Drivers of Ecosystem Change: an Overview". In: *Ecology and Society* 11(2).

Nicholson, A. E. (1992). "Monitoring discrete environments using dynamic belief networks." PhD thesis. University of Oxford.

Nojavan, F., Qian, S., and Stow, C. (2017). "Comparative analysis of discretization methods in Bayesian networks". In: *Environmental Modelling & Software* 87, pp. 64–71. DOI: 10.1016/j.envsoft.2016.10.007.

Norkko, J., Reed, D. C., Timmermann, K., Norkko, A., Gustafsson, B. G., Bonsdorff, E., Slomp, C. P., Carstensen, J., and Conley, D. J. (2012). "A welcome can of worms? Hypoxia mitigation by an invasive species". In: *Global Change Biology* 18(2), pp. 422–434. DOI: 10.1111/j.1365-2486.2011.02513.x.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. Statistics in practice. John Wiley & Sons.

O'Hagan, A. (2012). "Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux". In: *Environmental Modelling & Software*. Thematic issue on Expert Opinion in Environmental Modelling and Management 36, pp. 35–48. DOI: 10.1016/j.envsoft.2011.03.003.

Pearl, J. (1986). "Fusion, Propagation, and Structuring in Belief Networks". In: *Artificial Intelligence* 29(3), pp. 241–288. DOI: 10.1016/0004-3702(86)90072-X.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press. 384 pp.

Pearl, J. (2010). "Causal Inference". In: *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*. Proceedings of Machine Learning Research, pp. 39–58.

Pollino, C. A., Woodberry, O., Nicholson, A., Korb, K., and Hart, B. T. (2007). "Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment". In: *Environmental Modelling & Software* 22(8), pp. 1140–1152. DOI: 10.1016/j.envsoft.2006.03.006.

Rachid, G., Alameddine, I., Najm, M. A., Qian, S., and El-Fadel, M. (2021). "Dynamic Bayesian Networks to Assess Anthropogenic and Climatic Drivers of Saltwater Intrusion: A Decision Support Tool Toward Improved Management". In: *Integrated Environmental Assessment and Management* 17(1), pp. 202–220. DOI: 10.1002/ieam.4355.

Regan, H. M., Colyvan, M., and Burgman, M. A. (2002). "A taxonomy and treatment of uncertainty for ecology and conservation biology". In: *Ecological Applications* 12(2), pp. 618–628. DOI: 10.2307/3060967.

Robinson, J. W. and Hartemink, A. J. (2009). "Non-stationary dynamic Bayesian networks". In: Proceedings of the Advances in neural information processing systems, pp. 1369–1376.

Ropero, R. F., Renooij, S., and Gaag, L. C. van der (2018). "Discretizing environmental data for learning Bayesian-network classifiers". In: *Ecological Modelling* 368, pp. 391–403. DOI: 10.1016/j.ecolmodel.2017.12.015.

Sáez, C., Rodrigues, P. P., Gama, J., Robles, M., and García-Gómez, J. M. (2015). "Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality". In: *Data Mining and Knowledge Discovery* 29(4), pp. 950–975. DOI: 10.1007/s10618-014-0378-6.

Sahin, O., Stewart, R. A., Faivre, G., Ware, D., Tomlinson, R., and Mackey, B. (2019). "Spatial Bayesian Network for predicting sea level rise induced coastal erosion in a small Pacific Island". In: *Journal of environmental management* 238, pp. 341–351. DOI: 10.1016/j.jenvman.2019.03.008.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index". In: *Computer Physics Communications* 181(2), pp. 259–270. DOI: 10.1016/j.cpc.2009.09.018.

Saraiva, S., Meier, H. E. M., Andersson, H., Höglund, A., Dieterich, C., Gröger, M., Hordoir, R., and Eilola, K. (2019). "Uncertainties in Projections of the Baltic Sea Ecosystem Driven by an Ensemble of Global Climate Models". In: *Frontiers in Earth Science* 6. DOI: 10.3389/feart.2018.00244.

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Kuemmerlen, M., Langhans, S. D., Martínez-López, J., and Vermeiren, P. (2019). "How to make ecological models useful for environmental management". In: *Ecological Modelling* 411, p. 108784. DOI: 10.1016/j.ecolmodel.2019.108784.

Scutari, M., Graafland, C. E., and Gutiérrez, J. M. (2019). "Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms". In: *International Journal of Approximate Reasoning* 115, pp. 235–253. DOI: 10.1016/j.ijar.2019.10.003.

Scutari, M. and Strimmer, K. (2011). "Introduction to Graphical Modelling". In: *arXiv:1005.1036 [math, stat]*. DOI: 10.48550/arXiv.1005.1036. arXiv: 1005.1036.

Scutari, M., Vitolo, C., and Tucker, A. (2019). "Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation". In: *Statistics and Computing* 29(5), pp. 1095–1108. DOI: 10.1007/s11222-019-09857-1.

Shin, Y.-J. and Shannon, L. J. (2010). "Using indicators for evaluating, comparing, and communicating the ecological status of exploited marine ecosystems. 1. The IndiSeas project". In: *ICES Journal of Marine Science* 67(4), pp. 686–691. DOI: 10.1093/icesjms/fsp273.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). "Visualizing Uncertainty About the Future". In: *Science* 333(6048), pp. 1393–1400. DOI: 10.1126/science.1191181.

Sprites, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction and search*. Vol. 81. Lecture Notes in Statistics. Springer Verlag.

Stelzenmüller, V., Lee, J., Garnacho, E., and Rogers, S. (2010). "Assessment of a Bayesian Belief Network–GIS framework as a practical tool to support marine planning". In: *Marine Pollution Bulletin* 60(10), pp. 1743–1754. DOI: 10.1016/j.marpolbul.2010.06.024.

Stritih, A., Rabe, S.-E., Robaina, O., Grêt-Regamey, A., and Celio, E. (2020). "An online platform for spatial and iterative modelling with Bayesian Networks". In: *Environmental Modelling & Software* 127, p. 104658. DOI: 10.1016/j.envsoft.2020.104658.

Suikkanen, S., Pulina, S., Engstrom-Ost, J., Lehtiniemi, M., Lehtinen, S., and Brutemark, A. (2013). "Climate change and eutrophication induced shifts in northern summer plankton communities". In: *PLOS ONE* 8(6), e66475. DOI: 10.1371/journal.pone.0066475.

Teixeira, H., Berg, T., Uusitalo, L., Fürhaupter, K., Heiskanen, A.-S., Mazik, K., Lynam, C. P., Neville, S., Rodriguez, J. G., Papadopoulou, N., Moncheva, S., Churilova, T., Kryvenko, O., Krause-Jensen, D., Zaiko, A., Veríssimo, H., Pantazi, M., Carvalho, S., Patrício, J., Uyarra, M. C., and Borja, À. (2016). "A Catalogue of Marine Biodiversity Indicators". In: *Frontiers in Marine Science* 3. DOI: 10.3389/fmars.2016.00207.

Tomassini, L., Reichert, P., Knutti, R., Stocker, T. F., and Borsuk, M. E. (2007). "Robust Bayesian Uncertainty Analysis of Climate System Properties Using Markov Chain Monte Carlo Methods". In: *Journal of Climate* 20(7), pp. 1239–1254. DOI: 10.1175/JCLI4064.1.

Trifonova, N., Karnauskas, M., and Kelble, C. (2019). "Predicting ecosystem components in the Gulf of Mexico and their responses to climate variability with a dynamic Bayesian network model". In: *PLOS ONE* 14(1), e0209257. DOI: 10.1371/journal.pone.0209257.

Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., and Tucker, A. (2015). "Spatio-temporal Bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology". In: *Ecological Informatics* 30, pp. 142–158. DOI: 10.1016/j.ecoinf.2015.10.003.

Trifonova, N., Maxwell, D., Pinnegar, J., Kenny, A., and Tucker, A. (2017). "Predicting ecosystem responses to changes in fisheries catch, temperature, and primary productivity with a dynamic Bayesian network model". In: *ICES Journal of Marine Science* 74(5), pp. 1334–1343. DOI: 10.1093/icesjms/fsw231.

Trifonova, N., Scott, B., De Dominicis, M., Waggitt, J., and Wolf, J. (2021). "Bayesian network modelling provides spatial and temporal understanding of ecosystem dynamics within shallow shelf seas". In: *Ecological Indicators* 129, p. 107997. DOI: 10.1016/j.ecolind.2021.107997.

Tucker, A. and Liu, X. (2004). "A Bayesian network approach to explaining time series with changing structure". In: *Intelligent Data Analysis* 8(5), pp. 469–480. DOI: 10.3233/IDA-2004-8504.

Uusitalo, L. (2007). "Advantages and challenges of Bayesian networks in environmental modelling". In: *Ecological Modelling* 203(3), pp. 312–318. DOI: 10.1016/j.ecolmodel.2006.11.033.

Uusitalo, L., Kuikka, S., and Romakkaniemi, A. (2005). "Estimation of Atlantic salmon smolt carrying capacity of rivers using expert knowledge". In: *ICES Journal of Marine Science* 62(4), pp. 708–722. DOI: 10.1016/j.icesjms.2005.02.005.

Uusitalo, L., Blanchet, H., Andersen, J. H., Beauchard, O., Berg, T., Bianchelli, S., Cantafaro, A., Carstensen, J., Carugati, L., Cochrane, S., Danovaro, R., Heiskanen, A.-S., Karvinen, V., Moncheva, S., Murray, C., Neto, J. M., Nygård, H., Pantazi, M., Papadopoulou, N., Simboura, N., Srėbalienė, G., Uyarra, M. C., and Borja, A. (2016). "Indicator-Based Assessment of Marine Biological Diversity–Lessons from 10 Case Studies across the European Seas". In: *Frontiers in Marine Science* 3(159). DOI: 10.3389/fmars.2016.00159.

Wu, P. P.-Y., Julian Caley, M., Kendrick, G. A., McMahon, K., and Mengersen, K. (2018). "Dynamic Bayesian network inferencing for non-homogeneous complex systems". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67(2), pp. 417–434. DOI: 10.1111/rssc.12228.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., Heikkinen, R. K., Heinänen, S., Jones, A. R., Krishnakumar, P. K., Lauria, V., Lozano-Montes, H., Mannocci, L., Mellin, C., Mesgaran, M. B., Moreno-Amat, E., Mormede, S., Novaczek, E., Oppel, S., Ortuño Crespo, G., Peterson, A. T., Rapacciuolo, G., Roberts, J. J., Ross, R. E., Scales, K. L., Schoeman, D., Snelgrove, P., Sundblad, G., Thuiller, W., Torres, L. G., Verbruggen, H., Wang, L., Wenger, S., Whittingham, M. J., Zharikov, Y., Zurell, D., and Sequeira, A. M. M. (2018). "Outstanding Challenges in the Transferability of Ecological Models". In: *Trends in Ecology & Evolution* 33(10), pp. 790–802. DOI: 10.1016/j.tree.2018.08.001.

Zuur, A., Ieno, E., and Smith, G. (2007). *Analysing Ecological Data*. Statistics for Biology and Health. Springer.