Institut für Erd- und Umweltwissenschaften
Allgemeine Geophysik

Applications of Bayesian Networks in Natural Hazard Assessments

Kumulative Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Allgemeine Geophysik"

eingereicht an der
Mathematischen-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Diplom-Mathematikerin
Kristin Vogel

Potsdam, September 2013

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2014/6977/
URN urn:nbn:de:kobv:517-opus-69777
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-69777

Sicher ist, dass nichts sicher ist.

Selbst das nicht.

Joachim Ringelnatz

Allgemeinverständliche Zusammenfassung

Obwohl Naturgefahren in ihren Ursachen, Erscheinungen und Auswirkungen grundlegend verschieden sind, teilen sie doch viele Gemeinsamkeiten und Herausforderungen, wenn es um ihre Modellierung geht. Fehlendes Wissen über die zugrunde liegenden Kräfte und deren komplexes Zusammenwirken erschweren die Wahl einer geeigneten Modellstruktur. Hinzu kommen ungenaue und unvollständige Beobachtungsdaten sowie dem Naturereignis innewohnende Zufallsprozesse. All diese verschiedenen, miteinander interagierende Aspekte von Unsicherheit erfordern eine sorgfältige Betrachtung, um fehlerhafte und verharmlosende Einschätzungen von Naturgefahren zu vermeiden. Dennoch sind deterministische Vorgehensweisen in Gefährdungsanalysen weit verbreitet.

Bayessche Netze betrachten die Probleme aus wahrscheinlichkeitstheoretischer Sicht und bieten somit eine sinnvolle Alternative zu deterministischen Verfahren. Alle vom Zufall beeinflussten Größen werden hierbei als Zufallsvariablen angesehen. Die gemeinsame Wahrscheinlichkeitsverteilung aller Variablen beschreibt das Zusammenwirken der verschiedenen Einflussgrößen und die zugehörige Unsicherheit/Zufälligkeit. Die Abhängigkeitsstrukturen der Variablen können durch eine grafische Darstellung abgebildet werden. Die Variablen werden dabei als Knoten in einem Graphen/Netzwerk dargestellt und die (Un-)Abhängigkeiten zwischen den Variablen als (fehlende) Verbindungen zwischen diesen Knoten. Die dargestellten Unabhängigkeiten veranschaulichen, wie sich die gemeinsame Wahrscheinlichkeitsverteilung in ein Produkt lokaler, bedingter Wahrscheinlichkeitsverteilungen zerlegen lässt.

Im Verlauf dieser Arbeit werden verschiedene Naturgefahren (Erdbeben, Hochwasser und Bergstürze) betrachtet und mit Bayesschen Netzen modelliert. Dazu wird jeweils nach der Netzwerkstruktur gesucht, welche die Abhängigkeiten der Variablen am besten beschreibt. Außerdem werden die Parameter der lokalen, bedingten Wahrscheinlichkeitsverteilungen geschätzt, um das Bayessche Netz und dessen zugehörige gemeinsame Wahrscheinlichkeitsverteilung vollständig zu bestimmen. Die Definition des Bayesschen Netzes kann auf Grundlage von Expertenwissen erfolgen oder – so wie in dieser Arbeit – anhand von Beobachtungsdaten des zu untersuchenden Naturereignisses. Die hier verwendeten Methoden wählen Netzwerkstruktur und Parameter so, dass die daraus resultierende Wahrscheinlichkeitsverteilung den beobachteten Daten eine möglichst große Wahrscheinlichkeit zuspricht. Da dieses Vorgehen keine Expertenwissen voraussetzt, ist es universell in verschiedenen Gebieten der Gefährdungsanalyse einsetzbar.

Trotz umfangreicher Forschung zu diesem Thema ist das Bestimmen von Bayesschen Netzen basierend auf Beobachtungsdaten nicht ohne Schwierigkeiten. Typische Herausforderungen stellen die Handhabung stetiger Variablen und unvollständiger Datensätze dar. Beide Probleme werden in dieser Arbeit behandelt. Es werden Lösungsansätze entwickelt und in den Anwendungsbeispielen eingesetzt. Eine Kernfrage ist hierbei die Komplexität des Algorithmus. Besonders wenn sowohl stetige Variablen als auch unvollständige Datensätze in Kombination auftreten, sind effizient arbeitende Verfahren gefragt. Die hierzu in dieser Arbeit entwickelten Methoden ermöglichen die Verarbeitung von großen Datensätze mit stetigen Variablen und unvollständigen Beobachtungen und leisten damit einen wichtigen Beitrag für die wahrscheinlichkeitstheoretische Gefährdungsanalyse.

Summary

Even though quite different in occurrence and consequences, from a modeling perspective many natural hazards share similar properties and challenges. Their complex nature as well as lacking knowledge about their driving forces and potential effects make their analysis demanding: uncertainty about the modeling framework, inaccurate or incomplete event observations and the intrinsic randomness of the natural phenomenon add up to different interacting layers of uncertainty, which require a careful handling. Nevertheless deterministic approaches are still widely used in natural hazard assessments, holding the risk of underestimating the hazard with disastrous effects. The all-round probabilistic framework of Bayesian networks constitutes an attractive alternative. In contrast to deterministic proceedings, it treats response variables as well as explanatory variables as random variables making no difference between input and output variables. Using a graphical representation Bayesian networks encode the dependency relations between the variables in a directed acyclic graph: variables are represented as nodes and (in-)dependencies between variables as (missing) edges between the nodes. The joint distribution of all variables can thus be described by decomposing it, according to the depicted independences, into a product of local conditional probability distributions, which are defined by the parameters of the Bayesian network.

In the framework of this thesis the Bayesian network approach is applied to different natural hazard domains (i.e. seismic hazard, flood damage and landslide assessments). Learning the network structure and parameters from data, Bayesian networks reveal relevant dependency relations between the included variables and help to gain knowledge about the underlying processes. The problem of Bayesian network learning is casted in a Bayesian framework, considering the network structure and parameters as random variables itself and searching for the most likely combination of both, which corresponds to the *maximum a posteriori* (MAP score) of their joint distribution given the observed data. Although well studied in theory the learning of Bayesian networks based on real-world data is usually not straight forward and requires an adoption of existing algorithms. Typically arising problems are the handling of continuous variables, incomplete observations and the interaction of both.

Working with continuous distributions requires assumptions about the allowed families of distributions. To "let the data speak" and avoid wrong assumptions, continuous variables are instead discretized here, thus allowing for a completely data-driven and distribution-free learning. An extension of the MAP score, considering the discretization as random variable as well, is developed for an automatic multivariate discretization, that takes interactions between the variables into account. The discretization process is nested into the network learning and requires several iterations. Having to face incomplete observations on top, this may pose a computational burden. Iterative proceedings for missing value estimation become quickly infeasible. A more efficient albeit approximate method is used instead, estimating the missing values based only on the observations of variables directly interacting with the missing variable. Moreover natural hazard assessments often have a primary interest in a certain target variable. The discretization learned for this variable does not always have the required resolution for a good prediction performance. Finer resolutions for (conditional) continuous distributions are achieved with continuous approximations subsequent to the Bayesian network learning, using kernel density estimations or mixtures of truncated exponential functions.

All our proceedings are completely data-driven. We thus avoid assumptions that require expert knowledge and instead provide domain independent solutions, that are applicable not only in other natural hazard assessments, but in a variety of domains struggling with uncertainties.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

ONE

INTRODUCTION

Earthquakes, tsunamis, flood events, landslides, volcanic eruptions – they all and many other natural hazards are quite different in their causes and effects, but from a modeling perspective they share a lot of common properties and challenges. Their underlying processes are complex and not completely understood. The number of influencing factors is large and their interactions anything but transparent, which makes an identification of the driving forces and a description of their single and joint effects demanding. This leads to a variety of model structures suggested in literature, revealing a great uncertainty about the framework to use. Additionally, the observations on which the natural hazard analysis is based are often sparse, inaccurate and/or incomplete, which adds another layer of uncertainty on top. Various sources of uncertainty accumulate, each either corresponding to a lack of knowledge, the *epistemic uncertainty*, or to the intrinsic and irreducible *aleatoric uncertainty*, which comes about the randomness of the natural phenomenon under study. Ignoring those uncertainties may have disastrous effects, since it often leads to an underestimation of the hazard.

Nevertheless deterministic approaches are widely used in natural hazard assessments. Tsunami early warning systems, e.g, evaluate pre-calculated synthetic databases and pick out the scenario 'closest' to the current situation to estimate its hazard (Blaser et al., 2011). Recently developed models for flood damage assessments (i.e., the FLEMOps+r model) use classification approaches, where the event under consideration is assigned to its corresponding class and the caused damage is estimated by taking the mean damage of all observed events that belong to the same class (Elmer et al., 2010). In seismic hazard analysis the usage of regression-based ground motion models is common practice, restricting the model to the chosen functional form, which is defined based on physical constrains (Kühn et al., 2011). Deterministic approaches provide usually no information or hardly any inside into the uncertainty related to their estimates, but uncertainty is carrier of information and ignoring it as some sort of error would be wrong.

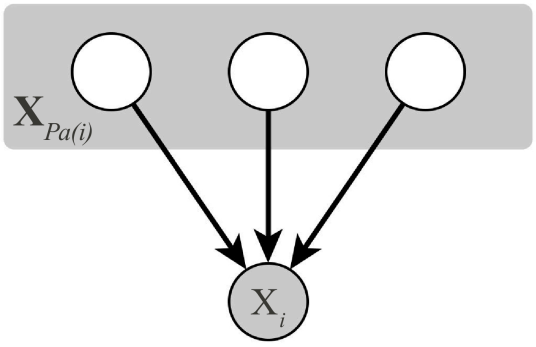
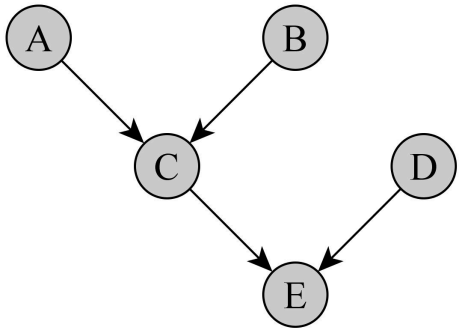


Figure 1.1: Illustration of a parent set $\mathbf{X}_{Pa(i)}$ of X_i.



$$P(A, B, C, D, E) = \mathsf{p}(E|C,D)\mathsf{p}(D)\mathsf{p}(C|A,B)\mathsf{p}(B)\mathsf{p}(A)$$

Figure 1.2: Example for the decomposition of the joint distribution according to a *DAG*.

Directed graphical models (DGMs), in particular Bayesian networks, pose a powerful formalism to capture and express uncertainties. In recent years they have successfully been employed in a wide range of earth science applications, including tsunami early warning, e.g. (Blaser et al., 2011), probabilistic seismic hazard analysis, e.g. (Kühn et al., 2011), and automatic detection and classification of seismic signals, e.g. (Riggelsen et al., 2007). DGMs treat all random quantities that pertain to a particular hazard domain as random variables, which are represented as nodes in a directed acyclic graph (DAG).

In the following random variables are indicated by uppercases (usually X_i), while vectors/sets of random variables are indicated by bold uppercases (usually $\mathbf{X}$). Realizations of the random variables are indicated by lowercases accordingly (x_i or $\mathbf{x}$). If not defined otherwise the set of all considered random variables is denoted by $\mathbf{X} = \{X_1, \ldots, X_k\}$. The dependency relations of the variables are encoded through the DAG structure, where arcs point from the variables in the parent set, $\mathbf{X}_{Pa(i)}$, to X_i (see Fig. 1.1), stating that X_i directly depends on $\mathbf{X}_{Pa(i)}$. Each random variable is associated with a conditional distribution $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$ and the joint distribution of all variables decomposes according to the DAG into a product of the conditionals: $P(\mathbf{X}) = \prod_i \mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$. Figure 1.2 shows an example. In the framework of this thesis we can only give a short introduction into DGMs, but there exist several textbooks on the topic, e.g. (Jensen and Nielsen, 2001; Koller and Friedman, 2009).

We consider three types of DGMs. The main focus is on the application of Bayesian networks (BNs), which treat all variables equally. Their only restriction on the graph structure is to form a directed acyclic graph. This allows for a construction of dependency relations that (are close to) reflect the reality and give insight into the underlying system. A BN is fully described by its DAG and its parameters, $\boldsymbol{\theta}$, that define the conditional distributions. For discrete variables the set of parameters corresponds to the conditional (point) probabilities for each combination of states: $\boldsymbol{\theta} = \bigcup\{\theta_{x_i|\mathbf{x}_{Pa(i)}} = \mathsf{p}(x_i|\mathbf{x}_{Pa(i)})\}$. For continuous variables depends the design of the parameters on the functional form of the conditional distributions.

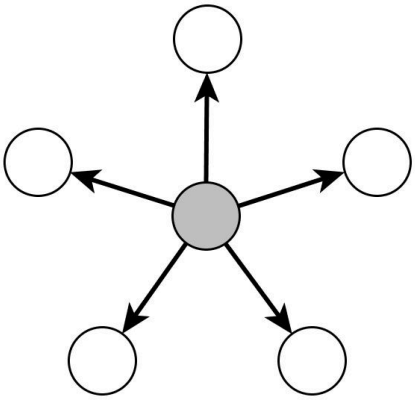


Figure 1.3: Illustration of a Naive Bayes network structure. The gray node indicates the class variable.

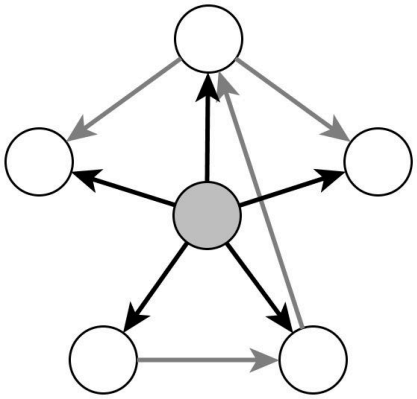


Figure 1.4: Illustration of a Tree Augmented Naive Bayes. The gray node indicates the class variable, the gray edges are the ones added to the Naive Bayes.

In the applications considered in this thesis network structure and parameters are learned from data, searching for the pair $(DAG, \boldsymbol{\theta})$ that is most likely to describe the data generating process. There is no expert knowledge included in the BN learning. Anyhow, if present, expert knowledge can be exploited to define (elements of) the BN or to set up a BN that reflects our prior belief and is updated based on the data.

Many natural hazard assessments have an increased interest in the prediction of a certain target variable. While a BN is designed to capture the joint distribution of all variables, the Naive Bayes classifier focuses on the variable of interest and may thus be more accurate in that region. Its network structure is simple and fixed: The target variable, often referred to as class variable, is the only parent of each other variable, the attributes, and has no parents itself (see Fig. 1.3 for illustration). Even though it does not reflect the real (in-)dependencies, the Naive Bayes usually performs well (competitive with or better than BNs) in classification tasks (Friedman and Goldszmidt, 1996a).

The Tree Augmented Naive Bayes is an extension of the Naive Bayes approach. It allows to assign one more parent, in addition to the class variable, to each attribute (see Fig. 1.4 for illustration). Maintaining the computational simplicity the Tree Augmented Naive Bayes classifier thus relaxes the independence assumptions made for Naive Bayes, but is usually still far from describing the real dependency relations. The improved classification performance thus does not come without costs. Concentrating on the variable of interest the (Tree Augmented) Naive Bayes classifier does not capture the joint distribution of all variables, which is needed to infer into all directions.

Although graphical models are well studied in theory, their application on real-world data is not straight forward. One of the most dominant problems is the handling of continuous variables. The main body of this thesis comprises four papers, where graphical models are applied in the domains of seismic hazard, flood damage and landslides. Each of them requires a treatment of continuous variables. To avoid assumptions on the families of distributions and

allow for a distribution free learning the continuous variables are discretized in all applications. Choosing an 'optimal' discretization that leads to a minimum of information loss is anything but trivial and a major issue in all four papers. The first paper concentrates on the development of surrogates for complex ground motion models used in the probabilistic seismic hazard analysis. The second and third paper investigate a flood damage data set collected after the 2002 and 2005/06 flood events in the Elbe and Danube catchments (Germany). Using the Bayesian network approach they aim to learn about damage causing and preventing factors and the interaction of those. The third paper additionally gives a suggestion how to deal with incomplete observations. The fourth paper adds a landslide analysis to the investigated natural hazard domains and stresses the benefits of the Bayesian network approach for natural hazard assessments. The full papers are reprinted in Chapters 2 to 5. In the following the manuscripts are shortly summarized.

PAPER 1

Graphical Models as Surrogates for Complex Ground Motion Models

Vogel, K., Riggelsen, C., Kühn, N., Scherbaum, F.; 2012. Published in *Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing*

One of the most critical elements in probabilistic seismic hazard analysis is the model that describes the ground motion caused by earthquakes. So-called *stochastic models* capture the characteristics of the ground motion well, but since they do not have nice analytical properties a simplified model is often used instead. This surrogate is usually defined by fitting a regression function to a synthetic data set generated by the stochastic model. This paper presents Directed Graphical Models as an alternative to the regression approach. A Bayesian network, a Naive Bayes and a Tree Augmented Naive Bayes classifier are learned based on a synthetic data set. Continuous variables are discretized for this purpose, using automatic discretization procedures that choose an 'optimal' discretization based on the observed data. In the (Tree Augmented) Naive Bayes approach the attributes are discretized depending on the class variable, using a variation of the class entropy to find a discretization that keeps the information loss small. The class variable itself is not discretized, but approximated with a Gaussian kernel density estimator. In the Bayesian network approach all variables are discretized simultaneously, using a multivariate discretization that takes the interaction of the variables into account. The proceeding is motivated by Monti and Cooper (1998) and is only briefly sketched in this paper. It will be adopted and enhanced in the following papers. Finally the prediction performance of the learned models is compared to the regression approach. To increase the precision of the Bayesian network, the discretization learned for the ground motion variable is ignored and its continuous distribution is, as for the (Tree Augmented) Naive Bayes, approximated with a kernel density estimator. Compared to the regression approach all three graphical models perform well in the ground motion prediction. The best prediction performance is delivered by the Naive Bayes classifier.

PAPER 2

Flood Damage and Influencing Factors: A Bayesian Network Perspective

Vogel, K., Riggelsen, C., Merz, B., Kreibich, H., Scherbaum, F.; 2012. Published in *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*

In this paper a Bayesian network is learned for flood damage assessments. In contrast to classical approaches, which often relate the flood damage to the inundation depth only, a variety of potential influencing factors is considered here, aiming to identify the driving forces and to learn about their single and joint effects. The investigated data set comprises 29 variables describing the flooding situation, building characteristics, precaution, warning, emergency measures, socio-economic factors and the damaged caused to residential buildings. Missing observations are for simplicity randomly replaced. A better justified prediction procedure for missing values is presented in the following paper. The discretization of the continuous variables is casted in a Bayesian framework, searching for the most likely triple of network structure, parameters and discretization given the observed data. The learned network reveals interactions between flood damage and the considered predictors that are widely neglected in flood damage assessments. The performance of the learned BN in terms of predicting the building damage is compared to models currently used for flood damage assessments (namely the stage-damage function and the FLEMOps+r model) as well as to a Naive Bayes and Tree Augmented Naive Bayes classifier. As for the ground motion application the approximation of the target variable is refined here using a kernel density estimator. The best prediction performance is given by the Naive Bayes and Tree Augmented Naive Bayes, but especially for the Tree Augmented Naive Bayes it stays open to which extend this is an effect of over-fitting. The prediction performance of the BN is comparable to the FLEMOps+r model, which is to our knowledge the best model currently in use. Additionally the BN captures the related uncertainty and allows for inference into all directions.

PAPER 3

Challenges for Bayesian Network Learning in a Flood Damage Assessment Application

Vogel, K., Riggelsen, C., Scherbaum, F., Schröter, K., Kreibich, H., Merz, B.; 2013. Published in *Proceedings of the 11th International Conference on Structural Safety & Reliability*

The flood damage application presented in the previous paper is picked up here again concentrating now on methodological issues not satisfyingly solved or not discussed so far. The multivariate discretization procedure is enhanced to come up with a discretization that is independent of the scaling of the variables (i.e. the discretization of a variable does not change, if it is considered on a logarithmic scale). Missing observations are now estimated based on the observations of variables that have direct impact on the missing variable. This proceeding leads to a change in the learned network structure revealing dependency relations that match better with expert opinions, especially in context with the rarely observed variables. Moreover, an approximation of the target variable distribution with mixtures of truncated exponentials (MTE) is suggested here as an alternative to the kernel density estimator used in the previous papers. Using the MTE approach the BN performs equally well in the target variable prediction, while the number of applied parameters reduces considerably compared to the kernel density estimation. Especially in complex networks the parameter reduction is of importance to keep the computational effort for inference in reasonable limits.

PAPER 4

The Application of Bayesian Networks in Natural Hazard Analyses

Vogel, K., Riggelsen, C., Korup, O., Scherbaum, F.; Submitted to *Natural Hazards and Earth System Sciences*

The last paper illustrates the flexible applicability of the BN approach and demonstrates its properties and benefits on way of exemplifications. The BN learning procedures developed so far and presented in the previous papers are applied here in different natural hazard settings. The seismic data set generated for the first paper is reused to learn a BN applying the revised discretization procedure presented in the third paper. To increase the BN's precision the MTE approximation for continuous distributions is applied subsequently to the BN learning. The flood damage assessment presented in the second and third paper is considered once more and an example for inference, investigating the impact of precaution on flood damage, is given. Adding a third natural hazard to the investigated domains a landslide model is learned based on a data set that compiles a number of geological, climatic and topographic metrics throughout the Japanese islands. The model uncertainty related to BN learning is discussed in this context.

If not mentioned otherwise, the research documented in all presented papers was carried out by the author of this thesis. Co-authors assisted in an advisory role. In paper 2 and 3 the prediction performance of the currently used flood damage models (the stage-damage function and the FLEMOps+r model) was evaluated by Kai Schröter, Heidi Kreibich and Bruno Merz. The interpretation of the learned landslide model in paper 4 was supported by Oliver Korup.

In addition to the above mentioned papers, the author also participated in the following publication, which is not included in the thesis:

How useful are complex flood damage models?

Schröter, K., Kreibich, H., Merz, B., **Vogel, K.**, Riggelsen, C., Scherbaum, F.; Submitted to *Water Resources Research*

The paper analyzes the prediction performance of several flood damage models of different complexity. The contribution to this work was the implementation of BNs based on the data set collected after the Elbe 2002 flood event. Two BNs, comprising 11 and 28 variables, were learned totally data-driven, while the structures of another two BNs, comprising the same sets of variables, were defined based on expert knowledge and only the parameters were learned from data. The prediction performance of the BNs was evaluated considering their ability to predict building damages caused by the same flood event and in a spatial and temporal transfer.

CHAPTER

TWO

GRAPHICAL MODELS AS SURROGATES FOR COMPLEX GROUND MOTION MODELS

Kristin Vogel, Carsten Riggelsen, Nicolas Kühn, Frank Scherbaum

Institute of Earth and Environmental Science, University of Potsdam
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

Published in
Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing,
Zakopane (Poland), 2012

ABSTRACT: In Probabilistic Seismic Hazard Analysis, which has become the basis of decision making on the design of high risk facilities, one estimates the probability that ground motion caused by earthquakes exceeds a certain level at a certain site within a certain time interval. One of the most critical aspects in this context is the model for the conditional probability of ground motion given earthquake magnitude, source-site-distance and potentially additional parameters. These models are usually regression functions, including terms modeling interaction effects derived from expert knowledge. We show that the framework of Directed Graphical Models is an attractive alternative to the standard regression approach. We investigate Bayesian Networks, modeling the problem in a true multivariate way, and we look into Naive Bayes and Tree-Augmented Naive Bayes, where the target node coincides with the dependent variable in standard ground motion regression. Our approach gives rise to distribution-free learning when necessary, and we experiment with and introduce different discretization schemes to apply standard learning and inference algorithms to our problem at hand.

2.1 Introduction

In the context of Probabilistic Seismic Hazard Analysis (PSHA) strong ground motion at a particular site, caused by an earthquake, is modelled by physical relationships between various parameters, usually dictated by physical principles. This requires accurate knowledge of the source process, of the properties of the propagation medium as well as of the subsurface under the site. In regions of well recorded seismicity the most popular modeling approach is to fit a regression function to the observed data, where the functional form is determined by expert knowledge. In regions, where we lack a sufficient amount of data, it is popular to fit the regression function to a data set generated by a so-called *stochastic model* (Boore, 2003), which distorts the shape of a random time series according to physical principles to obtain a time series with properties that match ground-motion characteristics. The stochastic model does not have nice analytical properties nor does it come in a form amenable for easy analytical handling and evaluation. In order to determine the ground motion the stochastic model is simulated, posing a time-consuming and computationally expensive challenge. Instead of using a stochastic model directly, a surrogate model, which describes the stochastic model in a more abstract sense (e.g. regression), is often used in PSHA.

In this paper we show how *Directed Graphical Models* (DGM) may be seen as a viable alternative to the classical regression approach. Graphical models have proven to be a "all-round" pre/descriptive probabilistic framework for many problems. The transparent nature of the graphical models is attractive from a domain perspective allowing for a better understanding and gives direct insight into the relationships and workings of a system. A possible application of DGMs for PSHA is already described in (Kühn et al., 2009).

In the following sections we give a short introduction into the ground motion domain and into DGMs. How the DGMs are learned for discrete variables is explained in Section 2.4. Discretization methods and how we deal with a continuous target variable are given in Section 2.5. In Section 2.6 we apply DGMs to a dataset simulated by a stochastic model and we end with the conclusions.

2.2 Ground Motion Models

Formally speaking, in ground motion modeling we want to estimate the conditional probability of a ground motion parameter Y such as (horizontal) peak ground acceleration (PGA) or spectral acceleration (PSA) given earthquake and site related predictor variables, $\mathbf{X}$. In the regression approach the ground motion parameter is usually assumed to be log-normally distributed, $\ln Y = f(\mathbf{X}) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Which predictor variables are used is a matter of choice; in thus sequel we have at our disposal, $\mathbf{X} = \{M, R, SD, Q_0, \kappa_0, V_S30\}$. The moment magnitude of the earthquake (M) and distance between source and site (R) traditionally have special status in PSHA, however, we treat them no differently than the other variables: Stress released during the earthquake (SD), attenuation of seismic wave amplitudes in deep layers (Q_0) and near the surface (κ_0), Average shear-wave velocity in the upper 30 m (V_S30). [1]

Seismological expert knowledge determines the functional form of the regressions function; in our case a reasonable form for a regression function is the following, which is based on the description of the Fourier spectrum of seismic ground motion (Boore, 2003),

$$\begin{aligned} f(\mathbf{X}) &= a_0 + a_1 M + a_2 M \cdot \ln SD + (a_3 + a_4 M) \ln \sqrt{a_5^2 + R^2} \\ &\quad + a_6 \kappa R + a_7 V_S30 + a_8 \ln SD \end{aligned} \tag{2.1}$$

with $\kappa = \kappa_0 + t^*$, $\quad t^* = \frac{R}{Q_0 V_{sq}} \quad$ and $\quad V_{sq} = 3.5 \frac{km}{s}$, where a_i is fitted to data simulated from the stochastic model.

2.3 Directed Graphical Models

DGM's describe a joint probability distribution of a set of variables, $\mathbf{X}$, decomposing it into a product of (local) conditional probability distributions $P(\mathbf{X}|DAG, \boldsymbol{\theta}) = \prod_i P(X_i|\mathbf{X}_{Pa(i)}) = \prod_i \theta_{X_i|\mathbf{X}_{Pa(i)}}$ according to a directed acyclic graph (DAG), with vertices X_i and edges pointing from the parent set, $\mathbf{X}_{Pa(i)}$, to X_i, encoding the conditional independences. The local conditional probability distributions, $P(X_i|\mathbf{X}_{Pa(i)})$ may be defined according to our prior knowledge, e.g., as Gaussians where the mean and the variance could be associated with corresponding vertices in the DAG. However often we want to make no such explicit assumptions, that is, we want to be able to model a wide range of distributions, because no prior knowledge may

[1] In the next sections we will sometimes include Y in $\mathbf{X}$; it will be clear from the context when this is the case.

be available. By adhering to categorical distributions we may approximate (by "histograms") any continuous distribution asymptotically; this would be called *distribution-free learning*. E.g. if estimated from observations, the parameters could be the maximum likelihood estimates, $\hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}} = \frac{n(x_i,\mathbf{x}_{Pa(i)})}{n(\mathbf{x}_{Pa(i)})}$ using the statistics $n(\cdot)$, the counts of a particular configuration from data. More about discretization follows in Section 2.5.

In contrast to classical regression, DGMs treat all random quantities, including co-variates, as random variables. This is not only reasonable, since the measure of the covariates is often defective, but also allows to infer in "all directions" and calculate any conditional distributions of interest. Furthermore DGMs offer a different perspective on how variables (including co-variates) relate, since no assumptions about the functional form for physical relationships between the variables have to be given. On the other hand, expert knowledge can be included by the usage of informative priors, both on structure and parameters. For a more detailed description of DGMs see (Edwards, 2000).

2.4 Learning Approaches for Discrete Variables

In the sections to come we assume that we have at our disposal an i.i.d. sample, $\mathbf{d}$, with n records (for now assume discretized data); this will in our case be the simulated data from the stochastic model, from which we want to learn. We investigate DGMs admitting to different decompositions/factorizations of the joint distribution, that is, the restrictions that are imposed by the DAG: Bayesian Networks (BNs), Naive Bayes (NBs) and Tree Augmented Naive Bayes (TANs).

2.4.1 Bayesian Networks

In contrast to the regression approach, for BNs we do not need to make any assumptions about any (functional or (in)dependence) relationship of the involved variables *a priori*. When learned from data, we automatically get a concise surrogate model. By inspecting the learned BN structure we may get an intuition about the workings of the underlying data generating system (the stochastic model) from an (in)dependence perspective. The BN at the same time enables for computing any marginal/conditional of interest.

BN learning involves traversing the space of BNs looking for the one yielding the highest score. As scoring function we use the Bayesian MAP scoring, introduced in (Riggelsen, 2008), assuming a joint uniform prior $P(DAG, \mathbf{\Theta}) = P(\mathbf{\Theta}|DAG)P(DAG)$, with $P(\mathbf{\Theta}|DAG)$ a uniform product Dirichlet distribution (with restricted hyper-parameters $\boldsymbol{\alpha}$ guaranteeing DAG scoring equivalence), and $P(DAG)$ uniform over BN structures too. This yields the MAP scoring metric which needs to be maximized,

$$S(DAG|\mathbf{d}) = \prod_i \prod_{\mathbf{x}_{Pa(i)}} \prod_{x_i} \hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}}^{n(x_i,\mathbf{x}_{Pa(i)})+\alpha(x_i,\mathbf{x}_{Pa(i)})-1} \times \text{regularization term.}$$

To traverse the (simulated) space of essential graphs we use a hill-climber algorithm, applying the Repeated Covered Arc Reversal operator (Castelo and Kocka, 2003), where arc addition and removal are the basic operations. Without going into detail, we note that the thus obtained structure also dictates the parameter estimates of $\boldsymbol{\theta}$ as the maximum likelihood (see previous section), but now based on $\alpha(\cdot) + n(\cdot) - 1$.

2.4.2 Naive Bayes

In the BN approach, the structure is learned, and all variables are treated equally; there is no dedicated "output" node. In the context of our surrogate mode there is however a variable of interest, Y. The network structure in Naive Bayes (NBs) is simple and fixed: The target variable Y, often referred to as class variable, is the only parent of each attribute X_i. Even though the assumed independence between the attributes is most likely violated, NBs usually perform well (competitive with or better than BNs) in classification tasks (Friedman and Goldszmidt, 1996a). Obviously, in contrast to the BNs, with NBs we lack the ability to gain insight into the relationships between the variables via inspection.

The (local) conditional distributions may be the usual maximum likelihood estimates; however, we use the smoothed maximum likelihood estimator given in (Friedman and Goldszmidt, 1996a) instead, $\hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}} = \alpha \frac{n(x_i, \mathbf{x}_{Pa(i)})}{n(\mathbf{x}_{Pa(i)})} + (1-\alpha)\frac{n(x_i)}{n}$ with $\alpha = \frac{n(\mathbf{x}_{Pa(i)})}{n(\mathbf{x}_{Pa(i)})+5}$.

2.4.3 Tree Augmented Naive Bayes

Tree Augmented Naive Bayes (TANs) are an extension of the NBs. They allow each attribute to have one more parent in addition to the target variable. This relaxes the independence assumption for the attributes made in NB, but maintains the computational simplicity. In a TAN construction we start off with the NB structure. To determine on the presence or absence of connections between the attributes, we use a score based on entropy, $\mathsf{Ent}_{\mathbf{d}}(\mathbf{X}) = -\sum_{\mathbf{x}} \frac{n(\mathbf{x})}{n} \log_2 \frac{n(\mathbf{x})}{n}$. Entropy measures the amount of information, needed to specify $\mathbf{X}$ in the dataset $\mathbf{d}$ with n records. We determine for each pair, $(X_i, X_j)_{i \neq j}$, the *explaining away residual* (EAR) (Pernkopf and Bilmes, 2005),

$$\begin{aligned}\mathsf{EAR}(X_i, X_j|Y) &= \mathsf{Ent}_{\mathbf{d}}(X_i, Y) + \mathsf{Ent}_{\mathbf{d}}(X_j, Y) - \mathsf{Ent}_{\mathbf{d}}(X_i, X_j, Y) - \mathsf{Ent}_{\mathbf{d}}(Y) \\ &\quad -\mathsf{Ent}_{\mathbf{d}}(X_i) - \mathsf{Ent}_{\mathbf{d}}(X_j) + \mathsf{Ent}_{\mathbf{d}}(X_i, X_j),\end{aligned}$$

which is high for pairs which are mutually informative conditioned on Y and at the same time not mutually informative unconditionally. In an undirected *maximum spanning tree* the weights of the edges are associated with the EAR, all edges with negative weights are deleted and for the remaining tree(s) we choose a root node and set the direction of the edges pointing away from the root. These are the edges, which are added to the NB, ultimately yielding the TAN. The estimation of $\boldsymbol{\theta}$ is done as described for NBs.

2.5 Discretization

For a distribution-free learning, we need to discretize the continuous variables of our data set. A discretization splits the range of $\mathbf{X}$ into (multidimensional) intervals and merges all real values of one interval into one state of a discrete variable, $\mathbf{X}'$. The number of intervals and their boundaries have to be chosen carefully, since essential information about the distributions and dependencies of the variables may be lost otherwise.

2.5.1 Bayesian Network

For BN's, where we are mainly interested in learning the dependency structure, we discretize all variables simultaneously, using a *multivariate discretization*, which takes the interaction between all connected variables into account. We use a method developed in (Monti and Cooper, 1998), assuming that the observed data, were generated in two steps. In the first step an interval is selected by drawing from $P(\mathbf{X}'|DAG)$. Afterwards we draw $\mathbf{X}$ from a uniform distribution over the selected interval, $P(\mathbf{X}|\mathbf{X}') = \prod_i P(X_i|X_i')$. According to (Monti and Cooper, 1998) we now seek a discretization $\mathbf{d}'$ of $\mathbf{d}$, which maximizes for a given structure $P(\mathbf{d}'|DAG)P(\mathbf{d}|\mathbf{d}')$; here $P(\mathbf{d}'|DAG)$ is the so-called *marginal likelihood*.

The optimal discretization depends on the BN structure and has to be adjusted dynamically as the structure changes. We do this in an iterative way, similar to (Friedman and Goldszmidt, 1996b), first learning the discretization for an initial network, which in turn is used to learn a new BN with the MAP-scoring function. The discretization and the BN-learning steps are repeated until we reach a local maximum of the MAP-score. Starting with different initial networks can lead to different results. We use the structure of a TAN to start with, but ideally different initial GMs should be tested.

2.5.2 Naive Bayes and Tree Augmented Naive Bayes

The above mentioned approach is not ideal for the NB and TAN approach. Here our attention is on the estimation of the target variable, and we discretize only the attributes, while the continuous target is approximated with a kernel density estimator.

Our method is based on the approach developed in (Fayyad and Irani, 1993), which is widely used for the discretization of continuous attributes in classification tasks *with a discrete class variable* Y. The discretization of each attribute X_i depends on Y, but is independent of the other attributes. It splits the total dataset $\mathbf{d}$ into subsets $\cup_{k=1}^{K}\mathbf{d}_k = \mathbf{d}$, where $\mathbf{d}_k$ includes all records, for which X_i falls into the k-th interval. We aim to choose interval boundaries that lead to a small *Minimum Description Length* (MDL). The MDL can be expressed as $\sum_k \frac{n_k}{n}\mathsf{Ent}_{\mathbf{d}_k}(Y) + cost$, where n_k is the number of records in $\mathbf{d}_k$, $\mathsf{Ent}_{\mathbf{d}_k}(Y)$ is the *class entropy* based on the dataset $\mathbf{d}_k$ and *cost* is a regularization term restricting the number of intervals.

The above method is only valid for a discrete target variable, but our target is $Y = \ln PGA$, i.e., continuous. To apply the class entropy, we replace the continuous Y with a discrete approximation Y', whose states, $y'_1, ..., y'_{n_Y}$, correspond to the interval midpoints, we get by splitting the range of Y into n_Y equidistant intervals of width Δ_Y. We choose a large number n_Y (e.g. $n_Y = 512$) to allow a precise approximation of Y. In order to estimate $\mathsf{Ent}_{\mathbf{d}_k}(Y)$ reliably, we now use a Gaussian kernel density estimator, $\hat{P}_{Y,\mathbf{d}}$, with a bandwidth according to Silverman's "rule of thumb" (Silverman, 1986) and set, $\hat{P}(y'_i) = \Delta_Y \cdot \hat{P}_{Y,\mathbf{d}}(y = y'_i)$. The class entropy $\mathsf{Ent}_{\mathbf{d}_k}(Y) \approx -\sum_{i=1}^{n_Y} \hat{P}(y'_i) \log_2 \hat{P}(y'_i)$ can now be used for the discretization of the attributes as described above.

The very fine discretization allows a precise estimation of Y, while its prediction would be limited to a couple of states, if we use a coarse discretization as we would get by applying the method described in Section 2.5.1. Anyhow a coarse discretization is often more effective to capture the essentials of the joint distribution.

2.5.3 Adopted Parameter Estimation

Working with a continuous variable or rather a discrete one with lots of states Y', also requires a transformed parameter estimation for the graphical model. Using the statistics $n(\cdot)$, would lead to weak maximum likelihood estimates $\hat{\theta}_{X_i|\mathbf{X}_{Pa(i)}}$ whenever $Y' \in \mathbf{X}_{Pa(i)}$, since they are based on only a few observations. Hence, in case of $Y' \in \mathbf{X}_{Pa(i)}$, we rewrite,

$$P(X_i|\mathbf{X}_{Pa(i)}) = \frac{P(X_i, \mathbf{X}_{Pa(i)})}{\sum_{x_i} P(x_i, \mathbf{X}_{Pa(i)})} = \frac{P(Y'|X_i, \mathbf{X}_{Pa(i)-Y'})\, P(X_i, \mathbf{X}_{Pa(i)-Y'})}{\sum_{x_i} P(Y'|x_i, \mathbf{X}_{Pa(i)-Y'})\, P(x_i, \mathbf{X}_{Pa(i)-Y'})},$$

with $\mathbf{X}_{Pa(i)-Y'} = \mathbf{X}_{Pa(i)} \setminus Y'$. Here $P(Y'|\mathbf{z})$ is again estimated with a kernel density estimator, $\hat{P}_{Y,\mathbf{d_z}}$, based on $\mathbf{d_z}$, which are all records matching $\mathbf{z}$ in $\mathbf{d}$. Thus we get $\hat{P}(y'_i|\mathbf{z}) = \Delta_Y \cdot \hat{P}_{Y,\mathbf{d_z}}(y = y'_i)$. The Gaussian kernel of the density estimator is smoothed by multiplying it with a symmetric $n_Y \times n_Y$-weight-matrix. The matrix entries are chosen in order to keep the mean squared error of the target variable prediction small.

2.6 Application Experiment

We generate $\mathbf{d}$ from the stochastic model (Boore, 2003), with $n = 10.000$ records. The predictor variables are either uniform or exponentially distributed within a particular interval:

$M \sim \mathcal{U}_{[5,7.5]}$, $R \sim \mathsf{Exp}_{[1\text{km},\,200\text{km}]}$, $SD \sim \mathsf{Exp}_{[0\text{bar},300\text{bar}]}$,
$Q_0 \sim \mathsf{Exp}_{[0\text{s}^{-1},5000\text{s}^{-1}]}$, $\kappa_0 \sim \mathsf{Exp}_{[0\text{s},0.1\text{s}]}$, $V_S30 \sim \mathcal{U}_{[600\,\text{m/s},2800\,\text{m/s}]}$

and the PGA is generated by the stochastic model. In the development of the GMs, we use $\ln PGA$ instead of PGA, as we also use $Y = \ln PGA$ in the regression model. The learned GMs are illustrated in Fig. 2.1.

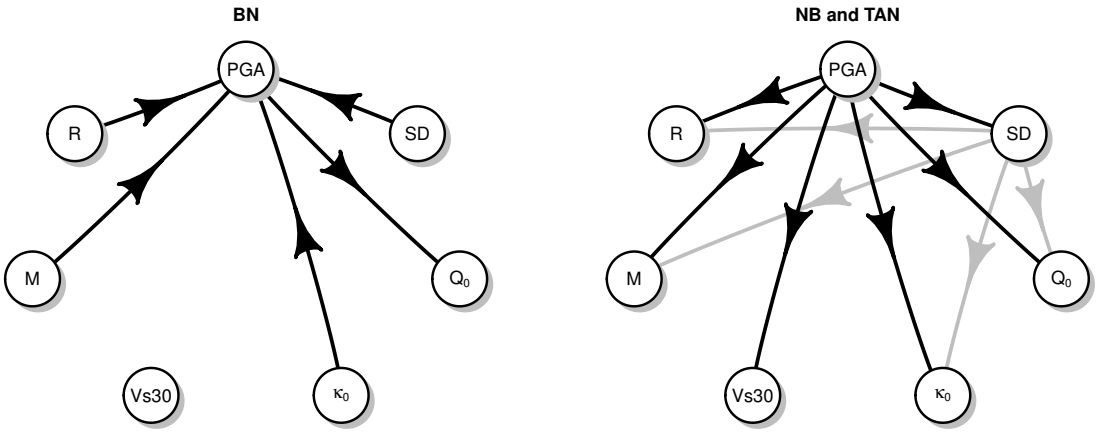


Figure 2.1: left: BN learned using MAP scoring metric; right: TAN learned with EAR, where the black edges show the NB and the gray ones the extension to TAN.

The BN structure gives insight into the data generating process. We already know, the learned direct independences between the attributes hold (no arcs), due to data construction. However, we observe conditionally induced dependences, which are reflected in the *v-connections*, e.g., $R \rightarrow \ln PGA \leftarrow SD$. The dependency between Q_0 and $\ln PGA$ is weak in our dataset and does not induce a conditional dependency to another co-variate. The influence of V_S30 may be ignored; apparently this variable is not essential in establishing $\ln PGA$.

An evaluation of the learned models may be done in terms of the performance of the $\ln PGA$ prediction. However, the BN has not been learned to perform well in terms of predicting $\ln PGA$ well; the BN captures the joint of all variables such that the performance is well "overall" for all variables. It is therefore somewhat unfair to judge the BN based just on the performance of one single variable. On the other hand, the TAN only captures the model structure in the sense that it yields the best performance of predicting $\ln PGA$. The performance test in terms of $\ln PGA$ prediction is done using a 5-fold cross validation and measured using mean squared errors (MSE); see Tab. 2.1.

For the prediction of $\ln PGA$ with the BN, we use the network structure and discretization of the covariates learned by applying the MAP scoring metric and the discretization method described in Section 2.5.1, but we ignore the discretization learned for $\ln PGA$. Instead, to allow a more precise prediction, we discretize $\ln PGA$ into $n_Y = 512$ equidistant intervals, as it is also done for NB and TAN (see Section 2.5.2) and recalculate the parameters of the affected variables as described in Section 2.5.3.

The prediction results of the GMs are quite well compared to the regression, using Eq. (2.1). There is only one case (4th dataset, BN) in which the GM performs worse than the regression model. In this case the algorithm failed to learn the connection to Q_0. The TAN and NB perform almost the same; the added flexibility of TANs, including the interaction effects of some predictor variables, does not seem to result in improvements in terms of MSE.

	1.	2.	3.	4.	5.	Avg.
BN	0.569	0.598	0.583	0.759	0.579	0.617
NB	0.488	0.489	0.566	0.592	0.473	0.522
TAN	0.509	0.525	0.566	0.597	0.494	0.538
Regression	0.666	0.679	0.688	0.681	0.650	0.673

Table 2.1: Mean squared errors of a 5-fold cross validation on a synthetic dataset using 8000 records for learning and a test set of 2000 records.

2.7 Conclusion

We presented an alternative to regression models for the construction of surrogates for ground motion models. Three GMs (BN, NB and TAN) were investigated along with schemes for discretization. On average they all perform better than the regression model in terms of predicting $\ln PGA$. Moreover, the entirely data-driven approach of learning the BN enables for a correct interpretation of the (in)dependences between the variables, as opposed to imposed algebraic interaction effects of the regression model. The advantages of GMs can help to tackle typical problems in PSHA. For instance are variables as κ_0 and Q_0 usually unknown and therefore not included in regression models. GMs offer the possibility to work with a distribution function instead of a precise value. This allows to deal with the uncertainty of these parameters in a more accurate way as in done in the general regression approach.

An obvious extension to NBs and TANs is to learn the entire Markov blanket of Y; this approach would yield an unrestricted *Bayesian Network classifier*. Hence, for vertex Y learn the parent set, children set and children's parent sets. Evaluation/model-selection would in that case rely on cross-validation of the predictive performance of Y (no direct scoring metric required).

CHAPTER

THREE

FLOOD DAMAGE AND INFLUENCING FACTORS: A BAYESIAN NETWORK PERSPECTIVE

Kristin Vogel [1], Carsten Riggelsen [1], Bruno Merz [2],
Heidi Kreibich [2], Frank Scherbaum [1]

[1] Institute of Earth and Environmental Science, University of Potsdam
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

[2] GFZ – German Research Centre for Geosciences,
Telegrafenberg, 14473 Potsdam, Germany

Published in
Proceedings of the 6th European Workshop on Probabilistic Graphical Models,
Granada (Spain), 2012

ABSTRACT: Classical approaches for flood risk assessment relate flood damage for a certain class of objects to the inundation depth, while other characteristics of the flooding situation and the flooded object are widely ignored. Observations on several discrete and continuous variables collected after the 2002 and 2005/2006 floods in the Elbe and Danube catchments in Germany offer a unique data mining opportunity in terms of learning a Bayesian Network. We take an entirely data-driven stance opting not to discretize continuous variables in advance; rather, we cast the problem in Bayesian framework, and consider the *maximum aposteriori* of the joint distribution of the triple, network structure, parameters and discretization, as the outcome of the analysis. Moreover, motivated by the work of Merz et al. (2010), who point out the need of an improved flood damage assessment, we re-define the discretization of the target variable, flood loss, once the network has been learned. Its domain is split into a large number of intervals and the associated parameters are estimated using a Gaussian kernel density estimator. Although the prediction of the relative flood loss is comparable to state-of-the-art methods, our approach benefits from capturing the joint distribution of all factors influencing flood loss.

3.1 Introduction

Graphical models have in recent years successfully been employed in earth sciences, giving rise to a wide range of applications, including Tsunami Early Warning, e.g. (Blaser et al., 2011), Probabilistic Seismic Hazard Analysis, e.g. (Kühn et al., 2011), and Automatic detection and classification of seismic signals, e.g. (Riggelsen et al., 2007). In this paper we embark on another problem: flood damage assessment of residential buildings. Typically, the damage to flooded objects is estimated by stage-damage functions which relate the relative or absolute damage for a certain class of objects to the water stage or inundation depth (Merz et al., 2010). Other characteristics of the flooding situation and of the flooded object are rarely taken into account, although it is clear that flood damage is influenced by a variety of factors such as inundation duration, contamination of flood water, or quality of external response in a flood situation. The single and joint effects of these parameters on the degree of damage are largely unknown and widely neglected in damage assessments. Moreover, the intrinsic uncertainties associated with these factors are largely ignored. *Bayesian Networks* (BN) pose an interesting formalism for capturing the interdependencies and the intrinsic uncertainty involved in flood risk assessment.

The paper is organized as follows: After giving a description of the data set in Section 3.2 and a brief introduction into learning BNs in Section 3.3, we show in Section 3.4, how we automatically discretize continuous variables, based on the observed data set, using a maximum a posteriori (MAP) score to search simultaneously for the best network structure, parameters and discretization. In Section 3.4.1 we point out, how we employ the learned BN to estimate the flood loss, using a very fine discretization of the target variable to allow a precise approximation of its continuous conditional distribution functions. Finally, the results are shown in Section 3.5 and we conclude in Section 3.6.

3.2 Variable Definitions and Dataset

We take a data mining perspective and aim for learning a BN from observational data. The observations are collected after the 2002 and 2005/2006 floods in the Elbe and Danube catchments in Germany. Results of computer-aided telephone interviews with 1135 flood affected households yield i.i.d. data, $\mathbf{d} = \cup_k \{\mathbf{x}^{(k)}\}$. Topics relate to various flood parameters (e.g. contamination, water depth), building and household characteristics, precautionary measures, and flood damage to buildings and contents. The raw data were supplemented by estimates of return periods, building values, loss ratio, i.e. the relation between the building damage and the building value, and indicators for flow velocity, contamination, flood warning, emergency measures, precautionary measures, flood experience and socioeconomic variables (Thieken et al., 2005; Elmer et al., 2010). Table 3.1 lists 28 candidate variables allocated to 5 domains and the predictand *rloss*, which is the direct damage to flooded residential buildings represented as relative value, i.e. fraction of the building value.

3.3 Bayesian Network Learning

Formally speaking a BN decomposes a joint probability distribution/density $P(\mathbf{X})$ into a product of (local) conditional probability distributions/densities $\mathsf{p}(\cdot|\cdot)$ as $P(\mathbf{X}) = \prod_i \mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$ according to a directed acyclic graph (DAG) with vertices X_i and directed edges from variables in the parent set $\mathbf{X}_{Pa(i)}$ to X_i. In case $\mathbf{X}$ is continuous and we do not know $\mathsf{p}(\cdot|\cdot)$ in advance, we may approximate the (local) conditional probability distributions by first discretizing the continuous variables and rely on contingency tables instead; the challenges that this involves is discussed in Section 3.4. For such a discrete BN we write $P(\mathbf{X}|DAG, \boldsymbol{\theta}) = \prod_i \theta_{X_i|\mathbf{X}_{Pa(i)}}$, where $\boldsymbol{\theta}$ (the *parameters*) are conditional probabilities derived from contingency tables. For the rest of this section we assume all variables $\mathbf{X}$ to be discrete.

BN model selection (learning the joint decomposition as well as the local conditional probabilities) is an exercise in traversing the space of BNs looking for the one which maximizes a given fitness score. As usual for model selection, regularization plays a role in this endeavor. We use the Bayesian BN MAP score (Riggelsen, 2008) shown to learn BN that are better than those derived via the marginal likelihood score (the BD-score). The BN is selected as the MAP of the joint posterior (here both DAG and parameter are being treated as true random variables)

$$P(DAG, \boldsymbol{\Theta}|\mathbf{d}) \propto P(\mathbf{d}|DAG, \boldsymbol{\Theta})P(\boldsymbol{\Theta}, DAG),$$

where the joint prior is a product, with $P(\boldsymbol{\Theta}|DAG)$ defined to be a product Dirichlet distribution and $P(DAG)$ defined to be uniform over DAGs (we may thus ignore this term when doing MAP estimation).

X_i	**Predictors**	**Scale and range**
	flood parameters	
wst	Water depth	C: 248 cm below ground to 670 cm above ground
d	Inundation duration	C: 1 to 1440 h
v	Flow velocity indicator	O: 0=still to 3=high velocity
con	Contamination indicator	O: 0=no contamination to 6=heavy contamination
rp	Return period	C: 1 to 848 yrs
	warning and emergency measures	
wt	Early warning lead time	C: 0 to 336 h
wq	Quality of warning	O: 1=receiver of warning knew exactly what to do to 6=receiver of warning had no idea what to do
ws	Indicator of flood warning source	N: 0=no warning to 4=official warning through authorities
wi	Indicator of flood warning information	O: 0=no helpful information to 11=many helpful information
wte	Lead time period elapsed without using it for emergency measures	C: 0 to 335 h
em	Emergency measures indicator	O: 1=no measures undertaken to 17=many measures undertaken
	precaution	
pre	Precautionary measures indicator	O: 0=no measures undertaken to 38=many, efficient measures undertaken
epre	Perception of efficiency of private precaution	O: 1=very efficient to 6=not efficient at all
fe	Flood experience indicator	O: 0=no experience to 9=recent flood experience
kh	Knowledge of flood hazard	N (yes / no)
	building characteristics	
bt	Building type	N (1=multifamily house, 2= semi-detached house, 3=one-family house)
nfb	Number of flats in building	C: 1 to 45 flats
fsb	Floor space of building	C: 45 to 18000 m^2
bq	Building quality	O: 1=very good to 6=very bad
bv	Building value	C: 92244 to 3718677 €
	socio-economic factors	
age	Age of the interviewed person	C: 16 to 95 yrs
hs	Household size, i.e. number of persons	C: 1 to 20 people
chi	Number of children (< 14 years) in household	C: 0 to 6
eld	Number of elderly persons (> 65 years) in household	C: 0 to 4
own	Ownership structure	N (1=tenant; 2=owner of flat; 3=owner of building)
inc	Monthly net income in classes	O: 11=below 500 €to 16=3000 €and more
socP	Socioeconomic status according to Plapp (2003)	O: 3=very low socioeconomic status to 13=very high socioeconomic status
socS	Socioeconomic status according to Schnell et al (1999)	O: 9=very low socioeconomic status to 60=very high socioeconomic status
	flood loss	
rloss	loss ratio of residential building	C: 0 = no damage to 1 = total damage

Table 3.1: Description of the candidate predictors (C: continuous, O: ordinal, N: nominal).

The simultaneous joint selection of DAG and parameters yields a DAG which equivalently can be found by maximizing the structure score

$$S(DAG|\mathbf{d}) = \prod_{i,\mathbf{x}_{Pa(i)},x_i} \hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}}^{n(x_i,\mathbf{x}_{Pa(i)})+\alpha(x_i,\mathbf{x}_{Pa(i)})} \underbrace{\prod_{i,\mathbf{x}_{Pa(i)}} \frac{\Gamma(\sum_{x_i}\alpha(x_i,\mathbf{x}_{Pa(i)}))}{\prod_{x_i}\Gamma(\alpha(x_i,\mathbf{x}_{Pa(i)}))}}_{regularization},$$

where the statistics $n(\cdot)$ are the counts of a particular configuration from the data/contingency table and $\alpha(\cdot)$ are the hyper-parameters of the Dirichlet (restricted as to guarantee uniform DAG scoring equivalence; see (Riggelsen, 2008)). The BN parameter estimates required for computing the above score are also the BN parameters, and are given in closed form by

$$\hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}} = \frac{n(x_i,\mathbf{x}_{Pa(i)}) + \alpha(x_i,\mathbf{x}_{Pa(i)})}{n(\mathbf{x}_{Pa(i)}) + \alpha(\mathbf{x}_{Pa(i)})}. \tag{3.1}$$

The BN is learned using a hill-climber approach in the space of DAGs based on the score given above where arc addition, removal and reversal are the basic operations. DAG equivalent classes are simulated using the Repeated Covered Arc Reversal operator (Castelo and Kocka, 2003).

3.4 Automatic Discretization

We want to adhere to an entirely data-driven approach for learning BNs and strive for making none or at least very weak assumption with regard to the functional form of the (local) conditional distributions, $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$. A sufficiently fine discretization of continuous variables, placing "counts" in contingency tables, enables us to approximate any other (local conditional) distribution, e.g., a Gaussian, but usually with a larger number of parameters.

To transform a continuous variable into a discrete one, the number of intervals and their boundaries have to be chosen carefully. A fine-grained discretization will result in a very sparsely connected BN due to regularization constraints, ultimately not reflecting the interactions we are interested in. On the other hand, too rough a discretization may not provide the "user" with the desired degree of resolution required for proper decision support. Various discretization approaches have been proposed in literature, (Friedman and Goldszmidt, 1996b; Monti and Cooper, 1998). Inspired by the latter we present a straightforward extension to the BN MAP learning score, allowing us to determine the "fitness" of a BN *and* discretization simultaneously, conditional on continuous data.

From now on, a superscript c denotes the continuous counterpart of a discretized variable/ configuration, e.g., $\mathbf{d}^c$ is the original continuous data and $\mathbf{d}$ a discretized version thereof. Let Λ define the discretization, that is, the set of interval boundary points for all variables; a configuration λ will thus "bin" the original data $\mathbf{d}^c$ yielding $\mathbf{d}$. Moreover, assume that we have $P(\mathbf{d}^c|\mathbf{d}, \Lambda)$; this is the generative model for the continuous data given some discretized version thereof as defined by Λ. Note that this model is unrelated to the BN; more on this distribution shortly. It follows that the likelihood for observing $\mathbf{d}^c$ for a given discretization, network structure and parameters (the BN) can be written as

$$P(\mathbf{d}^c|DAG, \mathbf{\Theta}, \Lambda) = P(\mathbf{d}|DAG, \mathbf{\Theta}, \Lambda)\ P(\mathbf{d}^c|\mathbf{d}, \Lambda).$$

Embedded in a Bayesian context, we are now seeking the MAP of the posterior

$$\begin{aligned} P(DAG, \mathbf{\Theta}, \Lambda|\mathbf{d}^c) &\propto P(\mathbf{d}^c|DAG, \mathbf{\Theta}, \Lambda)\ P(DAG, \mathbf{\Theta}, \Lambda) \\ &= \underbrace{P(\mathbf{d}|DAG, \mathbf{\Theta}, \Lambda)}_{1}\ \underbrace{P(\mathbf{d}^c|\mathbf{d}, \Lambda)}_{2}\underbrace{P(\mathbf{\Theta}|DAG, \Lambda)}_{3}\ \underbrace{P(\Lambda|DAG)}_{4}\ \underbrace{P(DAG)}_{5}. \end{aligned} \tag{3.2}$$

The product of the terms 1, 3 and 5 is equivalent to the (joint) BN posterior as introduced in Section 3.3 (terms 1 and 3 now of course depend on the discretization). Let term 4 be uniform on the space of all possible discretizations, allowing us to ignore this factor in the MAP estimation.

We define $P(X_i^c|X_i, \Lambda_i)$ according to the following considerations: the discrete X_i is associated with several interval boundaries, such that each state x_i has a lower λ_{x_i} and upper boundary λ^{x_i}. In between this interval X_i^c is distributed uniformly, outside it is zero; x_i thus "picks" the uniform interval in which X_i^c can lie. Effectively we arrive at term 2

$$P(\mathbf{d}^c|\mathbf{d}, \Lambda) = \prod_i \prod_{x_i} \left(\frac{1}{\lambda^{x_i} - \lambda_{x_i}} \right)^{n(x_i)},$$

which leads to a preference of small intervals, while term 1 and 3 counteract the formation of a high number of intervals.

Discretization and BN learning are nested iteratively: learn a BN for a given discretization, followed by learning a new discretization, and so on. For a given discretization maximizing Eq. (3.2) with respect to the BN-pair (DAG, $\mathbf{\Theta}$) is equivalent to maximizing the MAP BN score alone (which is the same as learning the DAG via $S(DAG|\mathbf{d}, \lambda)$ also implying the BN parameter estimates) because term 2 is independent of the BN. To find the interval boundaries for the discretization, the variables are discretized iteratively until Eq. (3.2) stops improving, or until a pre-defined number of iterations has been reached. At each step we select a variable and employ a binary search for the "best" discretization fixating the intervals for the other variables.

3.4.1 A Single Continuous Target

The "optimal" discretization will not necessarily result in the required resolution for a particular target variable of interest. In our case *rloss* is of primary interest, and the discretization described in the last section leads to a discretization into 5 intervals. To achieve a finer resolution we treat, once we learned the BN, the target variable X_i as discrete with a very large number of states, defined by splitting the range of X_i^c into n_{X_i} intervals, e.g., $n_{X_i} = 512$. For simplicity we use equidistant intervals of width Δ_{X_i}. The states of X_i are the midpoints of the corresponding intervals. Because of the large number of states, the estimator Eq. (3.1) for $\theta_{X_i|\mathbf{x}_{Pa(i)}}$, or $\theta_{X_j|\mathbf{x}_{Pa(j)}}$ when $X_i \in \mathbf{X}_{Pa(j)}$, is based on very few observations leading to weak estimates.

To avoid this problem, we use a Gaussian kernel density estimator to adopt the parameter estimation in the following manner. For $\theta_{X_i|\mathbf{x}_{Pa(i)}}$ we set

$$\hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}} = \Delta_{X_i} \tilde{P}_{X_i^c|\mathbf{x}_{Pa(i)}}(x_i),$$

where

$$\tilde{P}_{X_i^c|\mathbf{x}_{Pa(i)}}(x_i) = \frac{1}{n(\mathbf{x}_{Pa(i)})\ \sqrt{2\pi}\ h} \sum_{k|\mathbf{x}^{(k)}_{Pa(i)}=\mathbf{x}_{Pa(i)}} \exp\left(-\frac{(x_i^{c(k)} - x_i)^2}{2h^2}\right)$$

is the Gaussian kernel density estimator, with a bandwidth, h, according to Silverman's "rule of thumb" (Silverman, 1986), over all observations of X_i^c, for which $\mathbf{X}_{Pa(i)} = \mathbf{x}_{Pa(i)}$.

For $X_i \in \mathbf{X}_{Pa(j)}$ we use Bayes theorem to rewrite

$$\begin{aligned}
\mathsf{p}(X_j|\mathbf{X}_{Pa(j)}) &= \frac{\mathsf{p}(X_j, \mathbf{X}_{Pa(j)})}{\sum_{x_j} \mathsf{p}(x_j, \mathbf{X}_{Pa(j)})} \\
&= \frac{\mathsf{p}(X_i|X_j, \mathbf{X}_{Pa(j)-X_i})\ \mathsf{p}(X_j, \mathbf{X}_{Pa(j)-X_i})}{\sum_{x_j} \mathsf{p}(X_i|x_j, \mathbf{X}_{Pa(j)-X_i})\, \mathsf{p}(x_j, \mathbf{X}_{Pa(j)-X_i})} \\
&= \frac{\mathsf{p}(X_i|X_j, \mathbf{X}_{Pa(j)-X_i})\, \mathsf{p}(X_j|\mathbf{X}_{Pa(j)-X_i})}{\sum_{x_j} \mathsf{p}(X_i|x_j, \mathbf{X}_{Pa(j)-X_i})\ \mathsf{p}(x_j|\mathbf{X}_{Pa(j)-X_i})},
\end{aligned}$$

and we set

$$\hat{\theta}_{x_j|\mathbf{x}_{Pa(j)}} = \frac{\Delta_{X_i} \tilde{P}_{X_i^c|x_j,\mathbf{x}_{Pa(j)-X_i}}(x_i)\ \ \hat{\theta}_{x_j|\mathbf{x}_{Pa(j)-X_i}}}{\sum_{x_j} \Delta_{X_i} \tilde{P}_{X_i^c|x_j,\mathbf{x}_{Pa(j)-X_i}}(x_i)\ \ \hat{\theta}_{x_j|\mathbf{x}_{Pa(j)-X_i}}}.$$

Because of the large number of states for the target variable, inference can become time and space consuming. For relatively small/sparse networks this is not a big issue *per se* and in our particular case it has not posed any significant problem.

Mixtures of truncated Exponentials (MTE) are an alternative to approximate continuous distributions (Moral et al., 2001). Moreover using MTEs efficient inference is possible (Langseth et al., 2009b). Methods for their construction are given by e.g. Rumí et al. (2006) and Langseth et al. (2010). However, finding an optimal MTE-representation for a conditional/multivariate distribution from data is no trivial task. Moreover, learning both network structure and MTEs simultaneously from data is even more challenging. We leave this task for future work.

3.5 Results

We apply the BN-learning and discretization methods which are described in the last sections to the data set of Section 3.2. Continuous and ordinal variables are treated in the same manner and both discretized. Thus, the numbers of states is reduced for continuous as well as for discrete variables. Only the number of states for the nominal variables (ws, kh, bt, own) remains as given.

For $rloss$ the majority of the observations is gathered close to the lower domain boundary; taking the logarithm of $rloss$ results in more equal spread over the domain. To avoid an infinite domain range, the lower boundary, which corresponds to buildings with no damage, was set to $log(5.5 \cdot 10^{-6})$, where $5.5 \cdot 10^{-5}$ is the minimal observed loss ratio of damaged buildings.

There are missing values in the data, likely *missing (completely) at random*, M(C)AR. For convenience, we for now simply replace them by sampling from the observed values of the corresponding variable. Principled iterative methods like Expectation Maximization (EM) are intractable for our purpose, since learning both discretization and the BN means collecting sufficient statistics via inference (in the E-step of EM) disproportionally often. In the future the Markov Blanket Predictor (Riggelsen, 2006) will be employed, which is a fast one-pass approximation to EM, by restricting the attention to predictors of the Markov Blanket of a variable with a missing observation.

Figure 3.1 shows the BN learned for all variables listed in Tab. 3.1, starting from an initial "Tree Augmented Naive Bayes" network that was obtained using the method described in (Vogel et al., 2012a) with $rloss$ as class variable. Some of the variables (wte and fe) have zero in/out-degree as the data apparently did not support any interactions with other variables. Variables that belong to the same sub-domain, are in most cases linked via a short path. Even though all domains are included in the network, there is a clear distinction of the domains visible. Building characteristics and socio-economic factors have only indirect impact on the relative building loss, while flood parameters and precautions are closely related to the target variable. This gives us an idea about the importance of the variables for the calculation of the relative building loss.

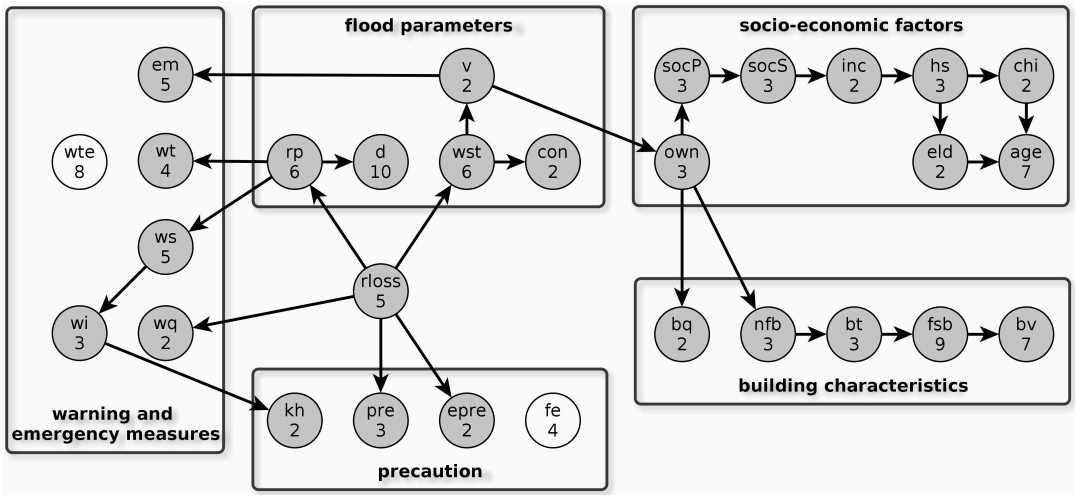


Figure 3.1: Bayesian network learned over the discretized data; Numbers in the nodes give the number of discrete states learned.

After the network was learned, *rloss* is selected as target variable and the number of states redefined viz. Section 3.4.1. Thus, we get an almost continuous approximation of the conditional probability function of *rloss*. Figure 3.2 illustrates the effect of the interval refinement. It shows the conditional distribution of *rloss* for the fine discretization in contrast to the coarse one for a flood event with water depth between 9 *cm* and 100 *cm*, a return period between 1 and 99 years and different precaution and warning levels (good: 1≤wq≤2, 13≤pre≤38, 1≤epre≤5; bad: 3≤wq≤6, 0≤pre≤2, epre=6).

We compare the performance of the BN in terms of the *rloss*-prediction to flood damage assessment approaches currently used in Germany, namely to the stage-damage-function approach and to FLEMOps+r (Elmer et al., 2010). For the stage-damage function approach, a root function is fitted to the damage data of certain object classes using least squares, i.e., the relative damage is a function of the water depth only. FLEMOps+r has been developed using the same data set and it has been shown to provide superior results compared to other approaches currently used in Germany. FLEMOps+r calculates the building loss ratio for private households using five classes of inundation depth, three intervals of flood frequency, three individual building types, two classes of building quality, three classes of contamination and three classes of private precaution. In essence, the data set is stratified into 27 subsamples and the average loss ratio is used as damage estimator (Elmer et al., 2010).

Additionally we compare the learned BN to the Naive Bayes (NB) and Tree Augmented Naive Bayes (TAN) learned from the same data set. These models are set up as restricted BNs with one single target variable in mind. We refer to (Vogel et al., 2012a) for a corresponding description of an automatic discretization and interval refinement according to Section 3.4.1. Of course the independence restrictions imposed by these models do not (unlikely) obey "reality" and can not be used to gain insight into the "workings" of the underlying system. However, in terms of predictions for a single target, they have shown to often outperform BNs.[1]

[1]In fact, discriminative models are even more likely to perform well, e.g., logistic regression.

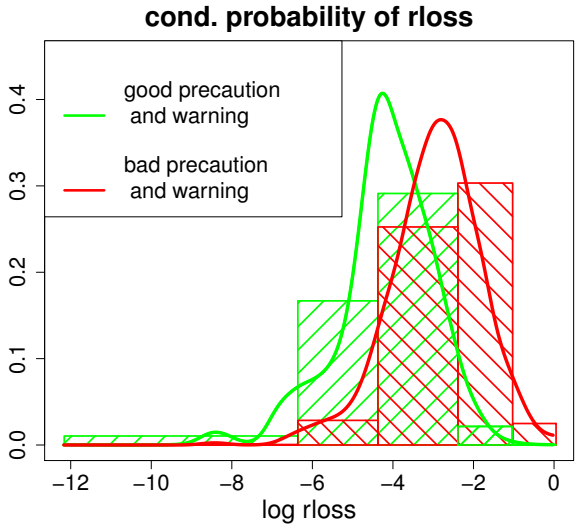


Figure 3.2: Conditional Probabilities of *rloss* for specific flood events using a coarse automatically learned discretization (shaded histograms) and interval refinement according to Section 3.4.1 (continuous lines).

To be able to compare to FLEMOps+r, we follow a evaluation policy commonly used in the field of hydrology: 100 bootstrap samples, each with 100 households, are drawn from the data set. We complete the results of the stage-damage-function and the FLEMOps+r model with the *rloss*-predictions we get from the BN, NB and TAN by using the expectation of the conditional *rloss*-distribution as predicted value. The predictions are quantified by the root mean squared error (RMSE) and the Pearson correlation coefficient.

It is important to stress that no separate test-sets are used: the bootstrap policy described uses parts of the training data for performance evaluation. This is not legitimate *per se*, and may influence the results considerably (optimistically). However, since the number of free parameters in the stage-damage function and the FLEMOps+r model are relatively small and the BN MAP criterion accounts for model complexity (it regularizes) these models will not over-fit and consequently performance testing on the training data will not yield overly optimistic results. Moreover, the BN MAP score is not a fitness measure of predictive performance of any target variable in particular, but rather, provides the predictive performance "overall" for all variables jointly. For the NB and TAN networks the number of free parameters is quite large and over-fitting might be a problem meaning that they on separate test-set may perform less well.

Figure 3.3 shows the performance measures for the 100 bootstrap samples in boxplots. It indicates that in terms of predicting *rloss* the BN performs well compared to the FLEMOps+r (the "best" method currently in use). The Naive Bayes and especially the Tree Augmented Naive Bayes show an improvement in the *rloss* prediction. However, it is important to note that the BN model in fact provides us with the joint distribution (the "correct" (in)dependence relationships) and it is therefore somewhat unfair to compare directly with approaches trying to improve upon the predictive performance of a single target variable only.

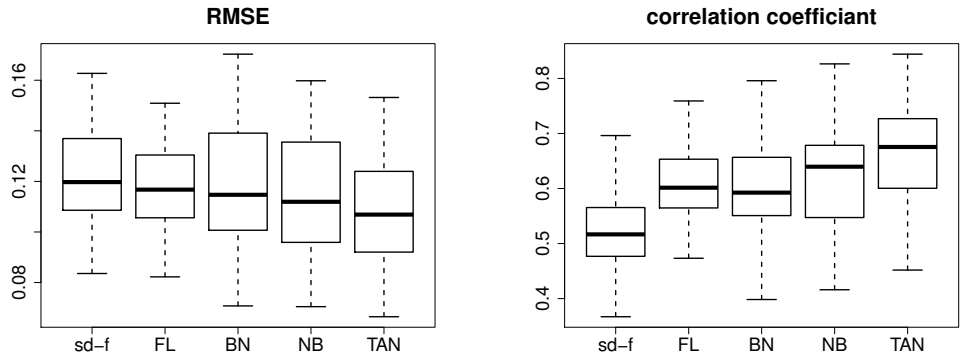


Figure 3.3: Comparison of flood damage estimation models (sd-f: stage damage function; FL: FLEMOps+r − model developed from same data set; BN: Bayesian Network; NB: Naive Bayes; TAN: Tree Augmented Naive Bayes).

3.6 Conclusion

A BN has been learned from real-life data describing flood related observations on 29 variables, the majority continuous and some discrete. In general continuous variables pose a challenge, and often discretization is performed as a pre-processing step prior to BN model selection. We have extended the BN MAP model selection metric to score not only BNs but simultaneously take the proper discretization into account, providing an entirely data-driven approach to learn from a Bayesian *maximum aposteriori* (MAP) perspective. From a data mining point of view, the BN indeed does reveal and confirm non-trivial interactions. The learned network captures the (in)dependencies revealing connectivity between flood loss and warning, emergency measures and socio-economic factors, which are widely neglected in flood risk assessment. Additionally, from a prediction point of view where the performance of a particular target is of interest, kernel estimation improves upon the BN "multivariate" view by increasing the degree of resolution (number of states) required for proper decision support (often derived/dependent on a single target variable). Compared to existing primarily deterministic flood damage estimation procedures, the BN shows a comparable performance with the added benefit of capturing and reasoning under uncertainty.

Acknowledgements

This work is supported by the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability, PROGRESS, a joint project of university and external organizations in the region of Potsdam-Berlin (Germany).

CHAPTER

FOUR

CHALLENGES FOR BAYESIAN NETWORK LEARNING IN A FLOOD DAMAGE ASSESSMENT APPLICATION

Kristin Vogel [1], Carsten Riggelsen [1], Frank Scherbaum [1]
Kai Schröter [2], Heidi Kreibich [2], Bruno Merz [2]

[1] Institute of Earth and Environmental Science, University of Potsdam
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

[2] GFZ – German Research Centre for Geosciences,
Telegrafenberg, 14473 Potsdam, Germany

Published in
Safety, Reliability, Risk and Life-Cycle Performance of Structures and Infrastructures,
G. Deodatis, B.R. Ellingwood and D.M. Frangopol, 2014

Challenges for Bayesian Network Learning
in a Flood Damage Assessment Application

ABSTRACT: Learning Bayesian networks from real world data poses challenges of different kinds. In this paper we address the principled handling of continuous variables in the light of incomplete observations using the example of flood damage assessment. The described proceeding can easily be transferred to other domains. Avoiding assumptions on the distributional family, we discretize the continuous variables thus allowing for distribution-free learning. The problem is cast in a Bayesian framework: instead of discretizing as a pre-processing step to the Bayesian network learning, we consider the *maximum aposteriori* of the joint distribution of the random variables: network structure, parameters and discretization. The incomplete observations raise an additional problem on top. Statistics used for the Bayesian network learning and discretization of the variables depend on the missing values. Iterative proceedings for the missing value prediction are infeasible, since they nest additional iterations into the already iterative algorithm for discretization and Bayesian network learning. Instead we use the non-iterative *Markov Blanket Predictor* (Riggelsen, 2006). This approach restricts attention to the Markov Blanket variables for which similar data records are used to derive several approximate predictive distributions. The flood damage assessment is based on a data set that involves 29 partly continuous variables and offers a unique opportunity for Bayesian network learning, taking not only the flooding parameters into account, but considering the warning situation, precautions, building characteristics and socio-economic factors as well. The resulting network shows, that the damage caused to a building by a flood event depends on a variety of factors. Especially flood and precaution parameters seem to have a crucial impact.

4.1 Introduction

In flood damage assessments we aim to estimate the damage to objects caused by a certain flood event. Similar to many other geo-scientific problems, we deal with complex processes where the driving forces are not well understood. The number of potential influencing factors is large and the single and joint effects of these parameters on the degree of damage are largely unknown and often neglected. The stage-damage function, which is still widely used in damage assessments, relates the damage for a certain class of objects to the water stage or inundation depth (Merz et al., 2010). Other characteristics of the flooding situation and of the flooded object are rarely taken into account, although flood damage is influenced by a variety of factors (Thieken et al., 2005). Moreover the intrinsic uncertainties associated with these factors and the modeling framework are largely ignored.

Bayesian Networks (BNs) are a probabilistic framework that allows for reasoning under- and propagation of uncertainty. A probabilistic approach proves to be especially important if the results are used downstream for decision support purposes (e.g. investment in risk mitigation), since a probability distribution carries significantly more information about the possible outcome than a single point estimate. Furthermore, BNs encode the (in-)dependencies of the involved variables and may thus explicitly be used to capture and illustrate the 'workings' of the underlying data generating process. This (in-)dependence structure can be obtained purely from observational data; no prior expert knowledge required, although if present, taking a Bayesian stance, it can be included into the learning procedure. A regularization term controls the complexity of the BN learned. Hence, over-fitting is avoided, even though the number of involved variables is large compared to the size of the data set.

In this paper we learn a BN based on a data set with many continuous variables that includes incomplete observations. A description of the used data set is given in Section 4.2, followed by a short introduction to BN learning in Section 4.3. To allow for a distribution-free learning we discretize the continuous values. The discretization is done parallel to the structure and parameter learning of the BN as described in Section 4.4. Another challenge we have to face is the handling of incomplete observations. In Section 4.5 we describe the Markov Blanket approach pertaining to missing values. Finally, in Section 4.6 we show and discuss the BN learned, followed by the conclusions in Section 4.7.

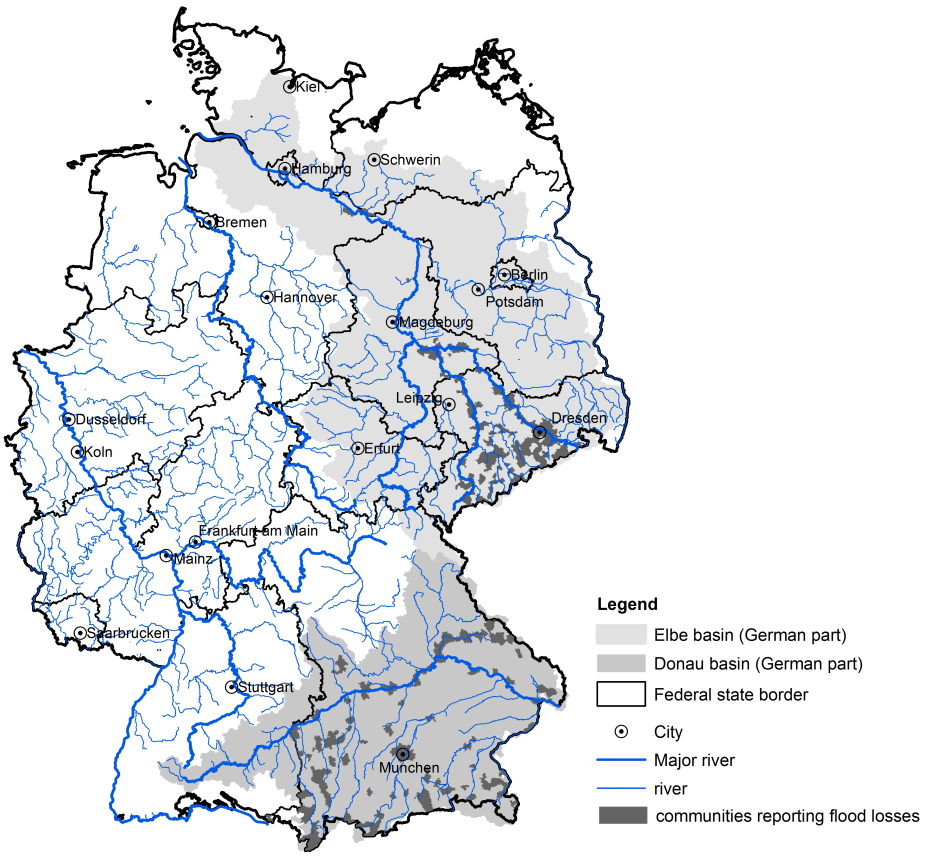


Figure 4.1: Investigated catchments and location of communities reporting losses from the 2002, 2005 and 2006 flood events in the Elbe and Danube catchments.

4.2 Variables and Data

The used data set $\mathbf{d} = \cup_{k=1}^{n}\{\mathbf{x}^{(k)}\}$ is the result of computer-aided telephone interviews with flood affected households in the Elbe and Danube catchments conducted after the 2002 and 2005/2006 flood events (Fig. 4.1) yielding $n = 1135$ records. The data contain a variety of discrete and continuous variables describing the flooding situation, building and household characteristics, precautionary measures and flood damage to buildings. The raw data were supplemented by estimates of return periods, building values and loss ratios and indicators for flow velocity, contamination, flood warning, emergency measures, precautionary measures, flood experience and socioeconomic variables (Thieken et al., 2005; Elmer et al., 2010). Table 4.1 lists the 28 considered variables allocated to 5 domains and the predictand *rloss*, which is the direct damage to flooded residential buildings represented as relative value, i.e. the relation between the building damage and the building value.

X_i	**Predictors**	**Scale and range**
	flood parameters	
wst	Water depth	C: 248 cm below ground to 670 cm above ground
d	Inundation duration	C: 1 to 1440 h
v	Flow velocity indicator	O: 0=still to 3=high velocity
con	Contamination indicator	O: 0=no contamination to 6=heavy contamination
rp	Return period	C: 1 to 848 yrs
	warning and emergency measures	
wt	Early warning lead time	C: 0 to 336 h
wq	Quality of warning	O: 1=receiver of warning knew exactly what to do to 6=receiver of warning had no idea what to do
ws	Indicator of flood warning source	N: 0=no warning to 4=official warning through authorities
wi	Indicator of flood warning information	O: 0=no helpful information to 11=many helpful information
wte	Lead time period elapsed without using it for emergency measures	C: 0 to 335 h
em	Emergency measures indicator	O: 1=no measures undertaken to 17=many measures undertaken
	precaution	
pre	Precautionary measures indicator	O: 0=no measures undertaken to 38=many, efficient measures undertaken
epre	Perception of efficiency of private precaution	O: 1=very efficient to 6=not efficient at all
fe	Flood experience indicator	O: 0=no experience to 9=recent flood experience
kh	Knowledge of flood hazard	N (yes / no)
	building characteristics	
bt	Building type	N (1=multifamily house, 2= semi-detached house, 3=one-family house)
nfb	Number of flats in building	C: 1 to 45 flats
fsb	Floor space of building	C: 45 to 18000 m^2
bq	Building quality	O: 1=very good to 6=very bad
bv	Building value	C: 92244 to 3718677 €
	socio-economic factors	
age	Age of the interviewed person	C: 16 to 95 yrs
hs	Household size, i.e. number of persons	C: 1 to 20 people
chi	Number of children (< 14 years) in household	C: 0 to 6
eld	Number of elderly persons (> 65 years) in household	C: 0 to 4
own	Ownership structure	N (1=tenant; 2=owner of flat; 3=owner of building)
inc	Monthly net income in classes	O: 11=below 500 €to 16=3000 €and more
socP	Socioeconomic status according to Plapp (2003)	O: 3=very low socioeconomic status to 13=very high socioeconomic status
socS	Socioeconomic status according to Schnell et al (1999)	O: 9=very low socioeconomic status to 60=very high socioeconomic status
	flood loss	
rloss	loss ratio of residential building	C: 0 = no damage to 1 = total damage

Table 4.1: Variables – C: continuous, O: ordinal, N: nominal.

4.3 Bayesian Network learning

Based on $\mathbf{d}$ a BN is learned that determines the influencing factors for flood damage to residential buildings. Reflecting the learned independencies between the variables a BN decomposes a joint probability distribution/density $P(\mathbf{X})$ into a product of (local) conditional probability distributions/densities $\mathsf{p}(\cdot|\cdot)$ as $P(\mathbf{X}) = \prod_i \mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$ according to a directed acyclic graph (DAG) with vertices X_i and directed edges from variables in the parent set $\mathbf{X}_{Pa(i)}$ to X_i. In case $\mathbf{X}$ is continuous and we do not know $\mathsf{p}(\cdot|\cdot)$ in advance, we may approximate the (local) conditional probability distributions by first discretizing the continuous variables and rely on contingency tables instead; the challenges that this involves is discussed in Section 4.4. For such a discrete BN we write $P(\mathbf{X}|DAG, \boldsymbol{\theta}) = \prod_i \theta_{X_i|\mathbf{X}_{Pa(i)}}$, where $\boldsymbol{\theta}$ (the *parameters*) are conditional probabilities derived from contingency tables. For the rest of this section we assume all variables $\mathbf{X}$ to be discrete.

BN model selection (learning the joint decomposition as well as the local conditional probabilities) is an exercise in traversing the space of BNs looking for the one which maximizes a given fitness score. As usual for model selection, regularization plays a role in this endeavor. We use the Bayesian BN MAP score (Riggelsen, 2008) shown to learn BNs that are better than those derived via the marginal likelihood score (the BD-score). The BN is selected as the MAP of the joint posterior (here both DAG and parameter are being treated as true random variables),

$$P(DAG, \boldsymbol{\Theta}|\mathbf{d}) \propto P(\mathbf{d}|DAG, \boldsymbol{\Theta})P(\boldsymbol{\Theta}, DAG),$$

where the joint prior is a product, with $P(\boldsymbol{\Theta}|DAG)$ defined to be a product Dirichlet distribution and $P(DAG)$ defined to be uniform over DAGs (we may thus ignore this term when doing MAP estimation). The simultaneous joint selection of DAG and parameters yields a DAG which equivalently can be found by maximizing the structure score,

$$S(DAG|\mathbf{d}) = \prod_{i,\mathbf{x}_{Pa(i)},x_i} \hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}}^{n(x_i,\mathbf{x}_{Pa(i)})+\alpha(x_i,\mathbf{x}_{Pa(i)})} \underbrace{\prod_{i,\mathbf{x}_{Pa(i)}} \frac{\Gamma(\sum_{x_i}\alpha(x_i,\mathbf{x}_{Pa(i)}))}{\prod_{x_i}\Gamma(\alpha(x_i,\mathbf{x}_{Pa(i)}))}}_{regularization}$$

where the statistics $n(\cdot)$ are the counts of a particular configuration from the data/contingency table and $\alpha(\cdot)$ are the hyper-parameters of the Dirichlet (restricted as to guarantee uniform DAG scoring equivalence; see (Riggelsen, 2008)). The BN parameter estimates required for computing the above score are also the BN parameters, and are given in closed form by $\hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}} = \frac{n(x_i,\mathbf{x}_{Pa(i)})+\alpha(x_i,\mathbf{x}_{Pa(i)})}{n(\mathbf{x}_{Pa(i)})+\alpha(\mathbf{x}_{Pa(i)})}$.

The space of BNs is explored using a hill-climber approach in the state-space of DAGs based on the score given above where arc addition, removal and reversal are the basic operations. DAG equivalent classes (essential graphs) are simulated using the Repeated Covered Arc Reversal operator (Castelo and Kocka, 2003).

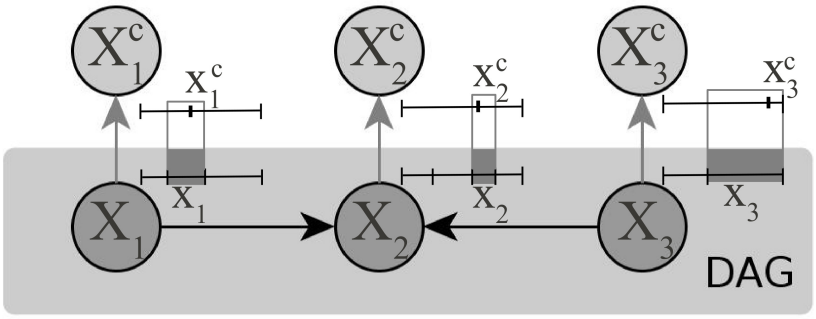


Figure 4.2: Example for the handling of continuous variables: the dependency structure of the variables is captured by their discrete representations. The continuous variable, X_i^c, depends only on its discrete counterpart, X_i and its realization x_i^c is a value in the interval that corresponds to the discrete realization, x_i.

4.4 Automatic Discretization

As mentioned above our data set contains continuous variables that we have to take care of. Since we strive to make none or at least very weak assumption with regard to the functional form of the (local) conditional distributions, $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$, we discretize the data, thus allowing for a distribution-free learning. To transform a continuous variable into a discrete one, the number of intervals and their boundaries have to be chosen carefully. A fine-grained discretization will result in a very sparsely connected BN due to regularization constraints, ultimately not reflecting the interactions we are interested in. On the other hand, a too rough discretization may not provide the "user" with the desired degree of resolution required for proper decision support.

Similar to our proceeding in (Vogel et al., 2012b), we use a Bayesian approach to extend the BN MAP score, that was introduced in Section 4.3, for discretization purposes. The extended score allows us to determine the "fitness" of a BN *and* discretization simultaneously while regularizing the total number of BN parameters. It thus avoids too fine discretizations or dense network structures.

Discretization and BN learning are nested iteratively in our method: learning a BN for a given discretization, followed by learning a new discretization, and so on. The discretization itself is an iterative process again: in each iteration we select a variable and employ a binary search for the "best" discretization of this variable, while the intervals of the other variables are kept fix. This is repeated until the score, that is described in the following, stops improving or until a pre-defined number of iterations has been reached.

To define the extended BN MAP score, we use following notations. From now on, a superscript c denotes the continuous counterpart of a discretized variable/configuration, e.g., $\mathbf{d}^c$ is the original continuous data and $\mathbf{d}$ a discretized version thereof. Let Λ define the discretization, that is, the set of interval boundary points for all variables; a configuration λ will thus "bin" the original data $\mathbf{d}^c$ yielding $\mathbf{d}$.

Based on an idea by Monti and Cooper (1998) we assume that all the interaction that takes place between the variables can be captured by their discrete representations. Thus it is sufficient to define a BN for the discrete variables and to add the continuous variables as nodes that depend only on their discrete counterpart (see Fig. 4.2). This means the generation (from a statistical perspective) of a continuous datum x_i^c works as follows: in the first step the discrete datum x_i is generated by a multinomial distribution modeled by the discrete BN. In the second step the continuous datum x_i^c is now generated conditional on the discrete x_i — from a generative perspective we "draw" a value from the given interval "picked" by x_i; this generation process is independent of the BN structure and BN parameters. The probability for observing $\mathbf{d}^c$ for a given discretization, network structure and parameters (the BN) can thus be decomposed into,

$$P(\mathbf{d}^c|DAG, \mathbf{\Theta}, \Lambda) = P(\mathbf{d}|DAG, \mathbf{\Theta}, \Lambda)\ P(\mathbf{d}^c|\mathbf{d}, \Lambda).$$

Embedded in a Bayesian context, we are now seeking the MAP of the posterior,

$$\begin{aligned} P(DAG, \mathbf{\Theta}, \Lambda|\mathbf{d}^c) &\propto P(\mathbf{d}^c|DAG, \mathbf{\Theta}, \Lambda)\ P(DAG, \mathbf{\Theta}, \Lambda) \\ &= \underbrace{P(\mathbf{d}|DAG, \mathbf{\Theta}, \Lambda)}_{1}\ \underbrace{P(\mathbf{d}^c|\mathbf{d}, \Lambda)}_{2}\underbrace{P(\mathbf{\Theta}|DAG, \Lambda)}_{3}\ \underbrace{P(\Lambda|DAG)}_{4}\ \underbrace{P(DAG)}_{5}. \end{aligned}$$

The product of the terms 1, 3 and 5 is equivalent to the (joint) BN posterior as introduced in Section 4.3 (terms 1 and 3 now of course depend on the discretization). Term 4 can be ignored under the assumption to be uniform on the space of all possible discretizations. The remaining term 2 was in (Vogel et al., 2012b) defined as $P(\mathbf{d}^c|\mathbf{d}, \Lambda) = \prod_i \prod_{x_i} \left(\frac{1}{\lambda^{x_i} - \lambda_{x_i}}\right)^{n(x_i)}$, where λ_{x_i} and λ^{x_i} are the lower and upper interval boundary of the interval that contains x_i. This definition has some weaknesses, we had to counteract in previous calculations:

1. Infinite domain ranges of X_i had to be delimited artificially to avoid infinite interval lengths and a convergence of term 2 towards zero.
2. Term 2 converges towards infinity, if interval boundaries are artificially chosen within an infinite small area around one observation. This effect was avoided by allowing interval boundaries only at the midpoints between two observations.
3. The learned discretization depends on the metric of the variable. The discretization of e.g. $\log X_i$ does not correspond to the one of X_i. Therefore the metric of the variable should be chosen carefully.

A new definition of $P(X_i^c|X_i, \Lambda_i)$ is motivated by the following considerations: let's define a metric for X_i^c such, that each observation of X_i^c corresponds to one unit of the metric. The discrete variable, X_i, is associated with several intervals, such that each state x_i corresponds to one interval of length $n(x_i)$ in the defined metric, since $n(x_i)$ this is the number of realizations

within the interval. The distribution function of X_i^c given x_i is assumed to be uniform within the interval and zero outside. The definition for term 2 is now,

$$P(\mathbf{d}^c|\mathbf{d},\Lambda) = \prod_i \prod_{x_i} \left(\frac{1}{n(x_i)}\right)^{n(x_i)}.$$

Analogue to the definition in (Vogel et al., 2012b) term 2 increases for an increasing number of intervals, while term 1 and 3 decrease at the same time and balance the resolution of the discretization.

Using the new definition for term 2, it is of no interest where exactly the interval boundaries lie. It is sufficient to know which two observations envelope the boundary. Thus we also get a natural limit for the number of possible discretizations.

In the procedures described so far all variables are treated with the same priority, but actually we have an increased interest in the estimation of the target variable *rloss* and may need a more precise resolution of *rloss*. More on this is addressed in Section 4.6.

4.5 Missing values

An other problem we have to face, is the handling of missing/incomplete values in our data. The scoring function we use in Section 4.3 and 4.4 for BN learning and discretization is a function of the sufficient statistics $n(\cdot)$, that are not defined for incomplete observations. To bypass the problem we predict the missing values using a predictive distribution, $P(\mathbf{U}|\mathbf{o},\boldsymbol{\theta},DAG)$, defined as a distribution of all missing values $\mathbf{u}$ in $\mathbf{d}$, given the observed values, $\mathbf{o}$ in $\mathbf{d}$, and a DAG. With this we can effectively compute the expected sufficient statistics $\mathbb{E}\left[n(x_i,\mathbf{x}_{pa(i)})\right] = \sum_{\mathbf{u}} n(x_i,\mathbf{x}_{pa(i)})\, P(\mathbf{u}|\mathbf{o},\boldsymbol{\theta},DAG)$ which is equivalent to (statistically correct) "summing out" the missing values in $\mathbf{d}$ for missing data mechanisms which are *ignorable* according to the *missing (completely) at random* (M(C)AR) criteria as formally defined in (Little and Rubin, 1987).

In the context of graphical models exists a variety of iterative approaches to determine the predictive distribution. For example the Structural Expectation-Maximization (Friedman, 1997, 1998) includes the model selection search within the Expectation-Maximization algorithm. A Bayesian method for learning is given in (Tanner and Wong, 1987), where a stochastic simulation-based approach for learning parameters, called Data Augmentation, is introduced. However, in our case we have to run several iterations of BN learning and discretization, each iteration requiring the estimation of the missing values. An iterative approach for missing value prediction will thus quickly become infeasible. We adhere to a more efficient albeit approximate method by applying the Markov Blanket Predictor (Riggelsen, 2006). Basically the algorithm consists of just two steps: in the *G-step* (generation step) we generate univariate predictive distributions; in the *E-step* (expectation step) we use the predictive distributions to compute the expected completions of the records and can consequently estimate the expected overall sufficient statistics.

Approximate predictive distributions

The basic idea behind the method is to predict a missing variable X_i in record l by concentrating only on its Markov Blanket (MB), $\mathbf{X}_{MB(i)}$, which are the variables that directly influence X_i, i.e. parents, children and parents of children. For illustration see Fig. 4.3, left. Assuming the MB is fully observed it blocks the influence from all other variables and the missing value depends only on the observed values in the MB,

$$P(X_i^{(l)}|\mathbf{o}^{(l)}, \boldsymbol{\theta}, DAG) = P(X_i^{(l)}|\mathbf{x}_{MB(i)}^{(l)}, \boldsymbol{\theta}, DAG). \tag{4.1}$$

When the MB is not completely observed, i.e. some variables therein are missing as well, it does not shield off the variable with the missing value. However, for approximation purposes we ignore this (which is not a big issue *per se*) and go ahead defining the approximate predictive distribution for each missing variable as above and may decompose Eq. (4.1) according to the DAG,

$$P(X_i^{(l)}|\mathbf{x}_{MB(i)}^{(l)}, \boldsymbol{\theta}, DAG) \propto \theta_{X_i|\mathbf{x}_{Pa(i)}} \prod_{j \in Ch(i)} \theta_{X_j|\mathbf{x}_{Pa(j)}},$$

where $Ch(i)$ are the variable indices of the children of X_i. The prediction of X_i requires simple inference, assuming correct parametrization of the BN, hence correct estimates $\hat{\boldsymbol{\theta}}$, *prior* to predicting the missing values; however, these we don't know and unfortunately can not reliably estimate for MAR missing data mechanisms without resorting to iterative estimation procedures.

Approximate parameter estimation

To avoid iterative approaches to find $\boldsymbol{\theta}$ we re-define the approximate predictive distribution as follows. We derive a network DAG' from DAG (see Fig. 4.3) by allocating edges, pointing to X_i, to each variable of the MB of X_i in DAG: $\mathbf{X}_{Pa(i)}^{DAG'} = \mathbf{X}_{MB(i)}^{DAG}$ and $\mathbf{X}_{Ch(i)}^{DAG'} = \emptyset$. The resulting DAG' preserves all dependencies given in DAG and can thus, approximating Eq. (4.1), be used to predict missing values (for any univariate variable of any record),

$$P(X_i|\mathbf{X}_{Pa(i)}^{DAG'}, \boldsymbol{\theta}^{DAG'}, DAG') \stackrel{def}{=} \theta_{X_i|\mathbf{X}_{Pa(i)}}^{DAG'}. \tag{4.2}$$

For the estimation of the parameters $\theta_{X_i|\mathbf{X}_{Pa(i)}}^{DAG'}$ in Eq. (4.2) we use a *similar cases* approach. The sufficient statistic $n(x_i, \mathbf{x}_{Pa(i)})$ is estimated with $s(x_i, \mathbf{x}_{Pa(i)})$, which is a weighted count of all records where $X_i = x_i$ and the observed part of the parents set $\mathbf{X}_{Pa(i)}$ matches with $\mathbf{x}_{Pa(i)}$. The counts are weighted with their degree of match. This means, for each record where $(X_i, \mathbf{X}_{Pa(i)})$ is fully observed we add 1 to $s(x_i, \mathbf{x}_{Pa(i)})$. For an incomplete observed parents set we count the possible completions of the record and add 1/number-of-possible-completions to $s(x_i, \mathbf{x}_{Pa(i)})$. This means we share the count for the record uniformly among all possible completions of the parents set. We now use $s(\cdot)$ to estimate the parameters from Eq. (4.2),

$$\hat{\theta}_{x_i|\mathbf{x}_{Pa(i)}}^{DAG'} = \frac{s(x_i, \mathbf{x}_{Pa(i)})}{\sum_{x_i} s(x_i, \mathbf{x}_{Pa(i)})}$$

that now fully define the approximate predictive distribution for X_i.

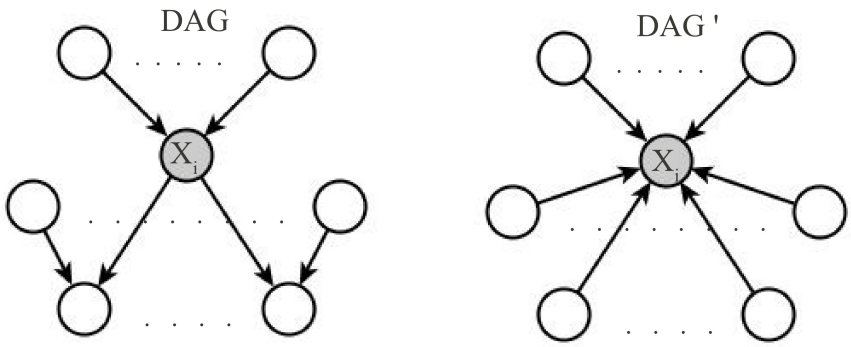


Figure 4.3: left: Markov Blanket of X_i; right: new model DAG' derived from DAG (all edges connecting variables in the Markov Blanket of X_i in DAG are directed towards X_i in DAG').

Selecting predictive variables

For small data samples and dense BN's we will end up with only few observations for each combination of parent states. The resulting estimates, $\hat{\boldsymbol{\theta}}^{DAG'}$, will thus be based on few observations and lead to poor results. To avoid this effect, the number of variables used for the predictive distribution can be reduced. Instead of defining the whole MB of X_i in DAG as parents of X_i in DAG', we select only the b best predictors from the MB. The predictors follow certain constraints and their predictive power can be determined in closed form. For a more detailed description we refer to (Riggelsen, 2006).

4.6 Application

Based on the data set of Section 4.2 we learn a BN by applying the BN-learning and discretization algorithm described in Section 4.3 and 4.4. The missing values are handled as suggested in Section 4.5 by using only the $b = 2$ best predictive variables for the estimation of the missing values. Continuous and ordinal variables are treated in the same manner and both discretized. Thus, the numbers of states is reduced for continuous as well as for discrete variables. Only the number of states for the nominal variables (ws, kh, bt, own) remains as given.

As initial network to start with, we choose a *Tree Augmented Naive Bayes* (TAN) network with $rloss$ as target variable. Thus we give some initial priority to the damage describing variable and the resulting BN has a higher probability of being dense in the surrounding of $rloss$. The initial TAN is obtained by applying the methods described in (Vogel et al., 2012a).

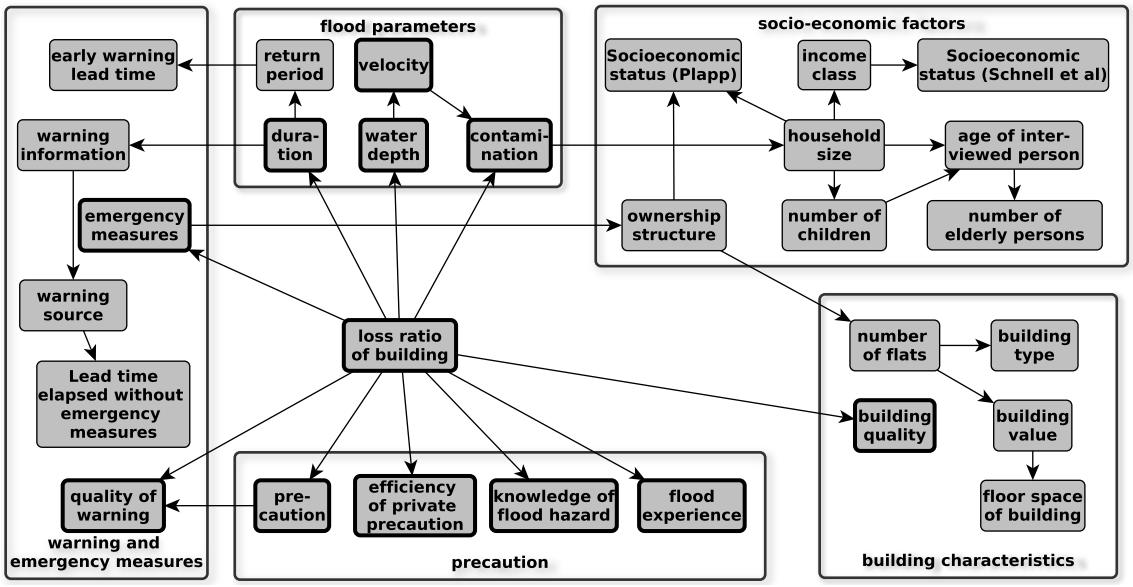


Figure 4.4: Bayesian network describing the damage to residential buildings caused by the 2002 and 2005/2006 flood events in the Elbe and Danube catchments; nodes marked with a thicker border belong to the MB of the relative building loss

The resulting BN, shown in Fig. 4.4, has a number of direct links connecting the damage describing variable directly with all subdomains except for socio-economic factors. Especially the domains 'precaution' and 'flood parameters' are densely connected with *rloss*, which seems to be reasonable and coincides with our intuition. Even though around half the records of the variable 'flood experience' are missing, the connection to the building damage, which experts assume to be valid, is found. This is in contrast to the network learned in (Vogel et al., 2012b), where missing values were replaced randomly. The found dependency is a hint for getting improved results by applying the missing value algorithm described in Section 4.5.

The developed model considers a large number of variables. Complete observations for all included variables are rare, but this is no drawback of the model. BNs are capable to use incomplete observations. By summing missing variables out, they exploit the information content that is present. The inclusion of many variables thus actually proves as an advantage, since it provides additional knowledge e.g. about missing variables that directly influence the target variable. Looking at the MB of *rloss* we find several variables (flood duration, water depth, flow velocity, contamination, building quality, flood experience, knowledge of hazard, precaution, efficiency of private precaution, warning quality and emergency measures) with direct influence on the building damage. This supports the assumption made in (Merz et al., 2010), that demands improved flood damage assessments, taking several variables into account. Especially precaution and flood parameters seem to have a high influence on the damage of the building.

Prediction performance

As mentioned we have an increased interest in the estimation of *rloss*, but the learned discretization will not necessarily result in a resolution that matches our requirements for prediction. One way of achieving a better resolution is to refine, once the BN is learned, the discretization of the target variable by splitting its range into many (e.g. 512) intervals and using a kernel density estimator re-estimate the BN parameters pertaining to *rloss*; for details refer to (Vogel et al., 2012b). However, the larger number of states is detrimental to BN inference which becomes both more time- and space consuming. An alternative is to use an approximation based on *mixtures of truncated exponentials* (MTE) (Moral et al., 2001; Langseth et al., 2010), allowing for efficient inference because MTEs are closed under probability propagation. Using MTEs we can define a conditional distribution $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$ by partitioning $\Omega_{(X_i,\mathbf{X}_{Pa(i)})}$ into hypercubes $D_1,\ldots,D_L$ and defining the density such that within each hypercube, D_l, it follows the form

$$\mathsf{p}_{\downarrow D_l}(X_i|\mathbf{X}_{Pa(i)}) = a_0 + \sum_{j=1}^{J} a_j \, \mathrm{e}^{\,b_j X_i + c_j^T \mathbf{X}_{Pa(i)}}. \tag{4.3}$$

To fit a (conditional) MTE density to observations, we have to address the following issues: the determination of the hypercubes into which $\Omega_{(X_i,\mathbf{X}_{Pa(i)})}$ is partitioned, the determination of the number of exponential terms in each hypercube and the estimation of the parameters. We use the maximum likelihood approach described in (Langseth et al., 2010) to solve these tasks and to define an approximation for the conditional distribution function of $X_i = rloss$. It should be recognized that *rloss* has no parents and Eq. (4.3) thus simplifies to a function that depends only on X_i. The found density has the form

$$\mathsf{p}(X_i) = 0.015 + 21.46\mathrm{e}^{-71.18\,X_i} + 4.71\mathrm{e}^{-6.88\,X_i}.$$

We now have to adapt the definitions for the conditional distributions of the children of *rloss*. Here we assume that the only effect of *rloss* on a child X_j is through the definition of the hypercubes. This means within an hypercube is $\mathsf{p}(X_j|\mathbf{X}_{Pa(j)})$ constant in *rloss*. We use the method described in (Langseth et al., 2009a) to find for each child variable the best partition of Ω_{rloss} to define the hypercubes.

The original discretization, that was learned parallel to the BN, splits the domain of *rloss* into 3 intervals. Figure 4.5 shows two examples of the effect of the resolution refinement for *rloss* using the kernel density estimator and the MTE approach.

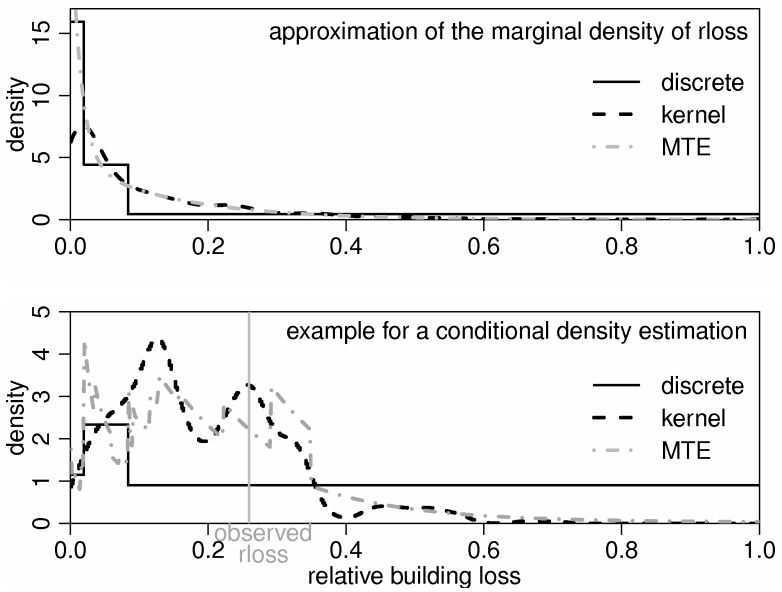


Figure 4.5: Top: Approximation of the marginal density of *rloss* with a discrete distribution for the learned discretization, with a kernel density estimator and the MTE approach. Bottom: Estimated conditional densities for a particular observation in the data.

	sdf	FLEMO	BN_{kernel}	BN_{MTE}	BN_{discr}
1	0.0111	0.0108	0.0101	0.0104	0.0456
2	0.0133	0.0114	0.0118	0.0116	0.0415
3	0.0161	0.0145	0.0146	0.015	0.0429
4	0.0200	0.0194	0.0165	0.0169	0.0496
5	0.0166	0.0150	0.0170	0.0163	0.044
Avg	0.0154	0.0142	0.014	0.014	0.0447

Table 4.2: Mean squared errors of a 5-fold cross-validation, where *rloss* is predicted with the three models stage-damage function, FLEMOps+r and Bayesian Networks.

We compare the performance of the learned BN in terms of predicting *rloss* – using the kernel density estimation and the MTE approach – to flood damage assessment approaches currently used in Germany, namely to the stage-damage function (sdf) approach and to FLEMOps+r (Elmer et al., 2010). For the sdf approach, a root function is fitted to the damage data of certain object classes using least squares, i.e., the relative damage is a function of the water depth only. FLEMOps+r has been developed using the same data set and it has been shown to provide superior results compared to other approaches currently used in Germany. FLEMOps+r calculates the building loss ratio for private households using five classes of inundation depth, three intervals of flood frequency, three individual building types, two classes of building quality, three classes of contamination and three classes of private precaution. In essence, the data set is stratified into 27 subsamples and the average loss ratio is used as damage estimator (Elmer et al., 2010).

To evaluate the prediction performance of the models we conduct a 5-fold cross-validation. The total data set is divided into 5 subsamples of same sizes. In each run we define one of these subsample to be the test set; the other four are used to learn the BN, sdf and FLEMOps+r model. The learned models are used to estimate the relative building loss for each record of the test set, given all observed values (besides the building loss itself). The sdf and the FLEMOps+r give no estimates for records where the used predictor variables are (partly) missing, while the BN has no difficulties to handle incomplete observations. The uncertainty concerning the unobserved variables is captured in the probability function as well as the uncertainty about the modeling framework and the intrinsic randomness (aleatory uncertainty). To include all records in our analysis, we define for sdf and FLEMOps+r: the estimated *rloss* for incomplete observations is the mean of all observed *rloss* values.

While sdf and FLEMOps+r give point estimates, the BN provides a distribution function for *rloss* (received with the kernel density estimation or with the MTE approach). It is a main benefit of BNs to keep track of these uncertainties. The identification and communication of uncertainty is especially important if the results are used for further computations and decision support. Anyhow, for the model comparison we condense the distribution function to its mean value, which we define to be the estimate for *rloss*. Thus we reduce provided information and disadvantage the BN. Nevertheless the BN provides good prediction results. Table 4.2 gives for all models the mean squared error of the estimated *rloss*. For the BN it shows additionally to the estimation results received with the refined resolutions the mean squared error we get using the originally learned course discretization. The prediction performance is clearly improved by the resolution refinement. The two refinements provide similar results – while the MTE approach requires less parameters – and their prediction performance is comparable to the one of the FLEMOps+r model. This is a promising result, since the BN is not developed to approximate the conditional distribution of the target variable, but aims to approximate the joint distribution of *all* variables and thus puts as much effort into the prediction of the building loss as e.g. into the prediction of the number of children. The sdf and FLEMOps+r aim directly at an optimal prediction of the building loss and do not have the flexibility of the BN, which allows inference in all directions and can thus be used for predictions of all (combinations of) variables of interest (e.g. which effect has precaution on the building loss).

4.7 Conclusion

A BN for flood damage assessment was learned based on a data set describing flood related observations on a mixture of discrete and continuous variables. Here, as in many real-life applications, continuous variables and incomplete observations pose a challenge for BN-learning. Instead of discretizing the continuous variables as a pre-processing step prior to the BN model selection, we extend the BN MAP model selection metric to score not only BNs, but take the discretization into account simultaneously. Thus we aim to find the best combination of BN structure, parameters and discretization; the combination that maximizes the joint posterior. For the handling of the missing values we use an MB approach. This alternative to iterative methods avoids an additional nesting of iterations and thus keeps the computational effort feasible. Additionally we showed how the learned BN can be modified for a more precise estimation of a chosen target variable.

Most of the methods applied here are derived from existing ideas developed and used for stand-alone problems. Interfacing these individual methods is challenging, because they are highly interacting. Also the computational efficiency plays an important role in this regard. The here described proceeding was coded in "R" in a generic fashion. It is a completely data-driven approach working without significantly human interaction. It can without much ado be applied to continuous incomplete data sets from other domains, where reasoning under- and propagation of uncertainty matters.

Acknowledgements

This work is supported by PROGRESS, a research cluster of university and external organizations in the region of Potsdam-Berlin (Germany). Carsten Riggelsen is funded by DFG project RI 2037/2-1.

CHAPTER

FIVE

THE APPLICATION OF BAYESIAN NETWORKS IN NATURAL HAZARD ANALYSES

Kristin Vogel, Carsten Riggelsen, Oliver Korup, Frank Scherbaum

Institute of Earth and Environmental Science, University of Potsdam
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

Submitted to
Natural Hazards and Earth System Sciences,
on August 14, 2013

ABSTRACT: In natural hazards we face several uncertainties due to our lack of knowledge and/or the intrinsic randomness of the underlying natural processes. Nevertheless, deterministic analysis approaches are still widely used in natural hazard assessments, with the pitfall of underestimating the hazard with potentially disastrous consequences. In this paper we show that the Bayesian network approach offers a flexible framework for capturing and expressing a broad range of different uncertainties as those encountered in natural hazard assessments. Although well studied in theory, the application of Bayesian networks on real-world data is often not straightforward and requires specific tailoring and adaption of existing algorithms. We demonstrate by way of three case studies (a ground motion model for a seismic hazard analysis, a flood damage assessment, and a landslide susceptibility study) the applicability of Bayesian networks across different domains showcasing various properties and benefits of the Bayesian network framework. We offer suggestions as how to tackle practical problems arising along the way, mainly concentrating on the handling of continuous variables, missing observations, and the interaction of both. We stress that our networks are completely data-driven, although prior domain knowledge can be included if desired.

5.1 Introduction

Natural hazards such as earthquakes, tsunamis, floods, landslides or volcanic eruptions all share – despite of differing causes, triggers, and effects – many of the same model- and decision theoretic questions. The underlying natural processes are often complex, while the number of potential influencing factors is large. The single and joint effects of the driving forces are not necessarily completely understood, potentially introducing a large degree of uncertainty, which impacts any quantitative analysis. Additionally, the observation on the basis of which inference is made is often sparse, inaccurate and incomplete, adding yet another layer of uncertainty on top. Thus various sources of uncertainty accumulate and interact in a non-trivial fashion. Indeed, this is one of the reasons why probabilistic approaches are often avoided. Such probabilistic approaches often seem intimidating given the large number of parameters required for such a proceeding, where the analytical tools needed for a rigorous handling of uncertainties are unknown or at least not readily available.

Deterministic approaches rarely provide information on the uncertainty related to parameter estimates beyond the use of statistical measures of dispersion such as standard deviations or standard errors about empirical means. However, uncertainty is a carrier of information to the same extent as a point estimate, and ignoring it or dismissing it as simply an error would be wrong. Ignoring uncertainties in quantitative hazard appraisals may have disastrous effects, since it often leads to over- or underestimates of certain event magnitudes. Yet, deterministic approaches are still the state of the art in many applications. For example, tsunami early warning systems evaluate pre-calculated synthetic databases and pick out the scenario that appears closest to a given situation in order to estimate its hazard (Blaser et al., 2011).

In this paper we consider Bayesian networks (BNs) that combine probability theory with graph theory, thus allowing for an intuitive, consistent, and rigorous way of quantifying uncertainties. The (in-)dependencies between the involved variables relevant to a particular hazard domain are translated into a graph structure that enables improved understandings and direct insights into the relationships and workings of a natural hazard system. Moreover, as BNs capture the joint distribution of all variables of a given domain, they may be used for expressing any conditional probability distribution of interest, thereby helping to answer quantitative questions on specific scenarios or process response chains.

In recent years BNs have already successfully been employed in a wide range of earth sciences, including automatic classifications of seismic signals e.g. (Riggelsen et al., 2007), tsunami early warning, e.g. (Blaser et al., 2009, 2011), Probabilistic Seismic Hazard Analyses e.g. (Kühn et al., 2011), and earthquake induced landslide susceptibility e.g. (Song et al., 2012). Here we additionally highlight three applications of BNs in natural hazard assessments that illustrate the flexibility of BNs. Alongside we discuss problems that arise when BNs are learned from real world data. The handling of continuous variables and incomplete observations is the chief problem here. Most approaches we use are derived from existing ideas and standard approaches, but their combination is challenging, since the individual methods are highly interacting.

In the next section we give an introduction into BNs, which is followed by several applications in natural hazards research. In Section 5.3 we develop a seismic ground motion model based on a synthetic data set, which serves to showcase some typical BN properties. We demonstrate an option to deal with continuous variables, making no prior assumptions about their distributional family. In Section 5.4 data collected after the 2002 and 2005/2006 flood events in the German Elbe and Danube catchments are used to learn a BN for flood damage assessments. This incomplete data set requires a treatment of missing observations, which is especially challenging in combination with continuous variables. A last application in Section 5.5 deals with a regional landslide susceptibility model for Japan, where we investigate how the same set of potential predictors of slope stability may produce nearly equally well performing, though structurally different, BNs, thus revealing important variable interactions that are often overlooked in landslide studies. This application further illustrates the model uncertainty related to BN learning. Conclusions and scope for future work perspectives are presented in Section 5.6.

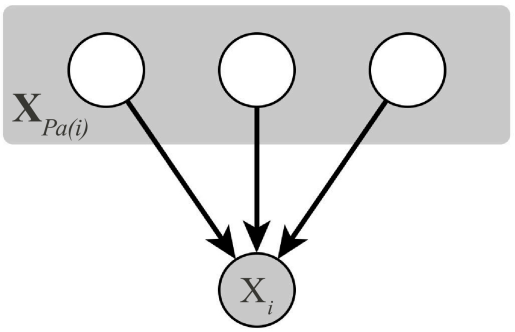
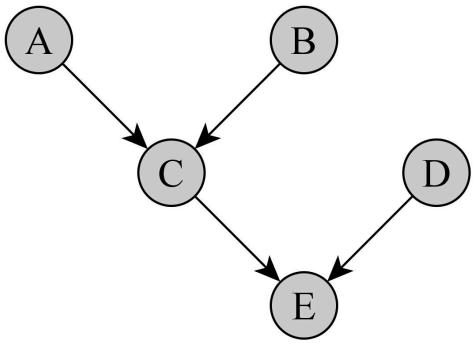


Figure 5.1: Illustration of a parent set $\mathbf{X}_{Pa(i)}$ of X_i in a BN.



$P(A, B, C, D, E) = \mathsf{p}(E|C,D)\mathsf{p}(D)\mathsf{p}(C|A,B)\mathsf{p}(B)\mathsf{p}(A)$

Figure 5.2: Example for the decomposition of the joint distribution according to a DAG.

5.2 Bayesian Networks

We operate in the world of probability theory and adhere to the corresponding probability axioms. We treat all random quantities pertaining to a particular hazard domain, including the covariates, as true random variables, each associated with a (conditional distribution), $\mathsf{p}(\cdot|\cdot)$. The interactions between those variables, be it measurables, observations or otherwise, therefore really boils down to the question of how the distributions of those random variables interact (at all times in accordance with the axioms of probability theory). A BN is a convenient framework for capturing such interactions, both from an intuitive, but also from a computational point of view. We refer to (Koller and Friedman, 2009) for a more detailed description of BNs in general and will restrict ourselves here to a few key aspects of the formalism.

5.2.1 Properties and Benefits

As previously mentioned, BNs combine probability theory with graph theory, representing all involved variables, $\mathbf{X} = \{X_1, \ldots, X_k\}$, as nodes in a directed acyclic graph, DAG; these variables will coincide with the specific variables of a hazard domain of interest. The arcs of the DAG point from the variables in the parent set, $\mathbf{X}_{Pa(i)}$, to X_i (see Fig. 5.1), stating that X_i directly depends on (is influenced by, is affected by, ...) $\mathbf{X}_{Pa(i)}$. The probability distribution of X_i is defined conditional on its parents set, $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$. Formally, a BN for $\mathbf{X}$ is a pair $(DAG, \boldsymbol{\theta})$, the structure, DAG and the parameters, $\boldsymbol{\theta} = \bigcup\{\theta_{x_i|\mathbf{x}_{Pa(i)}} = \mathsf{p}(x_i|\mathbf{x}_{Pa(i)})\}$,[1] which describes a joint distribution factorizing as,

$$P(\mathbf{X}|DAG, \boldsymbol{\theta}) = \prod_{i=1}^{k} \mathsf{p}(X_i|\mathbf{X}_{Pa(i)}). \tag{5.1}$$

[1] Here we assume discrete variables for which the set of parameters corresponds to the conditional (point) probabilities for each combination of states. For continuous variables the design of the parameters depends on the particular densities $\mathsf{p}(\cdot|\cdot)$.

Figure 5.2 is an example of a simple BN. Although not immediately clear from Eq. (5.1), a BN is characterized by many attractive properties which we may profit from in a natural hazard setting; here we mention a few:

Prop. 1 **- Graphical representation:** The interactions of the variables of the entire "system" are encoded through the DAG. The BN structure thus provides information about the underlying processes and the way various variables communicate and share "information" as it spreads around the network. Formally speaking the DAG encodes independence assumptions between variables.

Prop. 2 **- Use prior knowledge:** The intuitive interpretation of a BN makes it possible to define the BN beforehand based on prior knowledge; alternatively it may be learned from data, or even a combination of the two (cast as Bayesian statistical problem; see later for details): pose a prior BN and update it based on observations.

Prop. 3 **- Identify relevant variables:** By learning the BN from data we may identify the variables that are (according to the data) relevant. If several "islands" or isolated single nodes (not connected) appear, then it indicates that these variables potentially may be irrelevant.

Prop. 4 **- Capture uncertainty:** Uncertainty can easily be propagated throughout the network, between any nodes/variables; we effectively compute/estimate probability distributions rather than single point estimates.

Prop. 5 **- Allow for inference:** Instead of explicitly modeling the conditional distribution of a predefined target variable, the BN captures the joint distribution of all variables. Via inference, we can express any/all conditional distribution(s) of interest, in any directions (including forensic and inverse reasoning), e.g. we may ask for an observed damage, what is a likely intensity of the damage causing event. A detailed example for reasoning is given in Section 5.4.3.

Inference in BNs is closed under restriction, marginalization and combination enabling for fast (close to immediate) and exact inference.

Prop. 6 **- Use incomplete observations:** During predictive inference (i.e., computing a conditional distribution) incomplete observations of data do not pose a problem to BNs. By the virtue of the probability axioms, it merely impacts the overall uncertainty involved.

In the course of the paper we will refer to these 1–6 properties to make clear what is meant.

For "real-life" modeling problems, including those encountered in natural hazard analysis, adhering strictly to the BN formalism is often a challenging task, and consequently the properties listed above at times seem rather theoretical/academic. In the sections to come we will by way of examples shed some light on how typical natural hazard problems can be formulated around BNs. We take a data-driven stance, and the aim is thus to learn BNs from observations collected; the next section will introduce some key aspects related to inducing BNs.

5.2.2 Learning Bayesian Networks

Data based BN learning can be seen as an exercise in finding a pair $(DAG, \boldsymbol{\theta})$ which according to the BN decomposition in Eq. (5.1) could have been "responsible for generating the data". For this we traverse the space of pairs looking for the BN which maximizes a given fitness score; this should however be done with careful consideration to the issues always arising in the context of model selection, i.e., over-fitting, generalization, etc. Several suggestions for BN fitness scoring appear in literature derived from different theoretical principles and ideas. We opt for a Bayesian approach to learn BNs[2]. This has several advantages and allows us to elegantly combine prior knowledge and data, i.e. we have encoded our prior belief on the BN space $(DAG, \boldsymbol{\Theta})$ as $P(DAG, \boldsymbol{\Theta})$, we observe data $\mathbf{d}$ and obtain the revised distribution $P(DAG, \boldsymbol{\Theta}|\mathbf{d})$ based upon which we select the most probable BN. More specifically, we may rewrite the joint posterior as,

$$\underbrace{P(DAG, \boldsymbol{\Theta}|\mathbf{d})}_{\text{posterior}} \propto \underbrace{P(\mathbf{d}|DAG, \boldsymbol{\Theta})}_{\text{likelihood}} \underbrace{P(\boldsymbol{\Theta}, DAG)}_{\text{prior}}. \tag{5.2}$$

The prior distribution allows us to compensate sparse data, artifacts, bias, etc. but also allows us to assign domain specific prior preferences to certain BNs before seeing data (Prop. 2). In the following applications we work with a non-informative prior, which at the same time acts as a penalty term that regularizes the DAG complexity and avoids over-fitting. The likelihood term decomposes according to Eq. (5.1). Detailed descriptions for prior and likelihood term are given in Appendix 5.A.1.

Using the BN MAP (Bayesian Network Maximum A Posteriori) score we can maximize the right side of Eq. (5.2) (the product of the prior and the likelihood as given above) as to obtain the "best" BN given data; see (Riggelsen, 2008) for details with regard to search strategy, hyper-parameter selection, etc.

In the following section we learn a ground motion model, which is used in probabilistic seismic hazard analysis, as a BN; the used data set is synthetically generated. This section serves as an illustration of the BN formalism "in action" and will also present some theoretical and practical problems along with potential solutions in the context of BN learning. In the sections thereafter we will use real data.

5.3 Seismic Hazard Analysis: Ground Motion Models

When it comes to decision making on the design of high risk facilities, the hazard arising from earthquakes is an important aspect. In probabilistic seismic hazard analysis (PSHA) we calculate the probability of exceeding a certain ground motion at a certain site within a certain time interval. One of the most critical elements in PSHA, often carrying the largest amount of uncertainty, is the ground motion model. It describes the conditional probability of a ground

[2]Note that BNs are not necessarily to be interpreted from a Bayesian statistical perspective.

X_i	**Description**	**Distribution**$_{[Range]}$
	predictors	
M	moment magnitude of the earthquake	$\mathcal{U}_{[5,7.5]}$
R	source-to-site distance	$\mathsf{Exp}_{[1\text{km},\,200\text{km}]}$
SD	stress released during the earthquake	$\mathsf{Exp}_{[0\text{bar},500\text{bar}]}$
Q_0	attenuation of seismic wave amplitudes in deep layers	$\mathsf{Exp}_{[0\text{s}^{-1},5000\text{s}^{-1}]}$
κ_0	attenuation of seismic wave amplitudes near the surface	$\mathsf{Exp}_{[0\text{s},0.1\text{s}]}$
V_S30	average shear-wave velocity in the upper $30m$	$\mathcal{U}_{[600\,\text{m/s},2800\,\text{m/s}]}$
	ground motion parameter	
PGA	horizontal peak ground acceleration	according to the *stochastic model*

Table 5.1: Variables used for the ground motion model.

motion parameter, Y, such as *(horizontal) peak ground acceleration*, given earthquake and site related predictor variables, $\mathbf{X}_{-Y}$. Ground motion models are usually regression functions, where the functional form is derived from expert knowledge and the ground motion parameter is assumed to be log-normally distributed: $\ln Y = f(\mathbf{X}_{-Y}) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The definition of the functional form of $f(\cdot)$ is guided by physical model assumptions about the single and joint effects of the different parameters, but also contains some ad hoc elements. Using the Bayesian network approach there is no prior knowledge required *per se*, but if present it can be accounted for by encoding it in the prior term of Eq. (5.2). Is no reliable prior knowledge available, we work with a non-informative prior and the learned graph structure provides insight into the dependence structure of the variables and helps for a better understanding of the underlying mechanism. Modeling the joint distribution of all variables, $\mathbf{X} = \{\mathbf{X}_{-Y}, Y\}$, the BN implicitly provides the conditional distribution $P(Y|\mathbf{X}_{-Y},\ DAG, \boldsymbol{\Theta})$, which gives the probability of the ground motion parameter for specific event situations needed for the PSHA.

5.3.1 The Data

The event situation is described by the predictor variables $\mathbf{X}_{-Y} = \{M, R, SD, Q_0, \kappa_0, V_S30\}$, which are explained in Tab. 5.1. We generate a synthetic data set consisting of 10.000 records. The ground motion parameter, Y, is the horizontal peak ground acceleration (PGA). It is generated by a so called *stochastic model* (Boore, 2003), which distorts the shape of a random time series according to physical principles and obtains time series with properties that match the ground-motion characteristics. The predictor variables are either uniform ($\mathcal{U}$) or exponentially (Exp) distributed within a particular interval (see Tab. 5.1).

Since the stochastic model does not have nice analytical properties and its usage is non-trival and time consuming, surrogate models, which describe the stochastic model in a more abstract sense (e.g. regressions), are usually used in PSHA. We show that BNs may be seen as a viable alternative to the classical regression approach. However, before doing so, we need to touch upon some practical issues arising when learning BNs from continuous data.

For continuous variables we need to define the distributional family for the conditionals $\mathsf{p}(\cdot|\cdot)$ and thus make assumptions about the functional form of the distribution. To avoid such assumptions and "let the data speak", we discretize the continuous variables, thus allowing for a completely data-driven and distribution-free learning. In the following subsection we describe an automatic discretization, which is part of the BN learning procedure and takes the dependencies between the single variables into account. However, the automatic discretization does not necessarily result in a resolution that matches the requirements for prediction purposes or decision support. To increase the potential accuracy of predictions we approximate, once the network structure is learned, the continuous conditionals with *mixtures of truncated exponentials* (MTE) as suggested by Moral et al. (2001). More on this follows in Section 5.3.3.

5.3.2 Automatic Discretization for Structure Learning

To keep the information loss during the discretization small the number of the intervals and their boundaries have to be chosen carefully. A fine-grained discretization will result in a sparsely connected BN due to regularization constrains, while a rough discretization may not provide the required degree of resolution.

Instead of discretizing as pre-processing step to the Bayesian network learning, we use a multivariate discretization approach, that accounts for the BN structure. The "optimal" discretization, Λ, thus depends on the DAG of the BN and the observed (continuous) data, $\mathbf{d}^c$. Analogue to Section 5.2.2, we again cast the problem in a Bayesian framework searching for the combination of (DAG, $\boldsymbol{\theta}$, Λ) that has the highest posterior probability given continuous data,

$$P(DAG,\boldsymbol{\Theta},\Lambda|\mathbf{d}^c) \propto P(\mathbf{d}^c|DAG,\boldsymbol{\Theta},\Lambda)P(DAG,\boldsymbol{\Theta},\Lambda). \tag{5.3}$$

Λ is defined by a set of interval boundary points for all variables and bins the original data $\mathbf{d}^c$ yielding $\mathbf{d}$. Expanding on an idea by Monti and Cooper (1998) we assume that all communication/information floss between the variables can be captured by their discrete representations. For given $\mathbf{d}$ and Λ is $\mathbf{d}^c$ consequently independent of DAG and $\boldsymbol{\Theta}$ and the likelihood for observing $\mathbf{d}^c$ (for a given discretization, network structure and parameters) can be written as,

$$P(\mathbf{d}^c|DAG,\boldsymbol{\Theta},\Lambda) = P(\mathbf{d}^c|\mathbf{d},\Lambda)P(\mathbf{d}|DAG,\boldsymbol{\Theta},\Lambda) \tag{5.4}$$

and Eq. (5.3) decomposes into

$$P(DAG,\boldsymbol{\Theta},\Lambda|\mathbf{d}^c) \propto \underbrace{P(\mathbf{d}^c|\mathbf{d},\Lambda)}_{\text{continuous data}} \quad \underbrace{P(\mathbf{d}|DAG,\boldsymbol{\Theta},\Lambda)}_{\text{likelihood (discrete)}} \underbrace{P(DAG,\boldsymbol{\Theta},\Lambda)}_{\text{prior}}.$$

The likelihood term is defined as for the separate BN learning for discrete data (Section 5.2.2) and we use a non-informative prior again. For the continuous data we assume that all continuous observations within the same, by Λ defined interval have the same probability. More information about the score definition can be found in the Appendix 5.A.1 and technical details are given in (Vogel et al., 2012b, 2013). In the following we discuss the BN and discretization learned from the synthetic seismic data set.

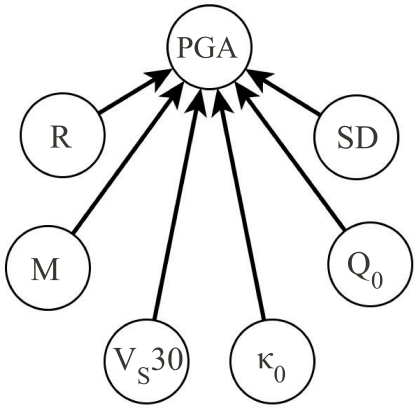


Figure 5.3: Theoretic BN, capturing dependencies of the data generating model.

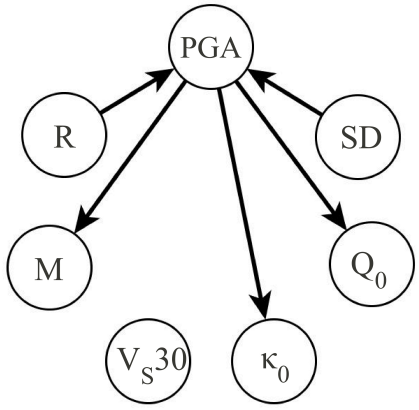


Figure 5.4: BN learned from the generated synthetic data.

Learned Ground Motion Model

Since we generated the data ourselves, we know which (in)dependencies the involved variables should adhere to; this is expected to be reflected in the BN DAG we learn from the synthetic data. Due to data construction are the predictors variables $M, R, SD, Q_0, \kappa_0, V_S30$ independent from each other and PGA depends on the predictors. Figure 5.3 shows the dependence structure of the variables. The converging edges at PGA indicate that the predictors become conditional dependent for given PGA. This means for given PGA they carry information about each other, e.g. for an observed large PGA value a small stress drop indicates a close distance to the earthquake. The knowledge about the dependence relations gives the opportunity to use the seismic hazard application for an inspection of the BN learning algorithm regarding to the reconstruction of the dependencies from the data, which is done in the following.

The network we found to maximize $P(DAG, \mathbf{\Theta}, \Lambda|\mathbf{d}^c)$ for the 10.000 synthetic, seismic data records is shown in Fig. 5.4. The corresponding found discretization is plotted in Fig. 5.5, which shows the marginal distributions of the discretized variables. The learned BN differs from the original one, mainly due to regularization constraints as we will explain in the following: As mentioned in Section 5.2 the joint distribution of all variables can be decomposed into the product of the conditionals according to the network structure. For discrete/discretized variables the number of parameters needed for the definition of $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$ in Eq. (5.1) corresponds to the number of possible state combinations for $(X_i, \mathbf{X}_{Pa(i)})$. Taking the learned discretization shown in Fig. 5.5, the BN of the data generating process (Fig. 5.3) is defined by 3858 parameters, 3840 needed alone for the description of $\mathsf{p}(PGA|M, R, SD, Q_0, \kappa_0, V_S30)$. A determination of that many parameters from 10.000 records would lead to a strongly over-fitted model. Instead we learn a BN, that compromises between model complexity and its ability to generate the original data. The BN learned under these requirements (Fig. 5.4) consists of only 387 parameters and still captures the most relevant dependencies.

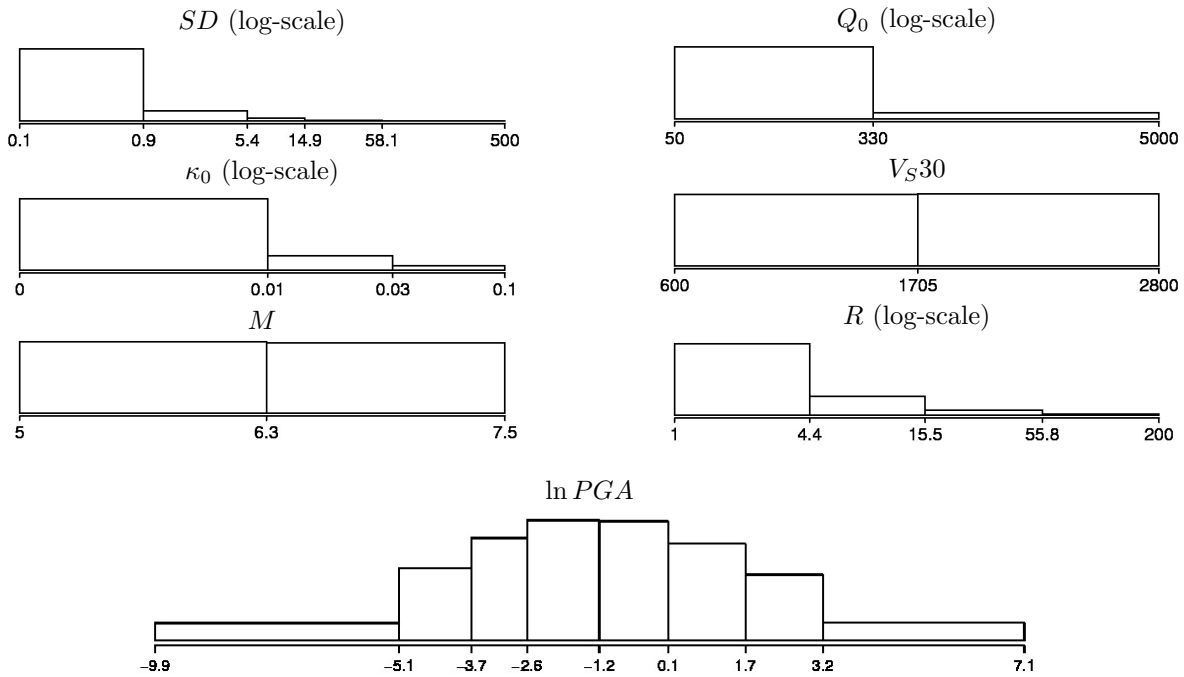


Figure 5.5: Marginal distribution of the variables, discretized according to the discretization found for the BN in Fig. 5.4. The number of intervals per variable ranges from 2 to 8.

Figure 5.6 shows the $\ln PGA$ values of the data set plotted against the single predictors. A dependence on stress drop (SD) and distance (R) is clearly visible. These are also the two variables with remaining converging edges on PGA, revealing that for given PGA SD contains information about R and vice versa. The dependencies between PGA and the remaining predictors are much less distinctive, such that the conditional dependencies between the predictors are negligible and the edges can be reversed for the benefit of parameter reduction. The connection to V_S30 is even neglected at all, since its impact on PGA is of minor interest compared to the variation caused by the other predictors.

Note, that the DAG of a BN actually expresses the independencies (not the dependencies) between the variables (Prop. 2). This means each (conditional) independence statement encoded in the DAG must be true, while encoded dependence relations must not hold per se. In turn this implies that each dependence holding for the data should be encoded in the DAG. The learning approach applied here fulfills the task quite well, detecting the relevant dependencies, while keeping the model complexity at a moderate level.

The model complexity depends not only on the DAG, but also on the discretization. A complex DAG will enforce a small number of intervals and a large number of intervals will only be chosen for variables with a strong influence on other variables. This effect is also visible for the learned discretization (Fig. 5.5). PGA is split into 8 intervals, distance and stress drop into 4 and 5, while the other variables consist of only 2 to 3 intervals.

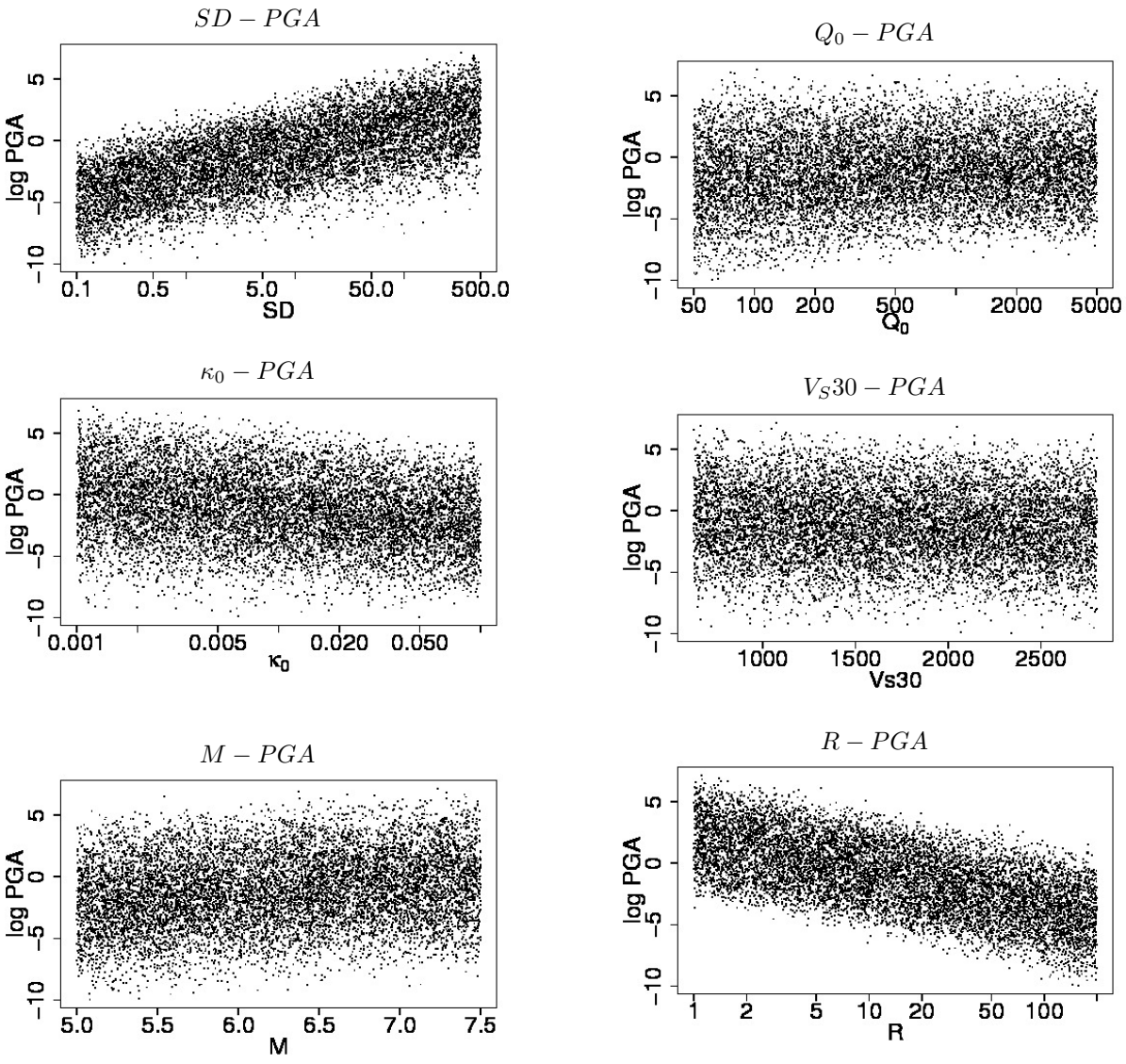


Figure 5.6: Dependencies between the predictor variables and $\ln PGA$.

5.3.3 Approximation of continuous distributions with Mixtures of Exponentials

A major purpose of the ground motion model is the prediction of the ground motion ($\ln PGA$) based on observations of the predictors; hence, although the BN captures the joint distribution (Prop. 5) of all involved variables, primary focus is in this context on a single variable, for which the accuracy of the prediction made with a BN is limited by the resolution of the discretization learned. For the BN shown above, the discretization of the target variable into 8 intervals enables a quite precise approximation of the continuous distribution, but this is not necessarily always the case. Complex network structures and smaller data sets used for the BN learning lead to a coarser discretization of the variables. To still enable precise estimations we may search for alternative approximations of the (or at least some, in particular the primary variable(s) of interest) continuous conditional distributions, once the BN has been learned.

Moral et al. (2001) suggest to use *mixtures of truncated exponentials* (MTEs) for this purpose, since they allow for the approximation of a variety of functional shapes with a limited number of parameters (Langseth et al., 2008) and they are closed under the operations used for BN inference: restriction, combination, marginalization (Langseth et al., 2009b). To approximate a conditional distribution $\mathsf{p}(X_i|\mathbf{X}_{Pa(i)})$ with MTEs, we partition the domain $\Omega_{(X_i,\mathbf{X}_{Pa(i)})}$ into hypercubes $D_1,\ldots,D_L$ and define the density within each hypercube, D_l, such that it follows the form

$$\mathsf{p}_{\downarrow D_l}(X_i|\mathbf{X}_{Pa(i)}) = a_0 + \sum_{j=1}^{J} a_j \, \mathrm{e}^{\,b_j X_i + c_j^T \mathbf{X}_{Pa(i)}}.$$

The determination of the hypercubes and the number of exponential terms in each hypercube as well as the estimation of the single parameters is done according to the maximum likelihood approach described in (Langseth et al., 2010). In the following we show how the MTE approximation improves the BN prediction performance compared to the usage of the discretized variables and we compare the results to those from a regression approach.

Prediction Performance

We conduct a 10-fold cross validation to evaluate the prediction performance of the BN compared to the regression approach: the complete dataset is divided into 10 disjoint subsamples of which in each trial one is defined as test set, while the others are used to learn the model (regression function or BN). The functional form of the regression function is determined by expert knowledge based on the description of the Fourier spectrum of seismic ground motion and follows the form

$$f(\mathbf{X}_{-Y}) = a_0 + a_1 M + a_2 M \cdot \ln SD + (a_3 + a_4 M) \ln \sqrt{a_5^2 + R^2} + a_6 \kappa R + a_7 V_S 30 + a_8 \ln SD$$

with $\kappa = \kappa_0 + t^*$, $\quad t^* = \frac{R}{Q_0 V_{sq}}$ $\quad$ and $\quad V_{sq} = 3.5 \frac{km}{s}$.

Averaged conditional density		
$BN_{discrete}$	BN_{MTE}	$regression$
0.237	0.320	0.331
0.240	0.297	0.329
0.239	0.298	0.331
0.218	0.255	0.323
0.216	0.260	0.339
0.222	0.257	0.339
0.215	0.252	0.332
0.243	0.317	0.330
0.212	0.249	0.328
0.243	0.315	0.331

Table 5.2: Results of a 10-fold cross validation to test the prediction performance of the BN (discrete and MTE approximations of the conditional distributions) and the regression approach. It shows for each trial the average of the conditional density for the observed $\ln PGA$ value.

Mean squared error		
$BN_{discrete}$	BN_{MTE}	$regression$
1.021	0.749	0.663
1.197	0.963	0.680
1.082	0.821	0.673
1.262	0.951	0.723
1.201	0.851	0.629
1.298	1.059	0.625
1.297	1.077	0.672
1.149	0.713	0.701
1.343	1.161	0.692
1.169	0.841	0.666

Table 5.3: Results of a 10-fold cross validation to test the prediction performance of the BN (discrete and MTE approximations of the conditional distributions) and the regression approach. It shows for each trial the mean squared error of the predicted $\ln PGA$.

We compare the regression approach in terms of prediction performance to the BN with discretized variables and with MTE approximations. For this purpose we consider for each model the conditional density distributions of $\ln PGA$ given the other variables and inspect how much probability each density assigns to the real $\ln PGA$ value. For the regression approach the conditional density follows a normal distribution, $\mathcal{N}(f(\mathbf{X}_{-Y}), \sigma^2)$, while it is defined via the DAG and the parameters $\boldsymbol{\theta}$ using the BN-models. Table 5.2 shows for each test set the conditional density of the observed $\ln PGA$ averaged over the individual records. An other measure for the prediction performance is the mean squared error of the estimates for $\ln PGA$ (Tab. 5.3). We define the mean values of the conditional densities as point estimates for $\ln PGA$. For the regression model the estimate i.e. corresponds to $f(\mathbf{x}_{-Y})$.

Even though the discretization of $\ln PGA$ is relative precise in the discrete BNs (8 intervals in each trial, except for the first trial, where $\ln PGA$ is split into 7 intervals), the MTE approximation of the continuous conditional distributions improves the prediction performance of the BN. Still it does not entirely match the precision of the regression function. Anyhow, the prediction performances are in the same order of magnitude and we must not forget, that the success of the regression approach relies on the expert knowledge used to define its functional form, while the structure of the BN is learned totally data driven. Further profits the regression approach in this example from the fact that the target variable ($\ln PGA$) is normally distributed, which is not necessarily the case for other applications. Focusing on the prediction of the target variable the regression approach also does not have the flexibility of the BN, which is designed to capture the joint distribution of all variables and thus allows for inference into all directions (Prop. 5), as exemplified in Section 5.4.3. Additional benefits of BNs, like their ability to make use of incomplete observations, will be revealed in the following sections, where we investigate real-world data.

5.4 Flood Damage Assessment

In the previous section we dealt with a fairly small BN (few variables/nodes) and a synthetic data set. In this section we go one step further and focus on learning a larger BN from real-life observations on damage caused to residential buildings by flood events. Classical approaches, so called stage-damage functions, relate the damage for a certain class of objects to the water stage or inundation depth, while other characteristics of the flooding situation and the flooded object are rarely taken into account (Merz et al., 2010). Even though it is known that the flood damage is influenced by a variety of factors (Thieken et al., 2005), stage-damage functions are still widely used. This is due to the fact that the number of potential influencing factors is large and the single and joint effects of these parameters on the degree of damage are largely unknown.

5.4.1 Real-life Observations

The data collected after the 2002 and 2005/2006 flood events in the Elbe and Danube catchments in Germany (Thieken et al., 2005; Elmer et al., 2010) offer a unique opportunity to learn about the driving forces of flood damage from a BN perspective. Figure 5.7 shows the investigated catchments. The data set consists of 1135 records of discrete and continuous variables describing the flooding situation, building and household characteristics, precaution, warning and emergency measures and building damage. Table 5.4 lists the 29 considered variables allocated to their domains. In Section 5.3.2 we already dealt with the issue of continuous data when learning BNs; here we will apply the methodology presented there. However, in contrast to the synthetic data from the previous section, many real world data sets are for different reasons lacking some observations for various variables, and for the data set at hand the percentage of missing values is below 20% for most variables yet for others it reaches almost 70%. In the next subsection we show how we deal with the missing values in the setting of the automatic discretization described in Section 5.3.2 when learning BNs.

5.4.2 Handling of Incomplete Records

To learn the BN we again maximize the joint posterior for the given data, Eq. (5.3). The calculation of the joint posterior requires for each variable X_i the number of counts for each combination of states for $(X_i, \mathbf{X}_{Pa(i)})$. However this is only given for complete data and for missing values it can only be estimated by using expected completions of the data. We note that a reliable and unbiased treatment of incomplete data sets (no matter which principled method is applied) is only possible for missing data mechanisms that are *ignorable* according to the *missing (completely) at random* (M(C)AR) criteria as defined in (Little and Rubin, 1987), i.e. the absence/presence of a data value is independent of the unobserved data. For the data sets considered in this paper we assume the MAR criterion to hold and derive the predictive function/distribution based on the observed part of the data in order to estimate the part which is missing.

Variable	Scale and range
flood parameters	
Water depth	C: 248 cm below ground to 670 cm above ground
Inundation duration	C: 1 to 1440 h
Flow velocity indicator	O: 0=still to 3=high velocity
Contamination indicator	O: 0=no contamination to 6=heavy contamination
Return period	C: 1 to 848 yrs
warning and emergency measures	
Early warning lead time	C: 0 to 336 h
Quality of warning	O: 1=receiver of warning knew exactly what to do to 6=receiver of warning had no idea what to do
Indicator of flood warning source	N: 0=no warning to 4=official warning through authorities
Indicator of flood warning information	O: 0=no helpful information to 11=many helpful information
Lead time period elapsed without using it for emergency measures	C: 0 to 335 h
Emergency measures indicator	O: 1=no measures undertaken to 17=many measures undertaken
precaution	
Precautionary measures indicator	O: 0=no measures undertaken to 38=many, efficient measures undertaken
Perception of efficiency of private precaution	O: 1=very efficient to 6=not efficient at all
Flood experience indicator	O: 0=no experience to 9=recent flood experience
Knowledge of flood hazard	N (yes / no)
building characteristics	
Building type	N (1=multifamily house, 2= semi-detached house, 3=one-family house)
Number of flats in building	C: 1 to 45 flats
Floor space of building	C: 45 to 18000 m^2
Building quality	O: 1=very good to 6=very bad
Building value	C: 92244 to 3718677 €
socio-economic factors	
Age of the interviewed person	C: 16 to 95 yrs
Household size, i.e. number of persons	C: 1 to 20 people
Number of children (< 14 years) in household	C: 0 to 6
Number of elderly persons (> 65 years) in household	C: 0 to 4
Ownership structure	N (1=tenant; 2=owner of flat; 3=owner of building)
Monthly net income in classes	O: 11=below 500 €to 16=3000 €and more
Socioeconomic status according to Plapp (2003)	O: 3=very low socioeconomic status to 13=very high socioeconomic status
Socioeconomic status according to Schnell et al (1999)	O: 9=very low socioeconomic status to 60=very high socioeconomic status
flood loss	
rloss − loss ratio of residential building	C: 0 = no damage to 1 = total damage

Table 5.4: Variables used for the flood damage assessment. C: continuous, O: ordinal, N: nominal.

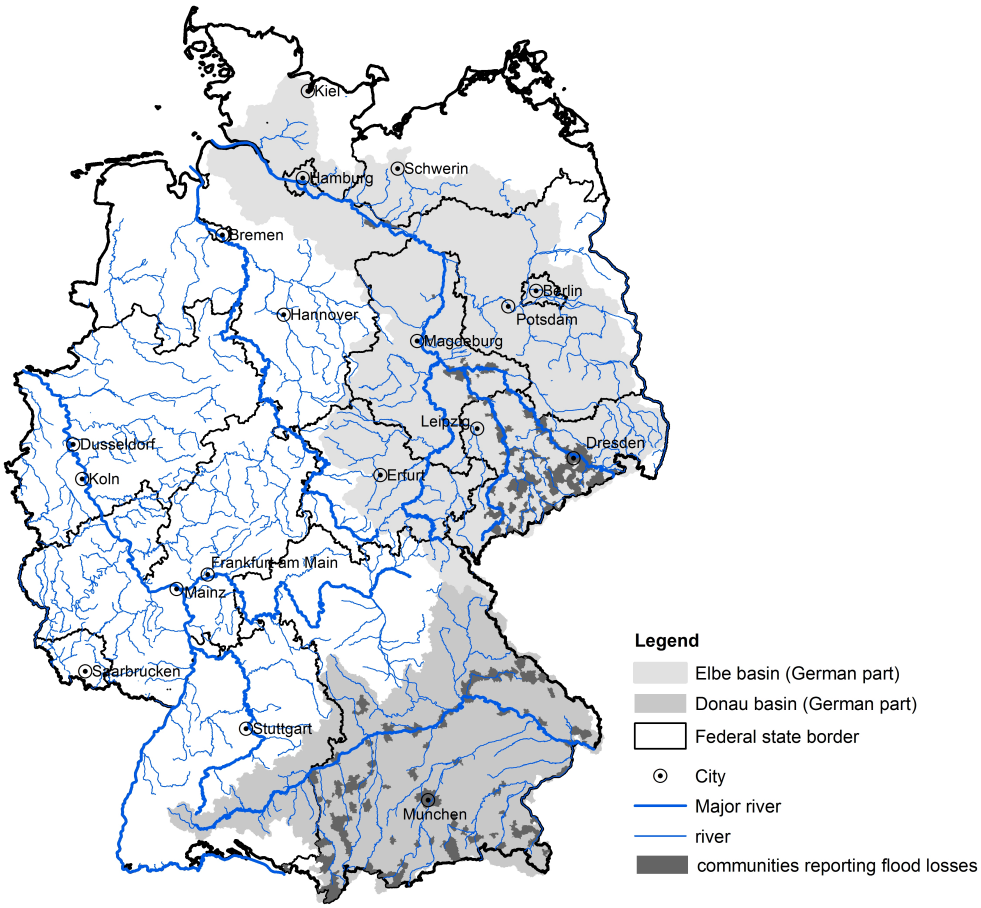


Figure 5.7: Investigated catchments and location of communities reporting losses from the 2002, 2005 and 2006 flood events in the Elbe and Danube catchments.

In the context of BNs a variety of approaches has been developed to estimate the missing values (so-called "imputation"). Most of these principled approaches are iterative algorithms based on Expectation-Maximization e.g. (Friedman, 1997, 1998) or stochastic simulations e.g. (Tanner and Wong, 1987). In our case we already have to run several iterations of BN learning and discretization, each iteration requiring the estimation of the missing values. Using an iterative approach for the missing value prediction will thus easily become infeasible. Instead we use a more efficient albeit approximate method, using the *Markov Blanket Predictor* developed by Riggelsen (2006).

The idea is to generate a predictive function which enables for predicting a missing variable X_i based on the observations of its Markov Blanket (MB), $\mathbf{X}_{MB(i)}$. The Markov Blanket identifies the set of variables directly influencing X_i, i.e. parents, children and parents of children. For illustration see Fig. 5.8, left. Assuming the MB is fully observed it effectively blocks influence from all other variables, i.e. the missing value depends only on its MB. When some of the

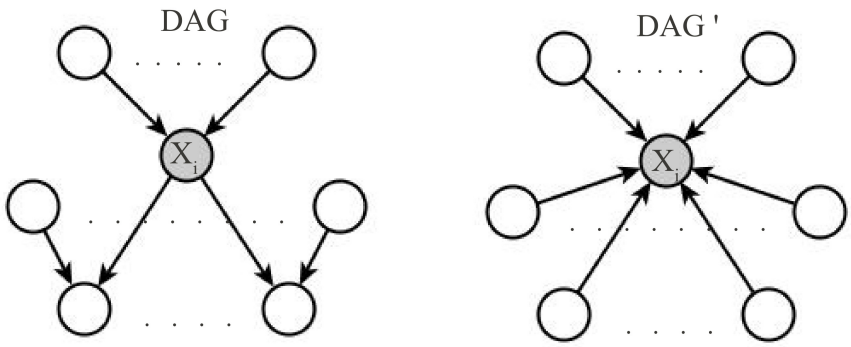


Figure 5.8: left: Markov Blanket of X_i; right: new model DAG' derived from DAG (all edges present in the Markov Blanket of X_i in DAG are directed towards X_i in DAG').

variables in the MB are missing, it does not shield off X_i; although maybe not immediately clear, this follows directly from the BN factorization of the joint distribution. However, for predictive approximation purposes we choose to always ignore the impact from outside the MB. Hence, the predictions for X_i based on observed data factorizes according to the DAG in Fig. 5.8, left, as,

$$P(X_i|\mathbf{X}_{MB(i)}, \boldsymbol{\theta}, DAG) \propto \theta_{X_i|\mathbf{X}_{Pa(i)}} \prod_{j \in Ch(i)} \theta_{X_j|\mathbf{X}_{Pa(j)}}, \tag{5.5}$$

where $Ch(i)$ are the variable indices of the children of X_i. The prediction of X_i requires inference in the BN (albeit very simple) where correct estimates of all $\boldsymbol{\theta}$ are assumed. However, these in general can't be given without resorting to iterative procedures. To avoid this we first define a slightly modified version of the predictive function based on DAG' which is derived from DAG by allocating edges, pointing to X_i, to each variable in its MB; see Fig. 5.8. The resulting DAG' preserves all dependencies given in DAG and can alternatively be used for the prediction of X_i,

$$P(X_i|\mathbf{X}^{DAG'}_{Pa(i)}, \boldsymbol{\theta}^{DAG'}, DAG') \stackrel{def}{=} \theta^{DAG'}_{X_i|\mathbf{X}_{Pa(i)}}. \tag{5.6}$$

For this predictive distribution we need to estimate the parameters $\theta^{DAG'}_{X_i|\mathbf{X}_{Pa(i)}}$. However, this we can approximately estimate without using iterative methods by a *similar cases* approach[3], described in Appendix 5.A.2. A detailed description for the generation of the predictive distribution is given in (Riggelsen, 2006; Vogel et al., 2013).

To avoid confusions, we want to stress that the MBs of the individual variables change during the BN-learning procedure. The prediction of the missing values has to be updated for each change of the DAG.

[3]Note that more parameters are required for the derived predictive distribution Eq. (5.6), but now at least *all* influencing variables are considered jointly, whereas in the former case Eq. (5.5) we would have estimated the parameters based on the parent sets only if a non-iterative similar cases approach would have been used.

5.4.3 Results

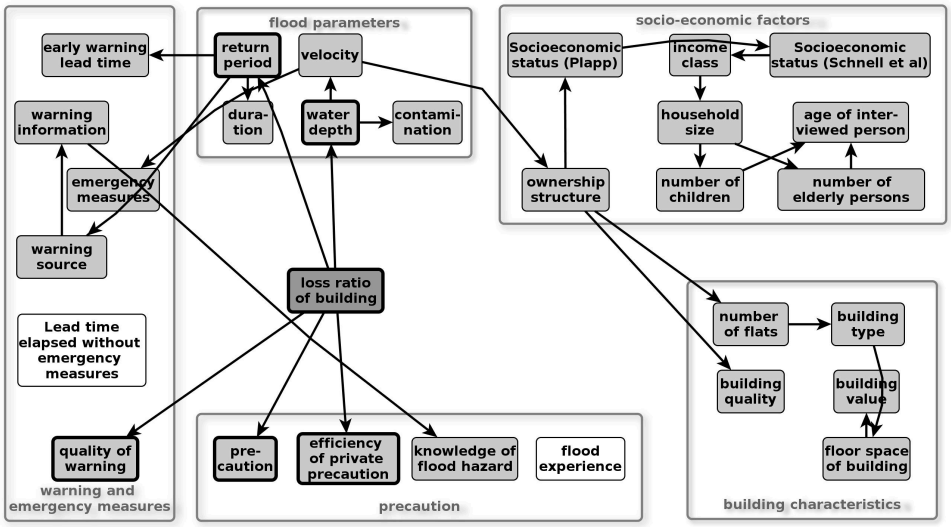
Coming back to the flood damage data, we have three variables with more than one third of the observations missing: flood experience (69% missing), warning quality (56% missing) and lead time elapsed without emergency measures (54% missing). In a first "naive" application (Vogel et al., 2012b) no special attention was on a proper treatment of missing values; the missing values were simply randomly imputed resulting in the isolation of two variables (flood experience and lead time elapsed) in the network; no connection to any other variable was learned (Fig. 5.9, top). Applying the Markov Blanket predictor the situation changes and a direct connection from the relative building damage, *rloss*, to flood experience is found as well as a connection between warning source and elapsed lead time (Fig. 5.9, bottom). These relations, especially the first one, match with experts expectations and speak for an improvement of the learned BN structure.

Using the graphical representation (Prop. 1), as mentioned in Section 5.2.1 the learned DAG (Fig. 5.9, bottom) gives insight into the dependence relations of the variables. It reveals a number of direct links connecting the damage describing variable with almost all subdomains. This supports the demand for improved flood damage assessments that take several variables into account (Merz et al., 2010). Moreover, the DAG shows which variables are the most relevant for the prediction of *rloss*. Especially the domains 'precaution' and 'flood parameters' are densely connected with the building damage and should be included in any damage assessment (Prop. 3).

Existing approaches for flood damage assessments usually consider fewer variables and an employment of a large number of variables is often considered as disadvantageous, since complete observations for all involved variables are rare. The requirement for complete observations does not hold for BNs (Prop. 6). The prediction of the building damage, e.g., depends only on the variables of its Markov Blanket (marked with a bold frame in Fig. 5.9). Is the observation of the Markov Blanket variables incomplete (not all variables are observed at inference time), information from outside the Markov Blanket 'flows' into the prediction by indirectly marginalizing (summing) missing variables out. The use of many variables thus now provides additional knowledge and proves to be an advantage.

The capability of BNs to predict from incomplete observations enables us to make predictions at an early stage of an event, employing only the information that is present at any given time. The prediction can subsequently be updated as new information becomes available.

BN learned with randomly replaced missing values



BN learned with missing values predicted from Markov Blanket

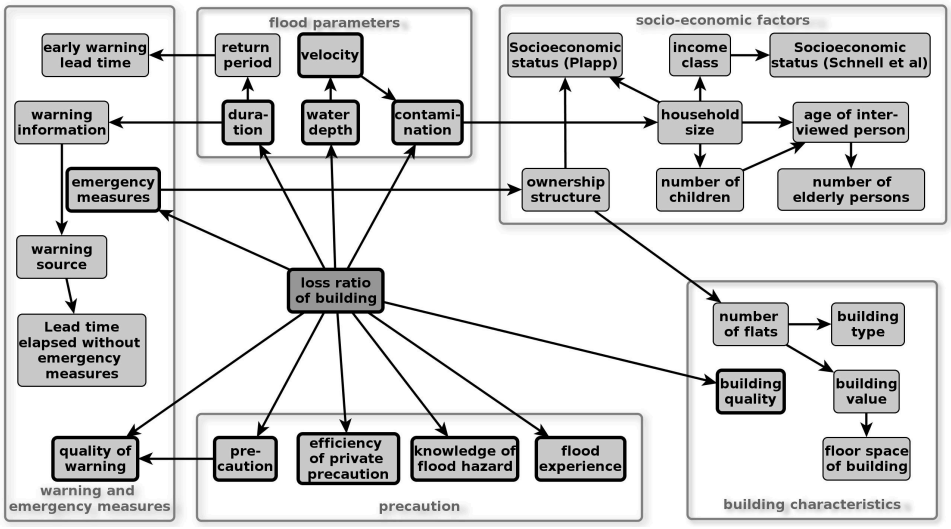


Figure 5.9: BNs learned for flood damage assessment. Top: missing values are randomly replaced. Bottom: Missing values are estimated using the Markov Blanket predictor. Nodes with a bold frame belong to the Markov Blanket of relative building loss and are thus assumed to have direct impact on the flood loss.

	sdf	FLEMOps+r	BN_{MTE}
1	0.0111	0.0108	0.0104
2	0.0133	0.0114	0.0116
3	0.0161	0.0145	0.015
4	0.0200	0.0194	0.0169
5	0.0166	0.0150	0.0163
Avg	0.0154	0.0142	0.014

Table 5.5: Mean squared errors of a 5-fold cross-validation, where the relative building loss is predicted with the three models stage-damage function, FLEMOps+r and Bayesian Networks.

Prediction Performance

The prediction of the relative building loss is of primary interest in flood damage assessments. Similar to our proceeding for the ground motion modeling, we approximate the distribution of the target variable with mixtures of truncated exponentials, thus achieving a better resolution for the distribution of interest. The resulting prediction performance of the BN is compared to currently used flood damage assessment approaches, namely the stage-damage function (sdf) and the FLEMOps+r model (Elmer et al., 2010), which was developed from the same data set, estimating the building damage based on water depth, flood frequency, building type, building quality, contamination and private precaution. While sdf and FLEMOps+r give point estimates, the BN gives us a distribution for *rloss* and thus reveals the uncertainty of the prediction (Prop. 4). Especially when it comes to decision making, the identification of uncertainty is a major advantage of the BN. However, to allow for model comparison, we reduce the distribution provided by the BN to its mean value, which we define to be the estimate of *rloss*. Table 5.5 shows the mean squared error of a 5-fold cross validation for the three model approaches. The prediction performance of the BN is comparable to the one of the FLEMOps+r, while the BN has the additional advantage to model the whole distribution of the target variable and to conduct the prediction even though not all variables are observed.

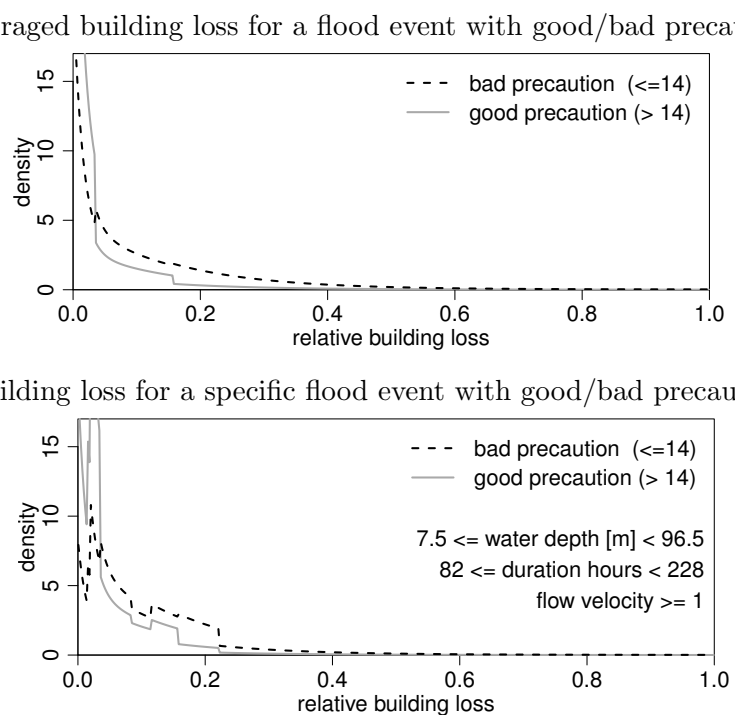


Figure 5.10: Top: Distribution of the building loss depending on the precaution. Flood specific parameters as well as other parameters are unknown and summed out. Bottom: Distribution of the building loss depending on the precaution for a specific flood situation, where water depth, duration and flow velocity are known. Other parameters are unknown and summed out.

Example for Inference: Impact of Precaution

As an example of reasoning (Prop. 5), we consider the effect of precaution on the building loss. Figure 5.10 shows the distribution of the building loss for a good precaution (precautionary measures indicator > 14) and a bad precaution (precautionary measures indicator ≤ 14) in a general case (top: all other variables are unknown and summed out) and for a specific flood event (bottom: $7.5\,m \leq$ water depth $< 96.5\,m$; $82\,h \leq$ duration $< 228\,h$; $1 \leq$ velocity). We may appreciate, how a good precaution increases the chance for no or only small building losses.

Similar investigations may support the identification of efficient precautionary measures, not only in the context of flood events, but for natural hazards in general. They may also help to convince authorities or private persons to undertake the suggested precautions. Using the flexibility of BNs and their ability to model specific situations, BNs may thus contribute to a better communication between scientists and non-scientific stakeholders. BNs can also be used for forensic reasoning, i.e. we can turn around the direction of reasoning in the just considered example and ask for a given observed damage in a specific or general event situation, what is a likely state of precaution. Forensic reasoning might be of interest e.g. for insurance companies.

5.5 Landslides

So far we assumed the existence of a unique model that explains the data best, but in practical problems there may be many models almost as good as the best, i.e. explaining the data similarly well. This results in an uncertainty about which BN structure to use. We consider this problem in our last application, where we work on landslides, which are an other ubiquitous natural hazard in many parts of the world.

A key theme in many landslide studies is the search for those geological, hydroclimatological, topographic and environmental parameters that are suitable for sufficiently predicting the susceptibility to slope failure in a given region. A wide range of multivariate data analysis techniques has been proposed to meet this challenge. Amongst the more prominent methods are logistic regression, artificial neural networks, and Bayesian Weights-of-Evidence. The popularity of such methods is only matched by their seeming success: A recent review of 674 scientific papers on the topic indicates that most reported success rates are between 75% and 95%, which raises the question why landslides still continue to cause massive losses despite this seemingly high predictive accuracy (Korup and Stolle, 2013). Moreover, success rates do not show any significant increase over the last ten years regardless of the number of landslide data or predictors used. An often overlooked key aspect in these analyses is the potential for correlated or interacting predictor candidates. Few studies have stringently explored whether this likely limitation is due to physical or statistical (sampling) reasons.

5.5.1 Data

We use a database of nearly 300,000 landslide deposit areas that derive mostly from deep-seated slow-moving failures and earthflows throughout Japan. This inventory was compiled by the National Research Institute for Earth Science and Disaster Prevention NIED (http://lsweb1.ess.bosai.go.jp/gis-data/index.html), and is one of the largest of its kind available. We have compiled a number of geological, climatic, and topographic metrics throughout the Japanese islands to test their influence on the average fraction of landslide-affected terrain that we computed within a 10-km radius. Most of our candidate predictors have been used in modified form in other studies (Tab. 5.6). While all of these candidate predictors may be physically related to slope instability, part of this choice of predictor composition is intentionally arbitrary in order to learn more about its effects on BN learning and structure. The final data set used for the BN learning consists of 553 records, where $\sim 0.4\%$ of the data are missing.

Name	Definition	Unit
Mean elevation	Average of elevation values within catchment boundaries	$[m]$
Catchment area	Log-transformed catchment area	$[a.u.]$
Catchment perimeter	Total length of catchment divides	$[m]$
Mean local topographic relief	Maximum elevation difference in a 10-km radius	$[m]$
Mean annual precipitation[1]	Based on interpolated rainfall station data (reference period 1980-2010)	$[mm]$
Mean coefficient of variation of annual precipitation[1]	Based on interpolated rainfall station data, with standard deviation divided by mean (reference period 1980-2010)	[1]
Mean coefficient of variation of monthly precipitation[1]	Based on interpolated rainfall station data, with standard deviation divided by mean (reference period 1980-2010)	[1]
Mean surface uplift 2001-2011[2]	GPS-derived accumulated surface uplift between 2001-2011	$[m]$
Mean surface uplift 2010-2011[2]	GPS-derived accumulated surface uplift between 2010-2011	$[m]$
Mean fraction of 10% steepest bedrock channels	Average fraction of 10% steepest channels per unit length of bedrock-river drainage network in a 10-km radius, based on an arbitrarily set reference concavity $\theta = 0.45$	[1]
Mean bedrock channel steepness	Average of channel steepness index per reach length, based on an arbitrarily set reference concavity $\theta = 0.45$	[1]
Regionalised river sinuosity	Average bedrock-channel sinuosity weighted by drainage network length in a 10-km radius calculated as the flow length of a given channel segment divided by its shortest vertex distance	[1]
Fraction of volcanic rocks[3]	Fraction of catchment area underlain by volcanic rocks	[1]
Fraction of lakes	Fraction of catchment area covered by lakes	[1]
Fraction of plutonic rocks[3]	Fraction of catchment area underlain by plutonic rocks	[1]
Fraction of sedimentary rocks[3]	Fraction of catchment area underlain by sedimentary rocks	[1]
Fraction of accretionary complex rocks[3]	Fraction of catchment area underlain by accretionary complex rocks	[1]
Fraction of metamorphic rocks[3]	Fraction of catchment area underlain by metamorphic rocks	[1]
Median area of landslide-affected terrain	Fraction of landslide terrain per unit catchment area within a 10-km radius calculated using an inventory of mostly prehistoric landslide-deposit areas	[1]

[1]Calculated using data provided by the Japan Meteorological Agency (JMA, www.jma.gov)
[2]Calculated from secular high-precision levelling data (Kimura et al., 2008)
[3]Calculated using the seamless digital geological map of Japan (1:200,000) available from the Geological Survey of Japan (https://gbank.gsj.jp/seamless).

Table 5.6: Variables used for the landslide model.

5.5.2 Uncertainty in BN structure

Ideally, a given model should adequately encapsulate natural phenomena such as the causes and triggers of slope instability. However, there may be several equally well poised, but competing, models because of the intrinsic uncertainty tied to the governing processes. In practice we also face other limitations that prevent us from focusing on one single best model. The finite number of observations we have at our disposal for learning, and the fact that it is unclear which relevant predictor variables to consider for landslide prediction implies that several models may be justifiable. This is a general problem when attempting to formally model natural systems. In our case this means that several BNs might explain the data (almost) equally well, i.e. receive a similar score according to Eq. (5.2).

An additional source of uncertainty can be attributed to the structure learning algorithm used to maximize the score defined in Eq. (5.2) or – for continuous variables – in Eq. (5.3). For infinite data sets the algorithm terminates according to *Meek's Conjecture* in the (unique) optimal equivalence class of DAGs (Chickering, 2002), but this does not necessarily hold for finite data sets, incomplete observations and a search space extended by the discretization. The algorithm for the traversal of the BN hypothesis space contains stochastic elements and may get stuck in local optima providing slightly different results for different runs.

To analyze this random behavior, we run the BN learning and discretization algorithm ten times on the same data set of landslide data. We do not expect to end up with the same BN in each trial, as the constraints to meet Meek's Conjecture are not fulfilled. Instead, we are more interested in documenting how strongly the results differ from each other.

Figure 5.11 gives a summarized representation of the BN DAG structures. The frequency with which an edge between two variables is learned is encoded by the widths (by scaling it accordingly). Despite of the difference in DAG structures, all learned BNs seem to model the data generating process almost equally well, which can be gathered from the score obtained by Eq.(5.3); for the BNs learned, we get a score between -64364.42 and -64253.98 . This is a promising result, since it indicates, that even though the algorithm gets stuck in local maxima, the quality of the results does not differ significantly. This supports the assumption that the quality of the learned BN is not seriously affected by random effects of the learning algorithm. Multiple runs of the algorithm on other data sets confirm this assumption.

In the literature on BN learning (and on model learning based on data in general) ideas of how to handle several competing, but all justifiable BNs have been investigated. Friedman et al. (1999) use bootstrap sampling to learn BNs from different variations of the data set. Based on those they develop a confidence measure on features of a network (e.g., the presence of an edge or membership of a node to a certain Markov Blanket). A Bayesian approach is presented by Friedman and Koller (2000) and Riggelsen (2005), who approximate the Bayesian posterior on the DAG space using a Markov Chain Monte Carlo approach. An adaptation of these methods for the extended MAP score introduced in this paper is left for future work.

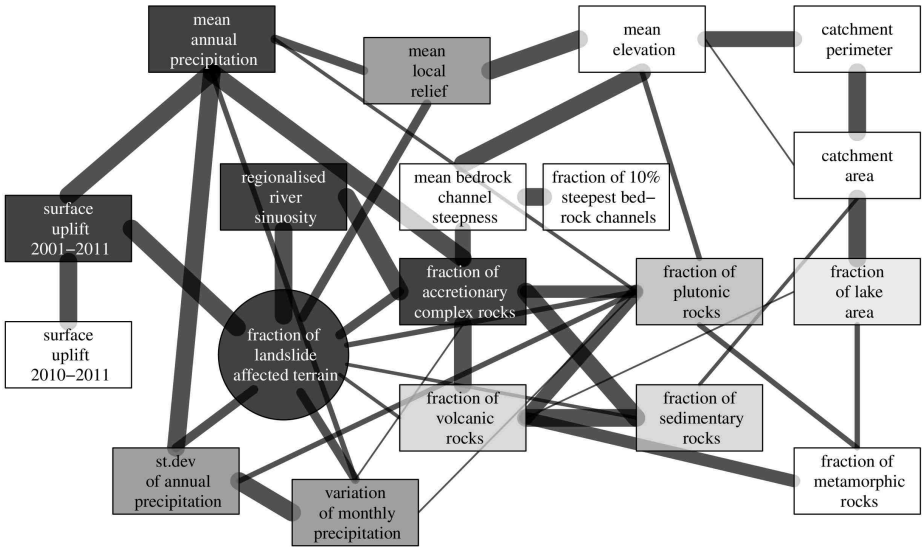


Figure 5.11: Summary of ten learned network structures, based on the same data set. Arrow widths between the variables are scaled to the number of times they occur in the learned BNs. Likewise, we color-coded the variables by the frequency with that they occur as part of the Markov Blanket of *fraction of landslide affected terrain* (circular node shape), where darker hues indicate more frequent occurrence.

5.5.3 Results

Despite of (or rather thanks to) the DAG structural differences we can glean some instructive insights from the learned BNs. The fact that we can learn something about the landslide-affected terrain from several BN structures indicates that the different predictors are highly interacting, and that a missed link between two variables can often be compensated by other interactions. To get an understanding which variables are most relevant for predicting landslide-affected terrain, we coded the variables in Fig. 5.11 by the frequency with that they occur as part of the target variable's Markov Blanket, where darker hues indicate more frequent occurrences.

Perhaps the most surprising aspect of the learned BNs is that only few of the predictors that have been invoked traditionally to explain landslide susceptibility are duly represented in the Markov Blanket. These include mean annual precipitation (part of the MB in each run) – including some derivatives such as precipitation variability (either annual or monthly variation is part of the MB) – and mean local topographic relief (part of the MB in half of the runs).

Instead, predictors such as regionalised bedrock river sinuosity or short-term (10-year cumulative) surface uplift derived from a dense network of GPS stations seem to provide relevant information about landslide-affected terrain in Japan. Bedrock river sinuosity may reflect the ability of rivers to carve more pronounced meanders in rocks with closely spaced defects. Therefore, sinuosity could be linked to first order to important rock-mass properties that govern the abundance of landslides. However, the link to contemporary surface uplift is less clear. Many of Japan's currently subsiding areas are limited to low-relief forearc areas, which feature

fewer landslides, hence an indirect link via topography seems plausible. Yet predictors such as mean elevation or bedrock channel steepness (as a proxy of fluvial erosion and undercutting of hillslopes) play largely subdued roles in the learned BNs part of the MB. Also, the role of lithology, expressed by the fractions of different rock types outcropping in a given area form a highly interacting cluster, where the information about accretionary complexes, i.e. heavily tectonized and welded remnants of former island arcs, seems to be of major importance for the landslide prediction. Alternatively it is either the fraction of plutonic, sedimentary or volcanic rocks that is part of the MB.

The learned BN structures are counter-intuitive compared to many other susceptibility models that traditionally emphasize hillslope inclination and topographic relief. Further studies must show, if the found dependencies are regional artifacts or valid on a larger scale. Nevertheless, these findings illustrate that the BN approach may reveal novel and unexpected insights into regional landslide prediction by highlighting unusual links between predictor variables that other multivariate models may not show as clearly. Equally important, BNs underscore which predictors may yield sufficient predictive potential should others not be available.

5.6 Conclusions

The Bayesian Network approach is a powerful framework to capture uncertainties and probabilistic elements in natural hazard assessments. We demonstrated its flexible applicability in seismic hazard, flood damage, and landslide susceptibility analyses. Alongside we discussed the handling of continuous data and incomplete observations as well as the uncertainty about the model structure, i.e. challenges that may arise when BNs are learned from real world data. Our suggested way of dealing with these problems is fully data driven and can thus easily be transfered to other domains.

Since the interest of most natural hazard assessment is in the prediction of a certain target variable, we compared the prediction performance of the BNs learned for the seismic hazard and flood damage application to currently used models. In both cases the BNs perform reasonable well. This is especially promising, since the BNs are designed to capture the joint distribution of all variables and thus put similar effort into the prediction of each variable, whereas alternative models focus on predicting the target variable solely. For a better prediction performance, we might think of different graphical models that share the focus on the target variable. These could include (tree augmented) Naive Bayes networks or an adapted score for the network learning that would weight more the target variable. Thus learned networks may also be more reliable in the identification of the variables relevant for the prediction, but will fail to capture the overall picture of dependence relations.

Working with BNs we profit from several attractive properties inherent to the BN framework. Learning the DAG structure and parameters from data, BNs require no prior domain knowledge. Yet if available, it can be exploited via the prior term, which is part of the scoring function. The discovered (in-)dependence relations help us to understand the underlying process and to identify (ir-)relevant variables. An intuitive understanding is supported by the graphical representation of BNs, although the same data may produce different graphs with comparable

performance. This highlights the potential for new insights into interactions between large sets of candidate predictors. The ability of BNs to predict from incomplete observations allows for hazard estimations at an early stage of an event. Using inference we can estimate missing values based on observations from related variables. The prediction can be updated as soon as new information about so far missing variables becomes available.

Capturing the uncertainty and providing a probability distribution instead of a point estimate, BNs provide a valuable contribution on the basis of which decision making should be made. Allowing for inference they also enable detailed examinations of specific scenarios. Thus Bayesian Networks may thus be used for an improved communication between scientist and public authorities and help for a better evaluation of hazards in general.

Acknowledgements

We acknowledge Bruno Merz, Heidi Kreibich and Kai Schröter from the GFZ Helmholtz Centre in Potsdam for providing their data and expertise about flood damage.

This work is supported and funded by PROGRESS, a research cluster of universities and external organizations in the region of Potsdam-Berlin (Germany). Additional funding comes from the DFG project RI 2037/2-1.

5.A Appendix

5.A.1 BN learning: MAP score

For a data based BN learning we search for the pair $(DAG, \boldsymbol{\theta})$ that has the highest probability for the observed data and thus maximizes the joint posterior

$$\underbrace{P(DAG, \boldsymbol{\Theta}|\mathbf{d})}_{\text{posterior}} \propto \underbrace{P(\mathbf{d}|DAG, \boldsymbol{\Theta})}_{\text{likelihood}} \underbrace{P(\boldsymbol{\Theta}, DAG)}_{\text{prior}},$$

which corresponds to the MAP score of the BN.

Investigating the single components of the score, we find that the likelihood term $P(\mathbf{d}|DAG, \boldsymbol{\Theta})$ is the product of Eq. (5.1) for every independent sample, hence, for complete data this is

$$P(\mathbf{d}|DAG, \boldsymbol{\theta}) = \prod_{l=1}^{|\mathbf{d}|} \prod_{i=1}^{k} \mathsf{p}(\mathrm{x}_i^{(l)}|\mathbf{x}_{Pa(i)}^{(l)}) = \prod_{i, x_i, \mathbf{x}_{Pa(i)}} \theta_{x_i|\mathbf{x}_{Pa(i)}}^{n(x_i, \mathbf{x}_{Pa(i)})}, \tag{5.7}$$

where $\mathrm{x}_i^{(l)}$ is the observed value of X_i in the l-th record and $\mathbf{x}_{Pa(i)}^{(l)}$ is the corresponding observation of the parent set. This reduces to a function of $n(x_i, \mathbf{x}_{Pa(i)})$, the number of counts (occurrences) of the values $(x_i, \mathbf{x}_{Pa(i)})$ in the data.

The joint prior distribution decomposes as $P(\boldsymbol{\Theta}, DAG) = P(\boldsymbol{\Theta}|DAG)P(DAG)$. For the applications in this paper we assume that all DAGs are equally likely and consequently define $P(DAG)$ to be uniform over the space of BN structures. $P(\boldsymbol{\Theta}|DAG)$ we define as a non-informative prior, acting at the same time as a penalty term. For discrete data this is a (product) Dirichlet distribution given by

$$P(\boldsymbol{\theta}|DAG) = \prod_{i=1}^{k} \prod_{\mathbf{x}_{Pa(i)}} \frac{\Gamma\left(\sum_{x_i} \alpha(x_i, \mathbf{x}_{Pa(i)})\right)}{\prod_{x_i} \Gamma\left(\alpha(x_i, \mathbf{x}_{Pa(i)})\right)} \prod_{x_i} \theta_{x_i|\mathbf{x}_{Pa(i)}}^{\alpha(x_i, \mathbf{x}_{Pa(i)})-1}, \tag{5.8}$$

where $\alpha(\cdot)$ are so-called hyper-parameters, primarily governing the regularization (to avoid over-fitting).

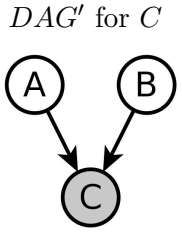
MAP score extension for continuous variables

Are continuous variables contained in the data set, we learn a discretization, Λ, parallel to the BN. The here used extended MAP score corresponds to the joint posterior of BN and discretization:

$$P(DAG,\boldsymbol{\Theta},\Lambda|\mathbf{d}^c) \propto P(\mathbf{d}^c|DAG,\boldsymbol{\Theta},\Lambda)P(DAG,\boldsymbol{\Theta},\Lambda),$$

Considering Eq. (5.4) we rewrite the equation above as

$$P(DAG, \boldsymbol{\Theta}, \Lambda|\mathbf{d}^c) \propto \underbrace{P(\mathbf{d}^c|\mathbf{d}, \Lambda)}_{\text{continuous data}} \underbrace{P(\mathbf{d}|DAG, \boldsymbol{\Theta}, \Lambda)}_{\text{likelihood (discrete)}} \underbrace{P(DAG, \boldsymbol{\theta}, \Lambda)}_{\text{prior}}.$$



Estimate $\hat{\theta}_{C=t|A=t,B=f}$:

$s(t,(t,f)) = \frac{7}{4}$

C	A	B	contribution to $s(t,(t,f))$
t	t	-	$\frac{1}{2}$
f	f	-	0
t	t	f	1
-	t	f	0
t	-	t	0
t	-	-	$\frac{1}{4}$

Figure 5.12: Illustration for the calculation of $s(\cdot)$ used for the parameter estimation in DAG'. The graph on the left shows a DAG' for the estimation of C depending A and B. The three variables take the values t and f. An exemplary data set is given in the table on the right together with the contribution for each record to $s\left((C=t,(A=t,B=f)\right)$.

The likelihood term is defined as in Eq. (5.7) and the prior decomposes into $P(\mathbf{\Theta}|DAG,\Lambda)P(\Lambda|DAG)P(DAG)$. For $P(\Lambda|DAG)$ and $P(DAG)$ we assume uniform distributions analogue as for the original MAP score and we define $P(\mathbf{\Theta}|DAG,\Lambda)$ as product Dirichlet again Eq. (5.8). For the continuous data term we define

$$P(\mathbf{d}^c|\mathbf{d},\Lambda) = \prod_i \prod_{x_i} \left(\frac{1}{n(x_i)}\right)^{n(x_i)},$$

which corresponds to the assumption, that all continuous observations are equally likely within the same interval. The joint posterior has a closed form as a function of $n(\cdot)$.

5.A.2 Similar cases approach

The estimation of a missing value as described in Section 5.4.2 requires, according to Eq. (5.6), the prediction of the parameter $\theta^{DAG'}_{X_i|\mathbf{X}_{Pa(i)}}$ from incomplete data. Instead of using the unobserved statistics $n(\cdot)$, we rely on counts of similar cases here. The statistics, $s(x_i,\mathbf{x}_{Pa(i)})$, is a weighted count of all records where $X_i = x_i$ and the observed part of the parents set $\mathbf{X}_{Pa(i)}$ matches with $\mathbf{x}_{Pa(i)}$; for each record where $(X_i,\mathbf{X}_{Pa(i)})$ is fully observed we add 1 to $s(x_i,\mathbf{x}_{Pa(i)})$. For an incomplete observed parents set we count the possible completions of the record and add 1/number-of-possible-completions to $s(x_i,\mathbf{x}_{Pa(i)})$. Refer to Fig. 5.12 for an example. We now estimate the parameters

$$\hat{\theta}^{DAG'}_{x_i|\mathbf{x}_{Pa(i)}} = \frac{s(x_i,\mathbf{x}_{Pa(i)})}{\sum_{x_i} s(x_i,\mathbf{x}_{Pa(i)})},$$

fully defining the predictive distribution for X_i Eq. (5.6).

CHAPTER

SIX

GENERAL CONCLUSIONS AND PERSPECTIVES

6.1 Summary and Context

In this thesis we presented graphical models, in particular Bayesian Networks (BNs), as flexible and powerful tools to capture and express uncertainties in natural hazard assessments. In all considered applications one or more BNs were learned from synthetically generated or observed data searching for the BN that explains the data best. There exist different interpretations, what "best" means in this context and accordingly different metrics for BN scoring are suggested in literature. A thorough treatment of different metrics can be found in Bouckaert (1995). Here we mention a few. The penalized (log-)likelihood approach stems from the frequentist perspective. The basic idea is to score the BN with its ability to fit the given database penalized by a term that accounts for the model complexity and favors simpler structures. The fit to the data is measured by the (log-)likelihood term, while the penalty term, e.g. the AIC criterion (Akaike, 1974) or the BIC criterion (Schwarz, 1978), usually includes the number of free model parameters. Another widely used approach, based on the Bayesian perspective, is the Bayesian Dirichlet scoring metric, often referred to as BD metric (Cooper and Herskovits, 1992; Heckerman et al., 1995). The BD metric makes use of the *marginal likelihood*, which is proportional to the posterior probability of the network structure given the data. It thus acts as a measure for the quality of the BN structure. The scoring metric used in this thesis, the BN MAP score, adheres to a Bayesian perspective as well. It was developed by Riggelsen (2008) and seeks for the most probable combination of BN structure and parameters for the given data. This is in contrast to the BD metric, which seeks only for the most probable BN structure. It was shown by Riggelsen (2008) that the BN MAP score selects BNs with a better fit to an independent test set than BNs chosen with the BD metric.

Working with fully observed and discrete data sets the BN MAP score can easily be calculated as a function of $n(\cdot)$, where $n(\mathbf{x})$ is defined as the number of occurrences of the state $\mathbf{x}$ in the data. Unfortunately in real-world applications the conditions are not that convenient. Continuous variables and incomplete observations require an adjustment of the data or the scoring function. Different suggestions for such adjustments are given in the course of the dissertation.

The methods developed in this thesis are summarized in the following and it is referred to their applications as well as to related research. The last section concludes with the benefits of applying BNs in natural hazard assessments and stresses the importance of a probabilistic proceeding. Additionally, it gives an outlook on future work perspectives.

6.1.1 Handling continuous variables – data-driven discretization

We face continuous variables in all natural hazard domains considered within this thesis (seismic hazard, flood damage, landslide susceptibility). Handling continuous variables requires assumptions about the families of distributions. This in turn requires expert knowledge, since the implied restrictions on the allowed distributions should be well-founded. Moreover, not all families of distributions grant for an easy analytical handling. To counteract those problems we discretize the variables instead, which allows for a distribution-free learning. There exist several discretization procedures (local vs. global, top-down vs. bottom-up, direct vs. incremental) resulting in different losses of information. Comparative studies, mainly focusing on classification tasks, are conducted e.g. by Dougherty (1995), Liu et al. (2002) and Yang and Webb (2002) with the finding that entropy based discretizations usually seem to perform best in classification tasks. We adopt and enhance the entropy based proceeding suggested by Fayyad and Irani (1993) to discretize attributes of a (tree augmented) Naive Bayes with a continuous target variable. The procedure is applied in Chapter 2 for the modeling of ground motion as well as in the context of flood damage for the (tree augmented) Naive Bayes used for comparison in Chapter 3. For the variable discretization in BNs the *minimal description length* (MDL) principle (Lam and Bacchus, 1994) can be considered as a generalization of the entropy based proceeding. It balances the complexity of discretization and network structure against how well the data are fitted. Enhancements of the MDL principle in connection with BNs are e.g. developed by Friedman and Goldszmidt (1996b) and Wang et al. (2006).

We take again a Bayesian stance instead and treat the discretization, Λ, as random variable as well as the network structure, DAG, and the parameters, $\boldsymbol{\Theta}$. We seek for the combination of those three with the largest posterior probability given the data. In this manner we extend the BN MAP score by adding the discretization to the variables of which we aim to maximize the joint posterior probability. For the definition of the extended BN MAP score we make use of the multivariate discretization procedure developed by Monti and Cooper (1998). The procedure is introduced and shortly explained in Chapter 2. The key idea is to assume that all interactions between the variables can be captured by their discrete representations. Thus, it is sufficient to define a BN on the discrete variables, while each continuous variable depends only on its discrete counterpart, i.e. it is independent of the BN. This allows us to decompose

the likelihood term of the continuous data, $\mathbf{d}^c$, into the likelihood term of the discrete data, $\mathbf{d}$, multiplied by the distribution of the continuous data given the discrete ones,

$$\underbrace{P(\mathbf{d}^c|DAG,\boldsymbol{\Theta},\Lambda)}_{\text{likelihood (continuous)}} = \underbrace{P(\mathbf{d}|DAG,\boldsymbol{\Theta},\Lambda)}_{\text{likelihood (discrete)}} \quad \underbrace{P(\mathbf{d}^c|\mathbf{d},\Lambda)}_{\text{continuous|discrete}} \quad .$$

We apply the above decomposition in Chapter 3 to define an extended version of the BN MAP score valid for continuous data. The continuous variables are in this context defined to be uniformly distributed over the interval which is determined by the discrete value. The discrete likelihood term is already part of the original BN MAP score and can be expressed as a function of $n(\cdot)$.

The definition of the continuous variables' distribution has some week points which are discussed in Chapter 4: First, it results in a discretization that depends on the scaling of the variables. Second, the extended BN MAP score converges towards zero, if the variables are defined on an infinite domain range. To avoid this, an artificial definition of boundaries is required for the variable distributions. Third, if an interval of the discretization is chosen infinitely small around an observation, the score converges towards infinity. Thus, the score favors discretizations with infinitely small intervals and we have to counteract their generation by limiting the allowed minimal size of an interval. In Chapter 4 we present an alternative definition of the extended BN MAP score that avoids the three mentioned problems. In contrast to the definition in Chapter 3 the continuous variables are not assumed to be uniformly distributed over the selected interval, but over all observations within the interval. The score thus depends on the number of observations per interval and not on the interval length. Consequently infinite or very small intervals pose no complications and the score is independent of the scaling of the variables. Additionally, the score does not depend on the exact position of the interval boundaries, instead it is sufficient to know which two observations envelope an interval boundary. The revised discretization is applied in Chapter 4 and 5 to the data sets already considered in Chapter 2 (ground motion) and Chapter 3 (flood damage) as well as to a further data set describing landslide susceptibility.

6.1.2 Avoid coarse resolutions – approximate continuous distributions

An adequate discretization is a trade-off between the number of intervals and the number of observations per interval. To identify the (in-)dependencies between the variables reliably, it is important to maintain a sufficient amount of observations in each interval (and in interval combinations of related variables). To fulfill this demand discretizations that are automatically learned often result in only a few intervals. This is especially the case for variables that depend on many other variables, since they are considered in combination with the states of their parents. The learned discretization may thus not reflect the desired resolution, in particular not if we are interested in the estimation of a certain target variable. In Chapter 2 we suggest therefore to approximate the distribution of the target variable of the learned ground motion model with a *Gaussian kernel density estimator*. The network structure learned before and the discretization for the other variables is kept fixed. The proceeding, including the required parameter adaption, is described in more detail in Chapter 3 and in this context applied to the

flood damage assessment. For both, the ground motion model and the flood damage model, the approximation with the kernel density estimator improves the prediction performance of the model, but at the same time it extensively increases the number of parameters used for the model definition. In more complex models or if we want to approximate more than one variable with the kernel density estimator the exploding number of parameters poses a computational burden. Alternatives that are less parameter intensive, but also less flexible, are approximations with continuous distributions. The families of distributions that allow for exact inference are limited. Conditional Gaussian distributions belong to those families and are well studied representations of continuous variable distributions in BNs (Lauritzen, 1992; Lauritzen and Jensen, 2001). However, besides of their quite confined functional shape, they are limited by the fact that they do not tolerate continuous parents for discrete variables. Koller et al. (1999) instead propose to model the distributions of discrete variables that have continuous parents as mixtures of exponentials, but this entails in the application of Monte Carlo methods for the conduction of inference.

Moral et al. (2001) suggest to apply mixtures of truncated exponentials (MTEs) for the representation of all variable distributions, discrete and continuous ones. MTEs are not only closed under the operations used for BN inference (Langseth et al., 2009b), but also allow for the approximation of a variety of functional shapes with a limited number of parameters (Langseth et al., 2008). We employ the MTE approach in Chapter 4 and 5 for the approximation of the target variable in the flood damage model. It is shown in Chapter 4 that the prediction performance of the flood damage model does not suffer under the replacement of the kernel density estimation with the MTE approach. At the same time the number of model parameters drops significantly. In Chapter 5 we also apply the MTE approach to the ground motion model, modeling the distributions of all variables as MTEs. For the ground motion prediction the MTE approach does not perform as well as the kernel density estimation applied in Chapter 2. However, for the sake of parameter reduction and computational efficiency the MTE approach might be preferable depending on the planned application of the ground motion model.

6.1.3 Handling of missing values

Another big issue in natural hazard assessments is the handling of incomplete data sets. Many real-world data lack observations for different reasons. Within this thesis we consider incomplete data sets in Chapter 3 to 5. Missing data generally pose a problem in BN learning, since the BN scoring metric does not longer have a closed form ($n(\cdot)$ is not defined for incomplete observations). In Chapter 3 we omit the problem while concentrating on the question of discretization simply by filling in missing values randomly drawing from the existing data. This proceeding distorts the original data set and garbles dependency relations between the variables. However, if the percentage of missing values is very small, the change in the data set is not crucial and the proceeding serves as a feasible solution. In Chapter 4 we consider the same data set as in Chapter 3 with a more attentive estimation of the missing values. It is important to note that reliable estimates, no matter which method is applied, are only possible for missing data mechanisms that are ignorable according to the *missing (completely)*

at random (M(C)AR) criteria as defined in (Little and Rubin, 1987). We assume the MAR criteria to hold for all incomplete data sets considered within this thesis.

First attempts of handling incomplete data in BN learning were restricted to the learning of parameters, while the network structure was kept fixed (Lauritzen, 1995). Efforts to learn parameters and structure from incomplete data are widely based on the *expectation maximization* (EM) algorithm (Dempster et al., 1977; Friedman, 1998). The iterative procedure predicts the missing values based on the current model (expectation step) and updates the model based on the completed data set (maximization step). As pointed out by Friedman (1998) the deterministic EM algorithm is prone to get stuck in nearest local optima. Multiple restarts are suggested to solve this problem.

The search space explored by stochastic search algorithms is more comprehensive. Laskey and Myers (2003) and Riggelsen (2005) suggest stochastic proceedings that make use of the Markov Chain Monte Carlo theory for the missing value prediction. Still, the proceedings are computational expensive. In our case we already have to run several iterations of BN learning and discretization, each iteration requiring the estimation of the missing values. Applying an iterative approach for the missing value prediction will thus quickly become infeasible. We adhere to a more efficient albeit approximate method instead and apply the Markov Blanket predictor developed by Riggelsen (2006). The basic idea is to predict the missing values based only on the observations of the corresponding Markov Blanket. A detailed description of the proceeding is given in Chapter 4, where the Markov Blanket predictor is applied in the flood damage assessment. Comparing the BNs learned in Chapter 3 and 4 from the same incomplete data set, we find that the BN structure learned with the Markov Blanket predictor matches better with expert expectations than the one learned with a random replacement of missing values. This is especially the case in regions where variables are observed not often and dependency relations are not recognizable from the data with random replacements. The Markov Blanket predictor apparently preserves dominant (in-)dependencies.

6.1.4 Model uncertainty

Working with real-world data there usually exist many models that explain the data equally well (receive a similar score). The applied algorithm picks only one of these likely models. Neither does the algorithm necessarily pick the most likely model nor does the solution, the chosen model, capture the uncertainty about the model structure. For the ground motion models and the flood damage models learned in this thesis, we assumed that the learned model is the one and only that explains the data best. Interpreting the model structures it should be considered that they do not pose the only explanation.

We consider the problem of model uncertainty in Chapter 5 where we run the BN learning algorithm ten times on the same data set for a landslide susceptibility study. The resulting network structures differ from each other, providing different explanations for the observed data. The different solutions occur due to stochastic elements in the learning procedure and the proneness of the algorithm to get stuck in local maxima when learning from a finite data set. However, despite of the different structures, all learned BNs model the data equally well (receive

a similar score). This result indicates that the quality of the learned BN is not significantly affected by random effects. Still, it stays the problem that there is no unique explanation of the data and a learned BN contains no information about alternative models and the model uncertainty. Different suggestions, how to deal with several competing, but all justifiable BNs are given in literature. Several restarts of the algorithm, for which the initial starting BN can be varied or the training data set is modified by bootstrap sampling (Friedman et al., 1999) give insight into the variety of explaining models. Friedman and Koller (2000) and Riggelsen (2005) alternatively suggest to use a Markov Chain Monte Carlo approach to simulate the posterior distribution of the BN structure. An adoption of these methods for the extended BN MAP score is beyond the scope of this thesis, but an interesting topic for further research.

6.2 Conclusion and future work perspectives

The benefits of applying BNs in natural hazard assessments are particularly stressed in Chapter 5 and picked up again in the following. The graphical representation supports an intuitive understanding of the underlying mechanism and reveals how variables share information and how it 'flows' through the network. This relieves the communication of the learned model to the user. It shows at one glance which variables are considered to be most relevant for the prediction of a certain variable of interest and which variables are assumed to be totally independent of that variable. The intuitive interpretation of the BN allows moreover to define the BN structure based on expert knowledge. This structure can be used directly as model structure or can be defined as a prior that is updated based on data. Alternatively, the BN can be learned purely data-driven as it was done for the applications considered within this thesis. The data based proceeding uncovers the relationships of the variables concealed in the data and may thus reveal unexpected interactions. However, it can not distinguish between artifacts in the data and true dependencies. If present, an exploitation of expert knowledge is hence especially advisable for sparse data sets. It can be included into the analysis via the prior term of the scoring function. How this is realized in detail is left for future work. It might e.g. be thought of a scoring of the edges according to the expert's belief in their justification. Another option is to score the presence/absence of a variable in another variables Markov Blanket according to the expert's belief.

An essential advantage of BNs is the probabilistic proceeding. It allows for the propagation and expression of uncertainties. This is extremely important in natural hazard assessments as demonstrated in the following seismic hazard example. An earthquake of small magnitude will usually cause a small ground motion, but in rare cases it can also cause a moderate or even a strong ground motion. A region with many, but only small earthquakes will be judged to be save based only on the expected values of ground motion, since cases where small magnitudes provoke large ground motions are ignored. In reality the frequent appearance of small magnitude events increases the chance for a large ground motion, but this can be only captured by a probabilistic proceeding. Still, deterministic proceedings are widely used in natural hazard assessments. This is not always due to a missing applicability of probabilistic methods, but has in some cases also political reasons. Deterministic statements, that transmit an impression of confidence, are preferred by a majority of people. Decision makers may also

favor deterministic results as they simplify the choice of action and a careful examination of uncertainty is (too) time consuming. The challenge is thus not only in the identification of uncertainty, but also in its communication. The need for the consideration of uncertainty has to be communicated as well as the uncertainty itself. Expressing uncertainty in a way that helps for decision support instead of raising confusion is a non-trivial task. An intuitively understandable suggestion in the context of tsunami early warning was delivered by Blaser et al. (2011, 2012). Considering different potentially tsunami causing situations the probabilities of four tsunami levels were calculated under the usage of BNs. The different levels were color-coded (no tsunami – green, small tsunami – yellow, moderate tsunami – orange, major tsunami – red) and their probabilities were displayed in a bar with four accordingly colored segments. The length of each segment corresponds to the probability of the associated tsunami level. The graphical representation supports a quick and intuitive understanding of the hazard and the related uncertainty. Still, to be used for decision support, it is essential to have the results quickly at hand. BNs allow to fulfill this demand enabling for fast, close to immediate inference. Moreover, inference can be conducted even though not all variables are observed yet. This allows for a hazard evaluation already at an early stage of the considered event. The estimation can be updated as soon as new information about so far missing variables becomes available.

Capturing the joint distribution of all variables and not only a specific conditional distribution, BNs allow to infer into all directions and to express any conditional distribution of interest. We may thus investigate individual dependency relations and specifically ask for the impact of a certain (set of) variable(s) on another (set of) variable(s). Such investigations clarify the dependency relations between the variables and may thus contribute to a better communication between scientist and non-scientific stakeholders. The effect of specific precautionary measures e.g. can be investigated in detail and considered for different conceivable scenarios. For instance in Chapter 5 the impact of precaution on the building damaged caused by flood events is examined. BNs can also be used for inverse reasoning. E.g. the intensity of an historic event can be estimated based on documented caused damages. BNs may thus help to reconstruct past events.

However, describing the joint distribution of all variables BNs might be less accurate in the prediction of a certain target variable. The scoring function used to learn BNs from data is designed to choose a BN that explains the complete data set. All variables are considered to be equally important. Yet, many natural hazard assessments are especially interested in the prediction of a certain target variable. As mentioned before (tree augmented) Naive Bayes classifiers focus on the prediction of the target variable. They thus often outperform BNs in classification tasks, although the network structure is fixed (or at least very limited in the allowed variations) and usually does not capture the true dependencies. This raises the questions if we can achieve a better prediction performance by considering the complete DAG space and choosing a BN that is evaluated according to its ability to predict the target variable. In this context we may think of an alternative scoring function for BNs. This function can e.g. give more weight to the target variable or score the BN's prediction performance. If we learn the BN from complete data, we can restrict the search space of all DAGs to the Markov Blanket of the target variable. How the search algorithm and the scoring metric look in detail is left for future work.

A further research perspective is the direct learning of BNs with MTEs representations for continuous variables. So far we discretized the variables in the learning procedure and searched for the best combination of discretization, network structure and parameters. In some cases we approximated the continuous distributions with MTEs afterwards to achieve a finer resolution. However, it is not evident that the selected network structure is the best choice in combination with MTEs. It would be more accurate to already consider the combination of network structure and MTEs during the learning phase as it is suggested by Romero et al. (2006). There, MTEs representations are learned for each DAG visited during the DAG traversal and the resulting BNs are scored using a penalized likelihood score. However, this proceeding is extremely demanding in computational costs. For this reason the number of exponential terms as well as the number of splits for the partitioning of the domain is kept constant. It stays to show if such or a similar proceeding is applicable to and beneficial for the applications considered within this thesis.

Another option to deal with continuous distributions in BNs is the application of copulas (Nelson, 2006). Copulas offer a general framework for constructing multivariate distributions by linking univariate marginals. Yet, constructing high-dimensional copulas is difficult and usually achieved by combining bivariate copulas. In practice those applications are generally limited to a small number of variables. Elidan (2010, 2013) suggests therefore the combination of copulas with the BN approach. The decomposition of the joint distribution into local terms provided by a BN reduces the dimensionality of the domain and thus enables the application of copulas in high-dimensional problems. The application of copulas in the context with BNs is beyond the scope of this thesis, but on the agenda for further research.

Despite of the above mentioned open issues the methods developed and applied in this thesis already pose an important contribution to the establishment of BNs in natural hazard assessments. Solutions for two dominant challenges for BN learning from real-world data are presented, namely the handling of continuous variables with unknown family of distributions and the handling of incomplete data. Since the algorithms that tackle the individual problems are highly interacting, computational efficiency is of special importance when the underlying training data comprise both, continuous variables and incomplete observations. The proceedings suggested in this thesis take care about computational feasibility. They thus enable the treatment of large data sets with continuous variables and missing data and make the benefits of BNs accessible for the considered natural hazard assessments. Additionally they are purely data driven and can thus easily be transferred to further natural hazard domains or other domains of the reader's interest.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716–723.

Blaser, L., Ohrnberger, M., Riggelsen, C., and Scherbaum, F. (2009). Bayesian Belief Network for Tsunami Warning Decision Support. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 757–768.

Blaser, L., Ohrnberger, M., Riggelsen, C., Babeyko, A., and Scherbaum, F. (2011). Bayesian networks for tsunami early warning. *Geophysical Journal International*, 185(3), 1431–1443.

Blaser, L., Ohrnberger, M., Krüger, F., and Scherbaum, F. (2012). Probabilistic tsunami threat assessment of 10 recent earthquakes offshore Sumatra. *Geophysical Journal International*, 188(3), 1273–1284.

Boore, D. (2003). Simulation of ground motion using the stochastic method. *Pure and Applied Geophysics*, 160(3-4), 635–676.

Bouckaert, R. (1995). *Bayesian belief networks: from construction to inference*. Universiteit Utrecht, Faculteit Wiskunde en Informatica.

Castelo, R. and Kocka, T. (2003). On inclusion-driven learning of Bayesian networks. *The Journal of Machine Learning Research*, 4, 527–574.

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

Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4), 309–347.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Twelfth International Conference on Machine Learning*, 194–202.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer.

Elidan, G. (2010). Copula bayesian networks. *Advances in neural information processing systems*, 24.

Elidan, G. (2013). Copulas in Machine Learning. *Copulae in Mathematical and Quantitative Finance*. Springer Berlin Heidelberg, 39–60.

Elmer, F., Thieken, A. H., Pech, I., and Kreibich, H. (2010). Influence of flood frequency on residential building losses. *Natural Hazards and Earth System Sciences*, 10(10), 2145–2159.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In *Fourteenth International Conference on Machine Learning*, 125–133.

Friedman, N. (1998). The Bayesian structural EM Algorithm. In *Fourteenth conference on Uncertainty in artificial intelligence*, 129–138.

Friedman, N. and Goldszmidt, M. (1996a). Building Classifiers using Bayesian Networks. In *National conference on artificial intelligence*, 1277–1284.

Friedman, N. and Goldszmidt, M. (1996b). Discretizing Continuous Attributes While Learning Bayesian Networks. In *Thirteenth International Conference on Machine Learning*, 157–165.

Friedman, N., Goldszmidt, M., and Wyner, A. (1999). Data analysis with Bayesian networks: A bootstrap approach. In *Fifteenth conference on Uncertainty in artificial intelligence*, 196–205.

Friedman, N. and Koller, D. (2000). Being Bayesian about network structure. In *Sixteenth conference on Uncertainty in artificial intelligence*, 201–210.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20(3), 197–243.

Jensen, F. and Nielsen, T. (2001). *Bayesian networks and decision graphs*. Springer, New York, NY, USA.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. The MIT Press.

Koller, D., Lerner, U., and Angelov, D. (1999). A General Algorithm for Approximate Inference and Its Application to Hybrid Bayes Nets. *Fifteenth conference on Uncertainty in artificial intelligence*, 324–333.

Korup, O. and Stolle, A. (2013). Landslide Prediction from Machine Learning. *submitted to Geology Today*.

Kühn, N., Riggelsen, C., and Scherbaum, F. (2009). Facilitating Probabilistic Seismic Hazard Analysis Using Bayesian Networks. In *Seventh Annual Workshop on Bayes Applications (in conjunction with UAI/COLT/ICML 2009)*.

Kühn, N. M., Riggelsen, C., and Scherbaum, F. (2011). Modeling the Joint Probability of Earthquake, Site, and Ground-Motion Parameters Using Bayesian Networks. *Bulletin of the Seismological Society of America*, 101(1), 235–249.

Lam, W. and Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, 10(3), 269–293.

Langseth, H. and Nielsen, T. D., Rumí, R., and Salmerón, A. (2008). Parameter estimation in mixtures of truncated exponentials. In *4th European Workshop on Probabilistic Graphical Models*, 169–176.

Langseth, H., Nielsen, T. D., Rumí, R., and Salmerón, A. (2009a). Maximum Likelihood Learning of Conditional MTE Distributions. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 240–251.

Langseth, H., Nielsen, T. D., Rumí, R., and Salmerón, A. (2009b). Inference in hybrid Bayesian networks. *Reliability Engineering & System Safety*, 94(10), 1499–1509.

Langseth, H., Nielsen, T. D., Rumí, R., and Salmerón, A. (2010). Parameter estimation and model selection for mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 51(5), 485–498.

Laskey, K. B. and Myers, J. W. (2003). Population Markov Chain Monte Carlo. *Machine Learning*, 50(1-2), 175–196.

Lauritzen, S. L. (1992). Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models. *Journal of the American Statistical Association*, 87(420), 1098–1108.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2), 191–201.

Lauritzen, S. and Jensen, F. (2001). Stable local computation with conditional Gaussian distributions. *Statistics and Computing*, 11(2), 191–203.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*, volume 4. Wiley, New York.

Liu, H., Hussain, F., Tan, C. L. I. M., and Dash, M. (2002). Discretization : An Enabling Technique. *Data mining and knowledge discovery*, 6(4), 393–423.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A. H. (2010). Assessment of economic flood damage. *Natural Hazards and Earth System Science*, 10, 1697–1724.

Monti, S. and Cooper, G. F. (1998). A multivariate discretization method for learning Bayesian networks from mixed data. In *Fourteenth conference on Uncertainty in artificial intelligence*, 404–413.

Moral, S., Rumí, R., and Salmerón, A. (2001). Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 156–167.

Nelson, R. B. (2006). *An introduction to copulas*. Springer Berlin.

Pernkopf, F. and Bilmes, J. (2005). Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In *22nd international conference on Machine Learning*, 657–664.

Riggelsen, C. (2005). MCMC learning of Bayesian network models by Markov blanket decomposition. In *Machine Learning: ECML 2005*, 329–340.

Riggelsen, C. (2006). Learning Bayesian Networks from Incomplete Data: An Efficient Method for Generating Approximate Predictive Distributions. In *SIAM International conf. on data mining*, 130–140.

Riggelsen, C. (2008). Learning Bayesian Networks: A MAP Criterion for Joint Selection of Model Structure and Parameter. In *ICDM, 2008. Eighth IEEE International Conference on Data Mining*, 522–529.

Riggelsen, C., Ohrnberger, M., and Scherbaum, F. (2007). Dynamic bayesian networks for real-time classification of seismic signals. In *PKDD 2007. 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, 565–572.

Romero, V., Rumí, R., and Salmerón, A. (2006). Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 42(1), 54–68.

Rumí, R., Salmerón, A., and Moral, S. (2006). Estimating mixtures of truncated exponentials in hybrid bayesian networks. *Test*, 15(2), 397–421.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464.

Silverman, B. (1986). *Density estimation for statistics and data analysis*, Vol. 26. Chapman & Hall/CRC.

Song, Y., Gong, J., Gao, S., Wang, D., Cui, T., Li, Y., and Wei, B. (2012). Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China. *Computers & Geosciences*, 42, 189–199.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398), 528–540.

Thieken, A. H., Müller, M., Kreibich, H., and Merz, B. (2005). Flood damage and influencing factors: New insights from the August 2002 flood in Germany. *Water resources research*, 41(12), W12430.

Vogel, K., Riggelsen, C., Kühn, N., and Scherbaum, F. (2012a). Graphical Models as Surrogates for Complex Ground Motion Models. In*11th International Conference on Artificial Intelligence and Soft Computing*, 188–195.

Vogel, K., Riggelsen, C., Merz, B., Kreibich, H., and Scherbaum, F. (2012b). Flood Damage and Influencing Factors: A Bayesian Network Perspective. In 6th European Workshop on Probabilistic Graphical Models, 347–354.

Vogel, K., Riggelsen, C., Scherbaum, F., Schröter, K., Kreibich, H., and Merz, B. (2013). Challenges for Bayesian Network Learning in a Flood Damage Assessment Application. In *11th International Conference on Structural Safety & Reliability*.

Wang, S.-C., Li, X.-L., and Tang, H.-Y. (2006). Learning Bayesian Networks Structure with Continuous Variables. In *Advanced Data Mining and Applications*, 448–456.

Yang, Y. and Webb, G. I. (2002). A Comparative Study of Discretization Methods for Naive-Bayes Classifiers. In *PKAW 2002. Pacific Rim Knowledge Acquisition Workshop*, 159–173.

ACKNOWLEDGEMENTS

During the preparation of this thesis I benefited from the valuable support of many people around me.

First of all, I would like to thank my supervisors Frank Scherbaum and Carsten Riggelsen for giving me the opportunity to do research in this interesting field where geophysics, math and computer sciences come together. I am thankful to Frank and Matthias Ohrnberger, who are responsible for my growing interest in geophysics during the end of my diploma studies and who informed me about this PhD topic. Special gratitude is dedicated to Carsten, who conveyed the topic of machine learning in particular graphical models to me. I benefited extensively from Carsten's willingness to share his competence in this field. Carsten provided me constantly with plenty of ideas and suggestions how to tackle specific problems and what else would be interesting to consider. He was always open for scientific (and private) discussions and encouraged me to present my results in publications. In this context he also helped me a lot with language issues. I keenly appreciate Carsten's perfect supervision.

Further I would like to thank Oliver Korup, Heidi Kreibich, Nicolas Kühn, Bruno Merz and Kai Schröter for their cooperation. Thanks to them this work comprises a variety of natural hazard domains. I am grateful for the data and the expertise they provided to me. Without their help the thesis would have missed the considered applications and stayed in a theoretical framework.

I would also like to thank all the people who shared an office with me as well as the whole geophysics group and the inhabitants of the Mühlenberg for providing an enjoyable working environment with a lot of cakes, barbecues, movies, caipirinhas, karaoke events, ...

I thank Conny, Sebastian, Dominik, Julia and Max for reading parts of the thesis and their feedback.

Thanks for the financial support to PROGRESS.

Finally, I am much obliged to my dear friends and in especially to my flatmates and Max for the entertainment outside of Golm and for backing me up. I am especially grateful to my family, in particular to my parents, for their constant support, even though this is strange stuff I am doing and – even more – it is in english. Thanks for being around.