12-9-2019 10:30 AM

Application of Bayesian Networks to Integrity Management of Energy Pipelines

Wei Xiang
The University of Western Ontario

Supervisor
Zhou, Wenxing
The University of Western Ontario

Graduate Program in Civil and Environmental Engineering
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Metal-loss corrosion and third-party damage (TPD) are the leading threats to the integrity of buried oil and natural gas pipelines. The pipeline industry is devoting significant efforts to manage the integrity of pipelines with respect to these threats. The reliability-based integrity management program is being increasingly adopted by pipeline operators to deal with uncertainties associated with the corrosion and occurrence of TPD events. This thesis employs Bayesian networks (BNs) and non-parametric Bayesian networks (NPBNs) to deal with four issues with regard to the reliability-based management program of corrosion and TPD.

The pipeline operators periodically perform in-line inspections (ILIs) to detect and size the corrosion defects on the pipelines. The first study integrates the quantification of measurement errors of the ILI tools, corrosion growth modeling and reliability analysis in a single dynamic Bayesian network (DBN) model, and employs the Expectation-Maximization (EM) algorithm in the context of the parameter learning to learn the parameters of the DBN model from the ILI-reported and filed-measured corrosion depths. In comparison with existing growth models, the integrating and graphical features of the developed model make the process of corrosion management more intuitive and transparent to users. The employment of parameter learning provides an objective and convenient approach to elicit the probabilistic information from ILI and field measurement data.

The second study develops the BN model to estimate the probability of a given pipeline being hit by third-party excavations by taking into account common preventative and protective measures. The EM algorithm in the context of parameter learning is employed to learn the parameters of the BN model from datasets that consist of individual cases of third-party activities but with missing information. The developed BN model is advantageous over the existing fault tree models in that it can handle the estimation of the probability of hit under different scenarios of available information. Moreover, the BN model and EM-based parameter learning proposed in this study allow pipeline operators to estimate the probability of hit by efficiently taking into account historical third-party excavation records in an objective manner.

The ILIs are infeasible for a portion of buried pipelines due to the reasons such as small pipe diameters, tight bends, or a lack of launching and receiving stations for ILI tools, which are known as unpiggable pipelines. To assist with the corrosion assessment for the unpiggable pipelines, the third study develops a non-parametric Bayesian network (NPBN) model to predict the corrosion depth on buried pipelines using the pipeline age and local soil properties as the predictors. The dependence structure and parameters of the NPBN model are extracted from Velázquez's dataset, which consists of 250 samples of corrosion depths, pipeline age, and such local soil properties as the water content, redox potential, and pH value.

The epistemic uncertainties in the basic random variables of reliability analysis of corroded pipelines introduce uncertainty into the calculated failure probability P_f, which may affect the decision making. The last study develops a sample size determination (SSD) method for collecting samples to reduce the epistemic uncertainties in the probabilistic distributions of basic random variables. This work first discretizes the continuous random variables and assign Dirichlet prior distributions to the probability mass functions (PMFs) to characterize the epistemic uncertainties. The total probability theorem is employed to express P_f in terms of PMFs of the discretized variables and conditional failure probabilities corresponding to given values of discretized variables. Then, the prior, posterior and pre-posterior analyses of P_f are carried out. The optimal sample size criterion to maximize the expected net gain of sampling (ENGS) is developed based on the result of the pre-posterior analysis of P_f and quadratic loss function. The developed method is applied to determining the sample size of the model error of a burst capacity model and determining the number of pipe joints to excavate for the corrosion assessment of unpiggable pipelines.

Keywords

underground pipeline, corrosion, third-party damage, Bayesian network, non-parametric Bayesian network, optimal sample size determination, value of information

Summary for Lay Audience

The buried pipelines are the most widely used mode to transport oil and natural gas. The metal-loss corrosion and damage from excavation activities can lead to pipeline incidents. To manage the pipeline safety, pipeline companies need to estimate the probabilities of occurrence of such incidents. This thesis uses graphical models known as Bayesian networks to enhance the current practice of corrosion and excavation damage management. A Bayesian network (BN) consists of circles to represent events and arrows to represent the relationship between the events. Once a part of the model is observed, the probabilities of the rest of the events can be calculated.

Pipeline companies routinely run inspection tools through the pipelines to detect and size corrosion defects. The thesis develops a BN model to forecast the growth of the corrosion depth and probability of failure at the specific corrosion defect using the corrosion depths reported by the inspection tools. However, the inspection tools are infeasible for a portion of pipelines due to the reasons such as small diameters and tight bends. To assist with the corrosion assessment of such pipelines, the thesis develops a BN model to predict the corrosion depth using the pipeline age and soil parameters.

To prevent the pipeline from excavation damage, the pipeline industry and regulatory agencies employ a series of measures such as patrols along the pipeline, warning signs on the pipelines and burial depth. The failures of all the preventative and protective measures can lead to the pipeline being hit by the excavation machine. The present thesis develops a BN model to estimate the probability of a given pipeline being hit by an excavation event. The probabilities of the preventative and protective measures are automatedly learned from the historical data collected by the pipeline industry.

To reduce the uncertainties in the corrosion management program, the pipeline industry often collects samples by performing experiments or field measurements, which are generally expensive. The fourth study in the thesis develops a method to determine the optimal sample size from the economic standpoint and apply it to two sample size determination problems in the context of corrosion management of pipelines.

Co-Authorship Statement

A version of Chapter 2, co-authored by Wei Xiang and Wenxing Zhou has been accepted by *Structure and Infrastructure* (accepted on August 11, 2019).

A version of Chapter 3, co-authored by Wei Xiang and Wenxing Zhou is under review for possible publication in *Reliability Engineering and System Safety*.

A version of Chapter 4, co-authored by Wei Xiang and Wenxing Zhou is under review for possible publication in *Corrosion (NACE)*.

A version of Chapter 5, co-authored by Wei Xiang and Wenxing Zhou is under review for possible publication in *Civil Engineering and Environmental System* (revision requested).

Dedication

To my parents and grandparents

Acknowledgments

First and foremost, I would like to express my sincere and deep gratitude to my supervisor Dr. Wenxing Zhou. His valuable guidance, profound insights into the field of integrity management of pipelines, and countless hours of revision of my technical writing are greatly appreciated. Without the illuminating instructions and consistent patience of my supervisor, it would be impossible for me to complete this thesis. It is a true privilege to study under his supervision.

Special thanks go to Dr. Han-Ping Hong. Working on research projects with him and taking his courses made me benefit a lot from his deep knowledge of structural reliability. I would also like to appreciate Dr. George Knopf, Dr. Gregory Kopp, Dr. Yong Li, and Dr. M.Reza Najafi for the time and efforts they put in reviewing my thesis, raising critical questions and providing constructive comments.

I would like to extend my thanks to all the fellow students in our research group and my dear friends at Western University. I have been fortunate enough to have them around during the past four years. Their friendship encouraged me to get through difficult times. The financial support provided by the Government of Ontario through the Ontario Trillium Scholarship (OTS), TC Energy (formerly TransCanada Ltd), Natural Science and Engineering Research Council (NSERC) of Canada, Western University and Dr. Wenxing Zhou is greatly appreciated.

My deepest gratitude always goes to my parents and grandparents for their endless love and unconditional support. To help me realize my dream, they have sacrificed too much. I am indebted to them forever.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations and Symbols

Abbreviations

AA	authorized activities
BN	Bayesian network
CDF	cumulative distribution function
CEPA	Canadian Energy Pipeline Association
CGA	common ground alliance
COV	coefficient of variation
CP	cathodic protection
CPT	conditional probability table
CSA	Canadian Standard Association
DAG	directed acyclic graph
DBN	dynamic Bayesian network
EGIG	European Gas Pipeline Incident Data Group
EM	Expectation-Maximization
ENGS	expected net gain of sampling
EVPI	expected value of perfect information
EVSI	expected value of sampling information
ILI	in-line inspection
KL	Kullback-Leibler
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MFL	magnetic flux leakage
NBS	National Bureau Standards
NPBN	non-parametric Bayesian network
PDF	probability density function
PHMSA	Pipeline and Hazardous Materials Safety Administration
PMF	probability mass function
ROW	right-of-way
SMYS	specified minimum yield strength
SSD	sample size determination

TPD	third-party damage
UA	unauthorized activities
VoI	value of information

Symbols

Chapter 2

D	outside diameter of the pipeline
M	Folias bulging factor
S	defect sate, e.g. survival or failure
X	actual defect depth
Y	ILI-reported defect depth
ΔX	annual growth of defect depth
l	defect length
o_p	operating pressure of the pipeline
pa	parent configuration of a node
r_b	burst pressure capacity of the pipeline at a given defect
r_c	total number of parent configurations of node C
w_t	actual pipe wall thickness
w_{tn}	nominal pipe wall thickness
α	parameters of the Dirichlet distribution
β	multiplicative bias associated with the ILI tool
γ	additive bias associated with the ILI tool
γ’	additive measurement error associated with the ILI tool
ε	scattering error associated with the ILI tool
θ	parameters of the BN model
κ	model error of the B31G Modified model
σ	standard deviation of ε
σ_y	yield strength of the pipe steel

Chapter 3

$A_{\#}$	nodes/random variables representing pipeline attributes
$B_{\#}$	basic events in the fault tree
$E_{\#}$	intermediate events in the fault tree
D_p	KL-divergence before parameter learning
D_π	KL-divergence after parameter learning
P_h	probability of hit

T_0	top event in the fault tree
pa	parent configuration of a node
b	states of $B_{\#}$
α	parameters of the Dirichlet distribution
θ	parameters of the nodes representing basic events

Chapter 4

$C(\bullet)$	copula
$F_{X_i}(x_i)$	marginal CDF of random variable X_i
U_i	inverse normal transformation of X_i
bc	bicarbonate
bd	bulk density
cc	dissolved chloride
d	corrosion depth
det(•)	determinant of the matrix
d_f	field-measured corrosion depth
pH	pH value
pp	pipe-to-soil potential
r	rank correlation between μ_d and d_f
re	soil resistivity
r_{ij}	rank correlation between U_i and U_j
rp	redox potential
sc	sulfate ion concentrations
wc	water content
$\mathbf{x}_e$	observations of the predictors
Σ_E	empirical rank correlation
Σ_N	empirical normal rank correlation
Σ_M	rank correlation matrix associated with the NPBN
$\Phi_n(\bullet)$	*n*-variate standard normal distribution function
$\Phi^{-1}(\bullet)$	inverse of the standard univariate normal distribution function
ρ	linear correlation between μ_d and d_f
ρ_{ij}	linear correlation coefficient between U_i and U_j

μ_d	predicted mean value of corrosion depth

Chapter 5

C	parameter of the quadratic loss function
C_F.	failure cost
C_s	unit sampling cost
C_κ	the cost of the full-scale burst test of a corroded pipe specimen
D	actual diameter of the pipeline
D_n	nominal diameter of the pipeline
Dir	Dirichlet distribution
$\mathrm{E}_N[\bullet]$	expectation with respect to $\mathbf{N}_Y$
$\mathrm{E}_{P_f}[L]$	expectation with respect to P_f
$L(\bullet)$	quadratic loss function
$L(\mathbf{w}_Y\|\mathbf{n}_Y)$	likelihood of $\mathbf{n}_Y$
M	Folias bulging factor
$\mathbf{N}_Y$	random vector corresponding to $\mathbf{n}_Y$
$N_{Y,j}$	random variable corresponding to $n_{Y,j}$
P_f,	failure probability
$\mathbf{W}_Y$	a vector representing the PMF of Y
$W_{Y,j}$	j-th element of $\mathbf{W}_Y$
Y	a discrete random variable
d	ILI-reported defect depth
l	ILI-reported defect length
$\mathbf{n}_Y$	a vector representing the number of samples for Y
n_{Y0}	total number of samples
$\mathbf{n}_{\mathrm{Y0-opt}}$	optimal sample size
$n_{Y,j}$	number of samples lying in j-th state of Y
p_e	point estimate of failure probability
o_{pn}	nominal operating pressure of the pipeline
o_p	actual operating pressure of the pipeline
w_{tn}	nominal wall thickness of the pipeline
w_t	actual wall thickness of the pipeline

y_j	j-th state of random variable Y
$\boldsymbol{\alpha}_Y$	parameter vector of the Dirichlet distribution
$\alpha_{Y,j}$	j-th element of $\boldsymbol{\alpha}_Y$
α_{Y0}	equivalent sample size of the Dirichlet distribution
σ_y	yield strength of the pipe steel
κ	model error of the B31G Modified model
$\mu_{N_{Y,j}}$	mean value of $N_{Y,j}$
$\mu_{P_f}^{\pi}$	prior mean value of P_f
$\mu_{P_f}^{p}$	posterior mean value of P_f
$\mu_{W_{Y,j}}^{\pi}$	prior mean value of $W_{Y,j}$
$\mu_{W_{Y,j}}^{p}$	posterior mean value of $W_{Y,j}$
$\xi_{N_{Y,j}}$	variance of $N_{Y,j}$
$\xi_{W_{Y,j}}^{\pi}$	prior variance of $W_{Y,j}$
$\xi_{W_{Y,j}}^{p}$	posterior variance of $W_{Y,j}$
$\xi_{P_f}^{\pi}$	prior variance of P_f
$\xi_{P_f}^{p}$	posterior variance of P_f
$\omega_{W_{Y,jk}}^{\pi}$	prior covariance between $W_{Y,j}$ and $W_{Y,k}$
$\omega_{W_{Y,jk}}^{p}$	posterior covariance between $W_{Y,j}$ and $W_{Y,k}$
$\omega_{N_{Y,jk}}$	covariance of $N_{Y,j}$ and $N_{Y,k}$ ($j, k = 1, 2, \ldots, m; j \neq k$)

1 Introduction

1.1 Background

1.1.1 Pipeline integrity management and issues to address

Pipeline systems have been widely recognized as the most efficient and safest mode to transport hydrocarbons (i.e. crude oil and natural gas) over long distances (Green and Jackson, 2015). The structural integrity of pipelines is subject to various threats that include external corrosion, internal corrosion, third-party damage (TPD), cracking, material failures, among others (Cosham et al., 2007). The data collected by the Pipeline and Hazardous Material Safety Administration (PHMSA) of the US Department of Transportation report 464 pipe-related incidents on onshore gas transmission pipelines between 2002 and 2013, of which the distribution by failure causes is depicted in Fig. 1.1 (Lam and Zhou, 2016). This figure indicates that external corrosion and third-party damage are the first two leading threats, which therefore are the focuses of this thesis. Since pipe-related incidents are generally associated with severe consequences in terms of human safety, property damage and environmental impact, the pipeline industry and regulatory agencies are devoting significant efforts to improve the safety of pipelines. The reliability-based pipeline integrity management program is increasingly adopted by the pipeline operators to deal with the uncertainties involved in the corrosion and occurrence of TPD activities (Adianto et al., 2018; Kariyawasam and Peterson, 2008; Tomic et al., 2018; Zhou, 2010).

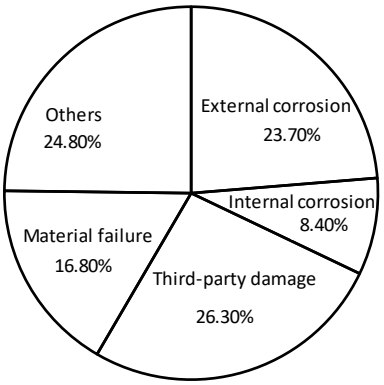


Figure 1.1 Distribution of pipe-related incidents between 2002 and 2013 by failure causes based on PHMSA data

While external coating and cathodic protection (CP) are widely employed to protect the pipeline from corrosion, corrosion may take place as a result of the breakdown of the protection systems. Inline inspection (ILI) tools are routinely used to detect and size corrosion defects. A typical magnetic flux leakage (MFL) based ILI tool, also known as "smart pig", is shown in Fig. 1.2. As the ILI tool travels through the pipeline, a magnetic flux field is imposed on the pipeline wall. The metal-loss corrosion can cause the distortion of the magnetic flux field as shown by Fig. 1.3, which can then be correlated with the size of a defect. The profile of a typical corrosion defect on the external surface of the pipeline characterized by maximum depth, length and width is given by Fig. 1.4. Note that the ILI tools can differentiate between the corrosion defects on the external and internal surfaces of pipelines.

Figure 1.2 A typical high-resolution MFL tool

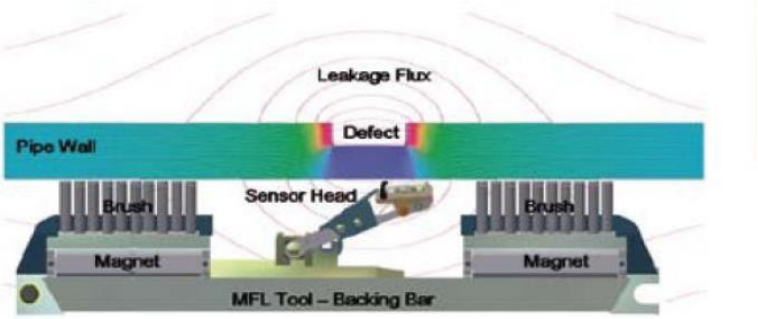


Figure 1.3 ILI detection of magnetic flux leakage from an external corrosion defect

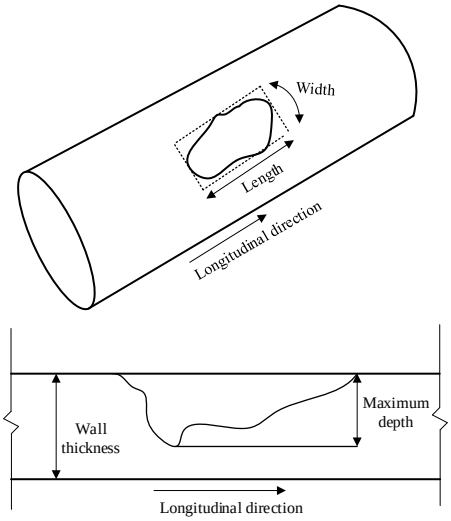


Figure 1.4 Dimensions of a typical corrosion defect on the external surface of pipeline

The reliability-based corrosion management program generally includes the periodical ILIs to detect and size corrosion defects on a given pipeline, engineering critical assessment of reported defects and mitigation of critical defects (Zhang, 2014). The accurate modeling of corrosion growth is of great importance to the time-dependent reliability evaluation and scheduling mitigation activities. Extensive studies have been reported in the literature to model the corrosion growth probabilistically to account for the inherent random nature of the corrosion growth (Ahammed, 1998; Caleyo et al., 2009a; Hong, 1999; Zhang, 2014; Zhou et al., 2017). The growth models developed in the hierarchical Bayesian framework are advantageous in that the ILI data can be incorporated to update the model parameters (Maes et al., 2010; Pandey et al., 2009; Zhang et al., 2013; Zhang et al., 2014). The errors

on the ILI data (i.e. biases and random scattering errors) are typically evaluated from regression or Bayesian analyses and then incorporated in the growth model as an input (Al-Amin et al., 2012). The growth models updated by ILI data can then be incorporated in the reliability-based defect assessment, e.g. evaluation of the time-dependent probabilities of failure of individual corrosion defects and/or system failure probability of a given pipe segment containing multiple active defects (Al-Amin and Zhou, 2014; Pandey et al., 2009). The three components, i.e. quantification of errors on ILI data, growth modeling of corrosion defects and time-dependent reliability analysis provide a reasonable framework to account for various uncertainties in the reliability-based corrosion management program. However, the implementation of these components in practice has the following difficulties: 1) the complexity of the hierarchical Bayesian models and Markov Chain Monte Carlo (MCMC) technique renders them difficult to use by non-specialists; 2) The quantification of errors on ILI data, Bayesian growth model updating and failure probability evaluation should be carried out in separated steps or models. It is therefore desirable from a practical perspective to combine the three components into a single integrated analysis and develop a tool more amenable to the corrosion management practice in the pipeline industry.

In practice, there are pipelines for which ILIs are infeasible or extremely difficult to conduct due to various reasons such as the tight bends, over- or under-size valves, complicated connections and a lack of launching and receiving stations for ILI tools (Rau and Kirkwood, 2016; Beauregard et al., 2018). Such pipelines are commonly known as unpiggable pipelines. The lack of inspection data presents significant challenge to the corrosion assessment of unpiggable pipelines. Since the corrosion deterioration on buried pipelines is greatly influenced by the corrosive properties of surrounding soils, characterizing the correlation of corrosion sizes with local soil parameters has received a great deal of attention in the research community (Velázquez et al., 2009; Caleyo et al, 2009b; Ricker, 2010; Melchers and Petersen, 2018). Velázquez et al. (2010) reported a corrosion dataset with 259 samples, of which each individual sample consists of the corrosion depth, pipeline age, and local soil parameters. Such dataset can be used to develop a model for predicting the corrosion depth using soil parameters as predictors. The

developed predictive model will be of great practical value for the corrosion assessment of unpiggable pipelines (Beauregard et al., 2018).

A third party is neither a pipeline operator nor a contractor hired by the operator to service the pipeline; in other words, a third party is an individual or organization unrelated to pipeline assets. Commonly used preventative and protective measures include, for example, the one-call system (third parties notify the pipeline operators through one-call centers before excavation), warning signs along the pipeline right-of-way (ROW), regular patrol of ROW, burial depth of pipelines and physical protection such as concrete slabs buried above the pipeline alignment. In the reliability-based pipeline integrity management program with respect to TPD, the fault tree model is widely employed to estimate the probability of hit (Chen and Nessim, 1999). In the fault tree model, the failures of the preventative and protective measures are known as basic events, and a pipeline being hit by a third-party activity is modeled as the result of occurrences of these basic events. In the practice of TPD management over the past few decades, pipeline operators have collected a substantial amount of TPD related data such as the individual TPD activities including the information of pipeline attributes, prevention measures and consequences of the TPD activities, and it is highly desirable to use the collected data to estimate the probabilities of basic events.

In the reliability analysis of corroded pipelines, the epistemic uncertainties in the probabilistic distributions of basic random variables introduce uncertainty into the calculated failure probability, which may affect the decision-making (Der Kiureghian, 1989). The epistemic uncertainties can be reduced by collecting samples of the basic random variables and using these samples to update the corresponding probability distributions. Since the sampling cost is in general high, the sample size should be determined by balancing between the cost and associated benefit. This is commonly known as the sample size determination (SSD). The existing methods can only address SSD problem for specific distributions (Nishijima and Faber, 2007; Higo and Pandey, 2016). It is therefore desirable to develop a general framework that can deal with SSD for a wide range of probability distributions by considering the impact of epistemic uncertainties in the distributions of basic random variables on the failure probability.

1.1.2 Research tools – Bayesian networks and non-parametric Bayesian networks

A Bayesian network (BN) is a graphical acyclic diagram (DAG) representing the joint distribution of a set of random variables. A BN consists of nodes symbolizing the random variables and arcs symbolizing causal relationships between the nodes. Given the observation on a subset of the nodes in a BN, the joint probability distribution of the rest of the nodes in the BN can be updated through Bayes' theorem. This is the so-called inference, the most important application of BNs. Various exact and approximate inference algorithms are described in many textbooks (e.g. Nielsen and Jensen, 2009; Pearl, 2014). BNs are generally applicable to discrete random variables (Langseth et al., 2009). The marginal and conditional distributions of discrete random variables are defined through the probability mass functions and conditional probability tables (CPT), respectively. The entries in the CPTs are called the parameters of the BN, which can either be specified by experts or extracted from data through the parameter learning (Heckerman, 1998). Due to the intuitive graphical nature and ability to efficiently handle the Bayesian updating of a large set of random variables, BNs have become increasingly popular in the engineering reliability and risk analysis during the last two decades, including using dynamic Bayesian networks (DBNs) to model the deterioration of structures (Luque and Straub, 2016; Straub, 2009), and utilizing BNs to evaluate and update the reliability of structures (Mahadevan et al., 2001; Straub and Der Kiureghian, 2010a; Straub and Der Kiureghian, 2010b).

Continuous random variables are generally discretized to be included in a BN. If a significant number of continuous random variables are however included in a BN, each of them discretized by a sufficiently large number of states to ensure the computational accuracy, the efforts for specifying the CPTs can become prohibitively burdensome. Moreover, carrying out inference in BNs with significantly large CPTs can be computationally prohibitive. The Non-parametric Bayesian network (NPBN) is developed to overcome the above-described drawbacks of the BN in dealing with continuous random variables (Kurowicka and Cooke, 2005). An NPBN is a DAG with nodes and arcs symbolizing a set of continuous random variables and dependence between them,

respectively. The dependence between any two nodes is quantified by the (conditional) Spearman's rank correlation, which is the correlation coefficient between ranks, i.e. cumulative distribution functions (CDFs), of the two variables. An NPBN characterizes the joint distribution of the continuous random variables involved by a copula. While any copula function can be used in NPBN, the Gaussian copula is of particular importance to NPBN mainly because it allows analytical inferences. The employment of Gaussian copula NPBNs has become increasingly popular for the high dimensional dependence modeling and risk analysis (Zilko et al., 2016; Morales-Napoles and Steenbergen 2014; Hanea et al., 2015; Morale-Napoles et al., 2014; Hanea et al., 2013; Lee and Pan, 2018; Wang et al., 2019).

A number of software tools are available to deal with BN modeling and inference (Mahjoub and Kalti, 2011), among which the commercial software tools Netica® and UNINET® are employed to deal with discrete Bayesian networks and non-parametric Bayesian networks, respectively. The modeling and parameter learning for BN models consisting of discrete/discretized random variables are implemented in the user-interface of Netica®. The NPBN model mining and updating based on a multivariate dataset are implemented in the software UNINET®.

1.2 Objective and research significance

The objectives of this thesis include: 1) integrate the quantification of measurement errors of ILI tools, corrosion growth modeling and reliability analysis in a single DBN model for the reliability-based corrosion management of oil and gas pipelines, and employ the Expectation-Maximization (EM) algorithm to learn the model parameters from ILI-reported and field-measured corrosion depths; 2) develop a BN model for evaluating the probability of hit given a third-party activity and employ the EM algorithm to learn the model parameters from historical data of third-party activities collected by the pipeline operators; 3) develop an NPBN model for predicting the corrosion depth using the pipeline age and soil parameters as predictors; 4) develop a methodology for determining the optimal sample size by balancing the sampling cost and associated benefit, which is then used to solve two SSD problems in the context of corrosion management of pipelines. It

is expected that the developed models and methodology in this thesis can benefit the integrity management of energy pipelines with respect to corrosion and third-party damage.

1.3 Scope of the study

Chapters 2 through 5 present four main topics, respectively. Chapter 2 integrates the quantification of measurement errors of in-line inspection (ILI) tools, corrosion growth modeling and reliability analysis in a single dynamic Bayesian network (DBN) model for the reliability-based corrosion management of oil and gas pipelines. The EM algorithm in the context of the parameter learning technique is employed to learn the parameters of the DBN model. The effectiveness of the parameter learning and the predictive accuracy of the DBN model are validated by the simulated and real corrosion data, respectively. Chapter 3 develops a BN model to estimate the probability of a given pipeline being hit by third-party excavations by taking into account common protective and preventative measures. The EM algorithm is employed to learn the parameters of the BN model from datasets that consist of individual cases of third-party activities but with missing information. The effectiveness of the parameter learning for the developed Bayesian network is demonstrated by a numerical example involving simulated datasets of third-party activities and a case study using real-world datasets obtained from a major pipeline operator in Canada. Chapter 4 develops an NPBN model to predict the corrosion depth on buried pipelines using the pipeline age and local soil parameters as predictors. The dependence structure and parameters of the NPBN model are extracted from a corrosion dataset in the open literature, which consists of individual samples of the corrosion depth, pipeline age together with a group of parameters characterizing the corrosive properties of local soil such as water content, redox potential, pH value. The 5-fold cross-validation is used to examine the predictive capability of the developed model. Chapter 5 establishes a methodology of SSD for collecting samples to update the distributions of basic random variables, thus reduce the epistemic uncertainty on the evaluated failure probability. The methodology is developed based on the pre-posterior analysis of the probability mass functions (PMFs) of basic random variables in the reliability analysis and the theory of value of information (VoI). The developed methodology is then applied to solve two SSD problems in the context of corrosion assessment of buried pipelines: determining the

sample size of the model error of a burst capacity model and determining the number of pipe joints to excavate for the corrosion assessment of unpiggable pipelines.

An overview of the remainder of the thesis is given in Fig. 1.5, based on which the link between the four research projects in Chapters 2 through 5 is described as follows. In the aspect of techniques, Chapters 2 and 3 both employ the BNs to model the dependence structure of a set of discrete/discretized random variables, and parameter learning technique to extract model parameters from the datasets collected by the pipeline industry. Chapter 4 employs NPBN and the model mining technique to construct both the dependence structure between a set of continuous random variables and model parameters (i.e. rank correlations) from a multivariate dataset. The SSD methodology presented in Chapter 5 is established on the basis of two key ideas: the discretization of continuous random variables and Bayesian pre-posterior analysis by exploiting the Dirichlet-Multinomial conjugate pair, which originate from the parameter learning theory of Bayesian networks. In the aspect of engineering practice, Chapters 2 and 4 deal with corrosion assessment of pipelines with ILI data and pipelines lacking ILI data (i.e. unpiggable pipelines), respectively. Chapter 3 discusses the problem of TPD management. Lastly, the SSD methodology developed in Chapter 5 can be applied for the sampling planning for reducing the epistemic uncertainty involved in the reliability analysis of corroded pipelines, which is illustrated by two numerical examples.

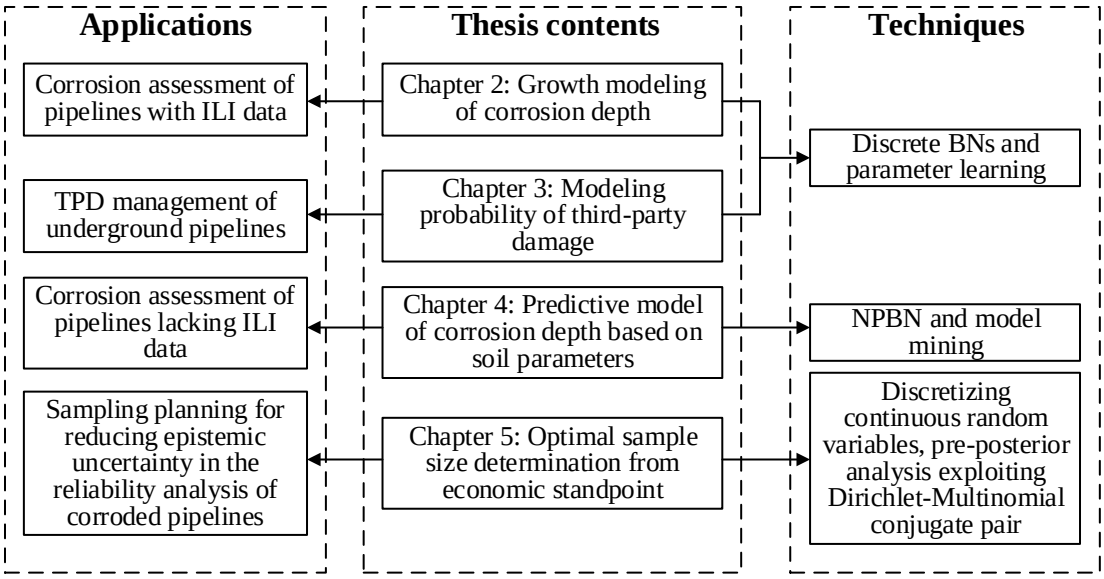


Figure 1.5 Overview of the research topics in the thesis

1.4 Thesis format

This thesis is prepared in an Integrated-Article Format as specified by the School of Graduate and Postdoctoral Studies at Western University, London, Ontario, Canada. Six chapters are included in the thesis. Chapter 1 presents the introduction of the thesis which includes the research background, objective and research significance, scope of the study and thesis format. Chapters 2 through 5 are the main body of the thesis, of which each chapter solves an individual topic. The main conclusions, limitations and recommendations for future research regarding the topics in the thesis are provided in Chapter 6.

References

Adianto, R., Nessim, M., Kariyawasam, S., and Huang, T. (2018). Implementation of Reliability-Based Criteria for Corrosion Assessment. In *Proceedings of the 12th International Pipeline Conference*. Calgary, Alberta, Canada.

Ahammed, M. (1998). Probabilistic estimation of remaining life of a pipeline in the presence of active corrosion defects. *International Journal of Pressure Vessels and Piping*, 75, 321-329.

Al-Amin, M., & Zhou, W. (2014). Evaluating the system reliability of corroding pipelines based on inspection data. *Structure and Infrastructure Engineering*, 10(9), 1161-1175.

Al-Amin, M., Zhou, W., Zhang, S., Kariyawasam, S., & Wang, H. (2012). Bayesian model for calibration of ILI tools. In: *Proceedings of the 9th International Pipeline Conference* (pp. 201-208), Calgary, Alberta, Canada.

Beauregard, Y., Woo, A., and Huang, T. (2018). Application of In-Line Inspection and Failure Data to Reduce Subjectivity of Risk Model Scores for Uninspected Pipelines. In *Proceedings of the 12th International Pipeline Conference* (pp. V002T07A028-V002T07A028). Calgary, Alberta, Canada.

Caleyo, F., Velázquez, J. C., Valor, A., and Hallen, J. M. (2009a). Markov chain modelling of pitting corrosion in underground pipelines. *Corrosion Science*, 51(9), 2197-2207.

Caleyo, F., Velázquez, J. C., Valor, A., and Hallen, J. M. (2009b). Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study. *Corrosion Science*, 51(9), 1925-1934.

Chen, Q., Nessim, M. A. 1999. Reliability-based prevention of mechanical damage to pipelines. submitted to Pipeline Research Council International, Inc., Catalogue, (L51816).

Cosham, A., Hopkins, P., and Macdonald, K. A. (2007). Best practice for the assessment of defects in pipelines–Corrosion. *Engineering Failure Analysis*, 14(7), 1245-1265.

Der Kiureghian, A. (1989). Measures of structural safety under imperfect states of knowledge. *Journal of Structural Engineering*, 115(5), 1119-1140.

Green, K. P., and Jackson, T. (2015). *Safety in the transportation of Oil and Gas: pipelines or rail?*. Vancouver, BC: Fraser Institute.

Hanea, A. M., Gheorghe, M., Hanea, R., and Ababei, D. (2013). Non-parametric Bayesian networks for parameter estimation in reservoir simulation: a graphical take on the ensemble Kalman filter (part I). *Computational geosciences*, 17(6), 929-949.

Hanea, A., Napoles, O. M., and Ababei, D. (2015). Non-parametric Bayesian networks: Improving theory and reviewing applications. *Reliability Engineering and System Safety*, 144, 265-284.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In: Jordan M. (Eds.), *Learning in graphical models* (pp. 301-354), Springer, Dordrecht.

Higo, E., and Pandey, M. D. (2016). Value of information and hypothesis testing approaches for sample size determination in engineering component inspection: a comparison. In: *Proceedings of ASME 2016 Pressure Vessels and Piping Conference* (pp. V005T10A009-V005T10A009), Vancouver, British Columbia, Canada.

Hong, H-P. (1999) Application of the stochastic process to pitting corrosion. *Corrosion*, 55(1):10-16.

Kariyawasam, S. and Peterson, W. (2008). Revised corrosion management with reliability based excavation criteria. In Proceedings of the 7th International Pipeline Conference. Calgary, Alberta, Canada.

Kurowicka, D., and Cooke, R. M. (2005). Distribution-free continuous Bayesian belief. *Modern statistical and mathematical methods in reliability*, 10, 309.

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

Lam, C., and Zhou, W. (2016). Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database. *Internal Journal of Pressurized Vessels and Piping*, 145, 29-40.

Lee, D., and Pan, R. (2018). A nonparametric Bayesian network approach to assessing system reliability at early design stages. *Reliability Engineering and System Safety*, 171, 57-66.

Luque, J., Straub, D. (2016). Reliability analysis and updating of deteriorating systems with dynamic Bayesian networks. *Structural Safety*, 62, 34-46.

Mahadevan, S., Zhang, R., and Smith, N. (2001). Bayesian networks for system reliability reassessment. *Structural Safety*, 23(3), 231-251.

Mahjoub, M. A., and Kalti, K. (2011). Software comparison dealing with bayesian networks. In *International Symposium on Neural Networks* (pp. 168-177). Springer, Berlin, Heidelberg.

Maes, M. A., Faber, M. H., and Dann, M. R. (2010). Hierarchical modeling of pipeline defect growth subject to ILI uncertainty. In *Proceedings of the 28th International Conference on Ocean, Offshore and Arctic Engineering* (pp. 375-384). Honolulu, Hawaii, USA.

Melchers, R. E., and Petersen, R. B. (2018). A reinterpretation of the Romanoff NBS data for corrosion of steels in soils. *Corrosion Engineering, Science and Technology*, 53(2), 131-140.

Morales-Nápoles, O., Delgado-Hernández, D. J., De-León-Escobedo, D., and Arteaga-Arcos, J. C. (2014). A continuous Bayesian network for earth dams' risk assessment: methodology and quantification. *Structure and Infrastructure Engineering*, 10(5), 589-603.

Morales-Nápoles, O., and Steenbergen, R. D. (2014). Large-scale hybrid Bayesian network for traffic load modeling from weigh-in-motion system data. *Journal of Bridge Engineering*, 20(1), 04014059.

Nielsen, T., and Jensen, F. (2009). *Bayesian networks and decision graphs*. New York, NY: Springer Science and Business Media.

Nishijima, K., and Faber M.H. (2007) Bayesian approach to proof loading of quasi-identical multi-components structural systems. *Civil Engineering and Environmental Systems*, 24 (2), 111-121.

Pandey, M.D., Yuan, X.-X., and van Noortwijk, J.M. (2009). The influence of temporal uncertainty of deterioration on life-cycle management of structures. *Structure and Infrastructure Engineering*, 5(2), 145-156.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Beliefs*. Morgan Kaufmann, San Mateo, Calif.

Rau, J., and Kirkwood, M. (2016). Hydrotesting and In-Line Inspection: Now and in the Future. In *Proceedings of the 11th International Pipeline Conference* (pp. V001T03A055-V001T03A055), Calgary, Alberta, Canada.

Ricker, R. E. (2010). Analysis of pipeline steel corrosion data from NBS (NIST) studies conducted between 1922–1940 and relevance to pipeline management. *Journal of research of the National Institute of Standards and Technology*, 115(5), 373.

Straub, D. (2009). Stochastic modeling of deterioration processes through dynamic Bayesian networks. *Journal of Engineering Mechanics*, 135(10), 1089-1099.

Straub, D., Der Kiureghian, A. (2010a). Bayesian network enhanced with structural reliability methods: methodology. *Journal of engineering mechanics*, 136(10), 1248-1258.

Straub, D., Der Kiureghian, A. (2010b). Bayesian network enhanced with structural reliability methods: application. *Journal of Engineering Mechanics*, 136(10), 1259-1270.

Tomic, A., Huang, T., and Kariyawasam, S. (2018). System Wide Risk Assessment in the 21st Century: TransCanada's Approach. In *Proceedings of the 12th International Pipeline Conference*. Calgary, Alberta, Canada.

Velázquez, J. C., Caleyo, F., Valor, A., and Hallen, J. M. (2009). Predictive model for pitting corrosion in buried oil and gas pipelines. *Corrosion*, 65(5), 332-342.

Velázquez, J. C., Caleyo, F., Valor, A., and Hallen, J. M. (2010). Field Study—Pitting Corrosion of Underground Pipelines Related to Local Soil and Pipe Characteristics. *Corrosion*, 66(1), 016001-016001.

Wang, F., Li, H., Dong, C., and Ding, L. (2019). Knowledge representation using non-parametric Bayesian networks for tunneling risk analysis. *Reliability Engineering and System Safety*, 106529.

Zhang, S. (2014). Development of probabilistic corrosion growth models with applications in integrity management of pipelines. Western University. London, Ontario, Canada.

Zhang, S., Zhou, W., Al-Amin, M., Kariyawasam, S., Wang, H. (2014). Time-Dependent Corrosion Growth Modeling Using Multiple In-Line Inspection Data. *Journal of Pressure Vessel Technology*, 136(4), 041202.

Zhang, S., Zhou, W., Qin, H. (2013). Inverse Gaussian process-based corrosion growth model for energy pipelines considering the sizing error in inspection data. *Corrosion Science*, 73, 309-320.

Zhou, W. (2010). System reliability of corroding pipelines. *International Journal of Pressure Vessels and Piping*, 87(10): 587-595.

Zhou, W., Xiang, W., and Hong, H. P. (2017). Sensitivity of system reliability of corroding pipelines to modeling of stochastic growth of corrosion defects. *Reliability Engineering & System Safety*, 167, 428-438.

Zilko, A. A., Kurowicka, D., and Goverde, R. M. (2016). Modeling railway disruption lengths with Copula Bayesian Networks. *Transportation Research Part C: Emerging Technologies*, 68, 350-368.

2 Integrated pipeline corrosion growth modeling and reliability analysis using the dynamic Bayesian network and parameter learning technique

2.1 Introduction

Historical failure data indicate that metal-loss corrosion is one of the leading threats to the structural integrity of underground oil and gas pipelines (CEPA, 2015). In the past few decades, in-line inspections (ILIs) have been widely adopted by the pipeline industry to detect and size the corrosion defects on pipelines (Kariyawasam and Peterson, 2010). The pipeline corrosion management program typically includes periodical ILIs to detect and size corrosion defects on pipeline segments, engineering critical assessment of the detected corrosion defects, and appropriate mitigation actions or scheduling the future ILIs. The accurate modeling of corrosion growth is of great importance to the corrosion management program. Critical corrosion defects may be missed by scheduled mitigation activities if the corrosion growth is significantly underestimated. On the other hand, overly conservative estimates of the growth may lead to unnecessary mitigation actions, which translates into significant cost penalties to pipeline operators.

It is advantageous to model the corrosion growth probabilistically to account for the inherent random nature of the corrosion growth. To this end, extensive studies have been reported in the literature, e.g. the linear and power-law growth models (Ahammed, 1998; Al-Amin and Zhou, 2014; Amirat et al., 2010), and stochastic process-based growth models (Hong, 1999; Pandey et al., 2009; Valor et al., 2007; Zhou et al., 2017). The hierarchical Bayesian models and Markov Chain Monte Carlo (MCMC) simulation technique have been employed to effectively estimate the parameters of corrosion growth models based on the ILI data (Pandey et al., 2009). However, the complexity of the hierarchical Bayesian model and MCMC technique renders them difficult to use by non-specialists in practice.

The defect depth reported by an ILI tool is typically assumed to be a linear function of the actual depth subjected to a random scattering error. The slope and intercept of the linear function are called the multiplicative and additive biases associated with the ILI tool,

respectively. The probabilistic characteristics of the multiplicative and additive biases, as well as the standard deviation of the scattering error are typically evaluated from regression or Bayesian analyses (Al-Amin et al., 2012; Nessim et al., 2008) and then incorporated in the growth model as an input (Al-Amin and Zhou, 2014). The updated growth model also becomes an input in the reliability-based defect assessment, e.g. evaluation of the time-dependent probabilities of failure of individual active corrosion defects and/or system failure probability of a given pipe segment containing multiple active defects (Al-Amin and Zhou, 2014). It hinders the wide application of the Bayesian growth model in practice that three separate models are employed for the ILI measurement error characterization, growth model updating, and reliability analysis. It is therefore desirable from a practical perspective to combine the three components into a single integrated analysis. This is the main motivation of the present study.

Due to the intuitive graphical nature and ability to efficiently handle the Bayesian updating of a set of random variables, Bayesian networks (BNs) have become increasingly popular in the engineering reliability and risk analysis during the last two decades. The applications closely relevant to the present study include using dynamic Bayesian networks (DBNs) to model the deterioration of structures (Luque and Straub, 2016; Straub, 2009), and utilizing BNs to evaluate and update the reliability of structures (Mahadevan et al., 2001; Straub and Der Kiureghian, 2010). As an important feature of BNs, the parameter learning technique (Heckerman, 1998; Spiegelhalter et al., 1993) provides an objective, efficient means to elicit model parameters in BNs from sparse data. While studies of parameter learning algorithms have been extensively reported in the field of Bayesian artificial intelligence (Heckerman, 1998; Masegosa et al., 2016; Spiegelhalter et al., 1993; Zhou et al., 2016), the application of this technique in civil engineering is scarce so far.

In the present study, an integrated DBN model is developed to characterize the growth of depths of corrosion defects based on the ILI data and evaluate the time-dependent failure probabilities of active corrosion defects. The novelty of the study is two-fold. First, three critical components in the pipeline corrosion management, i.e. ILI measurement error characterization, growth model updating, and reliability analysis are integrated into a single DBN model. Second, the parameter learning technique is employed to evaluate the

probabilistic characteristics of the measurement errors associated with the ILI tools and parameters of the defect growth model, based on the ILI-reported and field-measured depths of corrosion defects. This allows the parameters of the DBN model to be quantified in an automated and objective manner. The proposed model is illustrated and validated through a numerical example involving simulated corrosion data, and applied to real ILI and field-measured corrosion data obtained from an in-service natural gas pipeline in Canada. The remainder of this chapter is organized as follows. The basics of BNs and parameter learning technique are briefly presented in Section 2.2. The proposed DBN model is described in Section 2.3. The application of the proposed model on simulated and real corrosion data is described in Section 2.4, followed by conclusions in Section 2.5.

2.2 Basics of Bayesian networks and parameter learning

A brief introduction of BNs and the parameter learning technique is presented in the following. More detailed discussions of BNs can be found in many textbooks (e.g. Nielsen and Jensen, 2009; Pearl, 2014). BNs are directed graphical models representing the joint probabilistic distribution of a set of random variables, which are symbolized by nodes in BNs. While BN modeling can handle both discrete and, in some special cases, continuous random variables (Langseth et al., 2009), BN models discussed in this work are limited to discrete random variables. Therefore, random variables that are inherently continuous will be discretized in the BN models. The discrete values of a given random variable are called states. The dependence between nodes is symbolized by directed arcs and quantified by conditional probability tables (CPTs) attached to them. The CPTs enable BNs to factor the high-dimensional joint probability distribution into local conditional probability distributions. The assignment of observed values to the corresponding nodes is called the instantiation of the nodes, which can lead to the Bayesian updating of the nodes that are dependent on the instantiated nodes. As an example, consider the BN model shown in Fig. 2.1. The nodes A and B with arcs pointing to node C are the parents of C, denoted by pa(C), and C is the child node of A and B. The nodes A and B are called the root nodes, as they do not have parent nodes. Note that the CPTs of root nodes coincide with their probability mass functions (PMFs). The joint PMF of all the random variables involved in this model is expressed as follows by the chain rule,

$$p(a,b,c,d) = p(a)p(b)p(c|a,b)p(d|c) \tag{2.1}$$

where *a*, *b*, *c*, *d* denote the states of *A*, *B*, *C*, *D*, respectively; $p(\bullet)$ denotes the PMF of node ●, and $p(\bullet|)$ denotes the conditional PMF for node ●. If the state of *C* is observed to be c_e, the posterior joint PMF of the rest of the nodes can be calculated based on Bayes' rule as follows,

$$p(a,b,d|c_e) = \frac{p(a)p(b)p(c_e|a,b)p(d|c_e)}{\sum_{A,B} p(a)p(b)p(c_e|a,b)} \tag{2.2}$$

Then, the posterior marginal distribution of each node can be calculated by summing out the rest of the nodes from the posterior joint PMF given in Eq. (2.2). The Bayesian updating can be made more efficient by transferring the BN model into a junction tree and carrying out the Bayesian inference with cliques as opposed to individual nodes. The junction tree algorithm (Nielsen and Jensen, 2009) has been the standard algorithm implemented in most BN software.

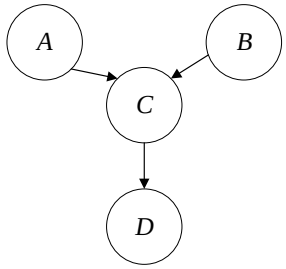


Figure 2.1 An example BN model

As a special case of BNs, DBNs have been used to model the stochastic deterioration of engineering structures (Luque and Straub, 2016; Rafiq et al., 2010; Straub, 2009). A DBN consists of a sequence of slices, each of which contains one or multiple nodes characterizing the system state at a specific temporal point. The dependence between different slices is symbolized by arcs that link nodes in different slices. If the dependence between nodes within a slice and between slices is identical for all the slices except for the first one, the DBN is referred to as a homogeneous DBN (Murphy, 2002; Straub, 2009). An example of expanded DBN is given in Fig. 2.2, while it can also be defined in a compact form with only the first two slices. When the nodes in a certain time-slice are instantiated,

the querying of previous, current and future time-slices in a DBN is termed as smoothing, filtering and prediction, respectively. The naïve inference algorithm on expanded DBNs is the same as the junction tree algorithm for BNs, i.e. treating the expanded DBN as a BN containing the nodes in all the time-slices. This is computationally inefficient for DBNs with a large number of time-slices. The special algorithms such as the Frontier Algorithm (Zweig, 1996) and Interface Algorithm (Murphy, 2002) were developed to adapt the junction tree algorithm to perform the inference for DBNs in more efficient manners.

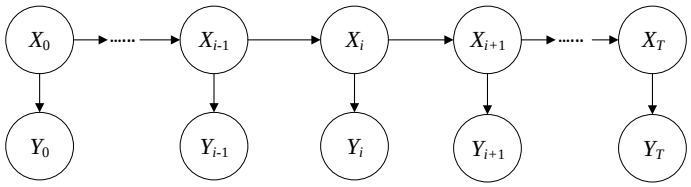


Figure 2.2 An example DBN model

The entries in the CPTs are called parameters of BNs, which are usually specified based on a combination of mathematical models, expert opinions, and collected data. The parameter learning is an effective approach to obtain Bayesian estimates of parameters of BNs from a set of observations on the nodes. A brief description of the parameter learning is provided below using parameters of node C in the BN model depicted in Fig. 2.1 as an example. Readers are referred to Heckerman (1998) for details. Let $p(c_j|\text{pa}(C)_k)$ (j = 1, 2, ..., r_c, k = 1, 2, …, r_{pa}) denote the parameters of C, i.e. the probability of the j-th state (c_j) under the k-th parent configuration ($\text{pa}(C)_k$), where r_c and r_{pa} are the total numbers of states and parent configurations of C, respectively. For notational simplicity, $p(c_j|\text{pa}(C)_k)$ is replaced by the shorthand notation $\theta_{j,k}$ hereafter. It follows that $\sum_{j=1}^{r_c} \theta_{j,k} = 1$ for k = 1, 2, …, r_{pa}. For a given parent configuration k, $\theta_{j,k}$ (j = 1, 2, ..., r_c) are considered as a vector of random variables following a Dirichlet distribution with parameters $\alpha_{1,k}$, $\alpha_{2,k}$, …, $\alpha_{r_c,k}$ (Heckerman, 1998). Before observations are obtained, the estimated value of $\theta_{j,k}$, denoted by $\hat{\theta}_{j,k}$, can be set to the corresponding mean of the Dirichlet distribution,

$$\hat{\theta}_{j,k} = \frac{\alpha_{j,k}}{\alpha_{0,k}} \tag{2.3}$$

where $\alpha_{0,k} = \sum_{j=1}^{r_c} \alpha_{j,k}$ is known as the equivalent sample size of the Dirichlet distribution (Heckerman, 1998).

Once a set of observations are obtained, the Bayesian updating of the distribution of $\theta_{j,k}$ (j = 1, 2, ..., r_c) is carried out. Consider first the simple scenario of complete (no missing) data, i.e. each of the observations containing values of *A*, *B*, *C* and *D*. The observations are considered drawn from a multinomial distribution (Heckerman, 1998). Given the Dirichlet-multinomial conjugate pair, the posterior distribution of $\theta_{j,k}$ is also a Dirichlet distribution with parameters $\alpha_{1,k} + n_{1,k}$, $\alpha_{2,k} + n_{2,k}$, …, $\alpha_{r_c,k} + n_{r_c,k}$, where $n_{j,k}$ (j = 1, 2, … r_c) is the number of observations of *C* in the *j*-th state under the *k*-th parent configuration. With the observations, $\hat{\theta}_{j,k}$ can be set to the mean of the posterior Dirichlet distribution, i.e.

$$\hat{\theta}_{j,k} = \frac{\alpha_{j,k}+n_{j,k}}{\alpha_{0,k}+n_{0,k}} \tag{2.4}$$

where $n_{0,k} = \sum_{j=1}^{r_c} n_{j,k}$. This completes the parameter learning for *C* under the complete data scenario.

Now consider the scenario of incomplete or missing data, which is often encountered in practice. Assume that there are a total of *n* sets of observations (i.e. *n* cases), each of which contains values of *A*, *B* and *D*, but misses the value of *C*. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is commonly employed to learn the parameters of *C*. To this end, the posterior distribution of $\theta_{j,k}$ is a Dirichlet distribution with parameters $\alpha_{1,k}$ + E[$n_{1,k}$], $\alpha_{2,k}$ + E[$n_{2,k}$], …, $\alpha_{r_c,k}$+ E[$n_{r_c,k}$], where E[$n_{j,k}$] (j = 1, 2, … r_c) is the expected number of observations of *C* in the *j*-th state under the *k*-th parent configuration. The value of E[$n_{j,k}$] is estimated as follows,

$$\mathrm{E}[n_{j,k}] = \sum_{l=1}^{n} p(c_j, \mathrm{pa}(C)_k | O_l) \tag{2.5}$$

where $p(c_j, \mathrm{pa}(C)_k | O_l)$ is the probability of c_j under pa(C)$_k$ given the *l*-th (l = 1, 2, …, n) case O_l, and can be obtained from the BN inference once the BN is instantiated by the evidence in O_l, i.e. corresponding values of *A*, *B* and *D*. The value of $\hat{\theta}_{j,k}$ is now given by

$$\hat{\theta}_{j,k} = \frac{\alpha_{j,k}+\mathrm{E}[n_{j,k}]}{\alpha_{0,k}+\sum_{j=1}^{r_c}\mathrm{E}[n_{j,k}]} \tag{2.6}$$

It follows that the evaluation of Eqs. (2.5) and (2.6) is an iterative process, as $\hat{\theta}_{j,k}$ obtained in the current iteration is used to estimate E[$n_{j,k}$] and thus leads to a new $\hat{\theta}_{j,k}$ in the next iteration. The iteration is terminated once the log-likelihood of the observations converges to a local maximum, and this completes the parameter learning for C under the missing data scenario.

2.3 Corrosion growth modeling by a dynamic Bayesian network

This section presents the development of the DBN-based growth model for the defect depth (i.e. in the through-pipe wall thickness direction) and procedures to learn the parameters of root nodes from ILI-reported and field-measured defect depths. The growth of the depth of an individual defect is assumed to follow a linear function of time with an uncertain annual growth rate, and is modeled by a DBN at discrete time points. The ILI-reported defect depths can be used to instantiate the corresponding nodes in the DBN for model updating. The description is based on a DBN model for a given corrosion defect as depicted in Fig. 2.3, which includes four time-slices. Time-slice 0 represents year 0, i.e. the time of the first ILI considered in the modeling, and each subsequent time-slice represents an increment of one year. Nodes X_0, X_1, X_2 and X_3 represent the defect depths at years 0, 1, 2 and 3, respectively. It follows that $X_1 - X_0 = X_2 - X_1 = X_3 - X_2 = \Delta X$, where ΔX is the depth increment within a year. It is assumed that ILI is carried out at years 0 and 2. Nodes Y_0 and Y_2 represent the ILI-reported defect depths at years 0 and 2, respectively. The relationship between Y_i and X_i (i = 0, 2) is defined by (Al-Amin et al., 2012),

$$Y_i = \beta_i X_i + \gamma_i + \varepsilon_i \ (i = 0, 2) \tag{2.7}$$

where β_i, γ_i and ε_i denote the multiplicative bias, additive bias and random scattering error, respectively, associated with the ILI tool employed at year i. Typically, β_i and γ_i are considered as deterministic quantities, and ε_i is assumed to follow a zero-mean Gaussian distribution with the standard deviation denoted by σ_i (Al-Amin et al., 2012). To reduce

the number of parameters involved in the parameter learning and thus increase its effectiveness, Eq. (2.7) is re-written as follows,

$$Y_i = \beta_i X_i + \gamma_i' \ (i = 0, 2) \tag{2.8}$$

where γ_i' follows the Gaussian distribution with the mean value and standard deviation equal to γ_i and σ_i, respectively. Based on Eq. (2.8), a single node γ_i' is used in the DBN model to account for the additive bias γ_i and random scattering error ε_i of the ILI tool.

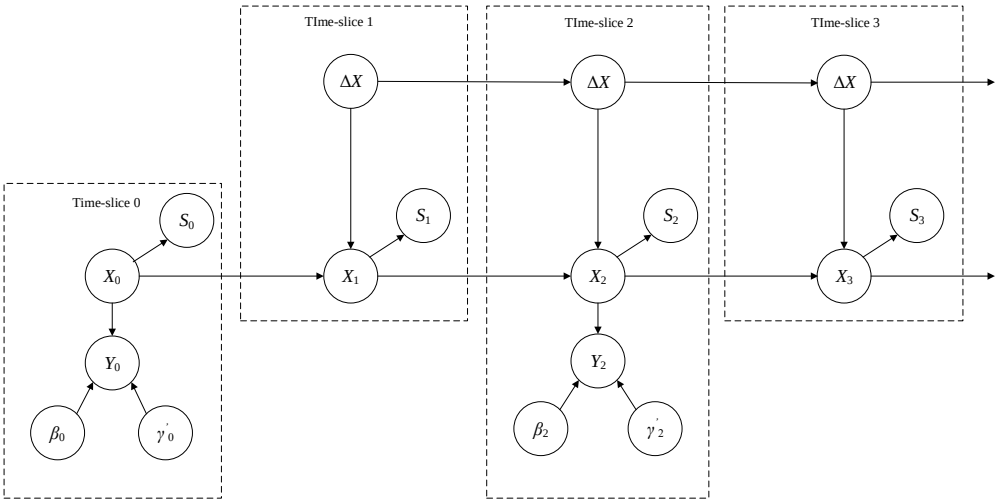


Figure 2.3 Conceptual DBN growth model of defect depth

Since all the random variables contained in the above-described DBN model are continuous in nature, they are first discretized. While the dynamic discretization technique (Marquez et al., 2010) can be employed to make the discretization adaptive to achieve the most accurate characterization of the high density regions of the distribution, the present study adopts a simple discretization scheme: each node is partitioned by a set of equal intervals, with the interval size pre-selected. The CPTs of X_j (j = 0, 1, 2, 3) and Y_i (i = 0, 2) are created using the Monte Carlo (MC) simulation (Straub, 2009; Straub and Der Kiureghian, 2010). For instance, the entries of the CPT of Y_2 conditioned on a given parent configuration are created as follows. The values of X_2, β_2 and γ_2' are assumed to be uniformly distributed between the lower and upper bounds of given states, from which the samples can be generated. The samples of Y_2 can then be calculated using Eq. (2.8). The

counts of samples of Y_2 lying in a certain state of Y_2 normalized by the total number of samples is the parameter associated with that state.

Although ΔX appears in time-slices 1, 2 and 3, it is emphasized that ΔX in time-slices 2 and 3 is a copy of ΔX in time-slice 1, thus consistent with the linear growth (i.e. constant growth rate) model adopted in this study. Symmetric Dirichlet distributions with the equivalent sample size equal to unity, corresponding to non-informative prior distributions (Zhou et al., 2016), are assigned to the PMFs of root nodes, i.e. β_0, β_2, γ_0', γ_2', X_0 and ΔX, prior to carrying out the parameter learning on them.

The node S_j (j = 0, 1, 2, 3) has binary states (i.e. survival and failure); the probability of the failure state is the cumulative failure probability of the defect under the internal pressure up to year j. Note that the value of S_3 is of primary interest, as it is the predicted cumulative failure probability up to year 3 by taking into account the defect growth model updated based on the ILI data at years 0 and 2. While both leak and burst failure modes (Zhou, 2010) can be considered, only the burst failure mode is included in the present model for simplicity.

The limit state function, g, considered in the evaluation of CPT of S_j is given by,

$$g = r_b - o_p \tag{2.9}$$

$$r_b = \kappa \frac{2w_t(\sigma_y + 68.95)}{D} \left[\frac{1 - 0.85\frac{d}{w_t}}{1 - 0.85\frac{d}{Mw_t}} \right] \tag{2.10}$$

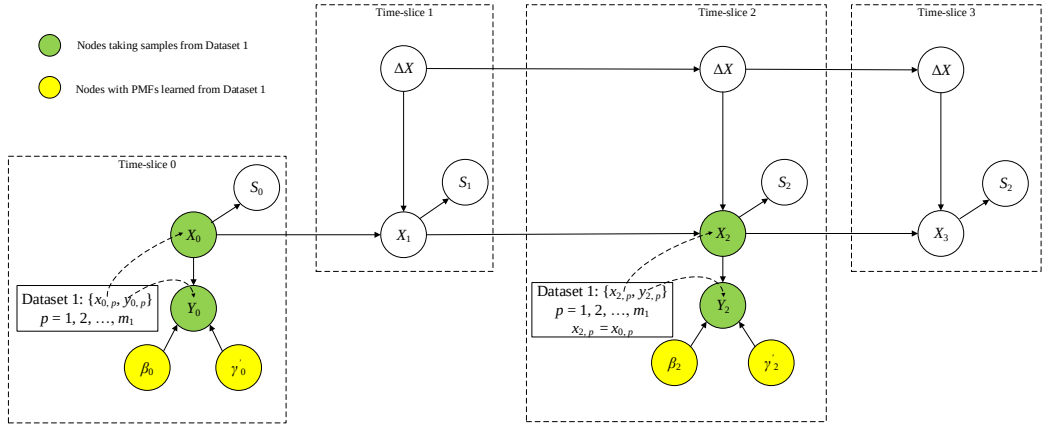
$$M = \begin{cases} \sqrt{1 + 0.6275\frac{l^2}{Dw_t} - 0.003375\left(\frac{l^2}{Dw_t}\right)^2} & l \le \sqrt{50Dw_t} \\ 3.3 + 0.032\frac{l^2}{Dw_t} & l > \sqrt{50Dw_t} \end{cases} \tag{2.11}$$

where o_p is the operating pressure of the pipeline; r_b is the burst pressure capacity of the pipe at the defect, evaluated using the B31G Modified model (Kiefner and Veith, 1989); d (i.e. X_j in the DBN model) is the actual defect depth; D is the pipe outside diameter; w_t is the actual pipe wall thickness; σ_y is the yield strength of the pipe steel; κ denotes the model error associated with the B31G Modified model; M is the Folias bulging factor, and l is the

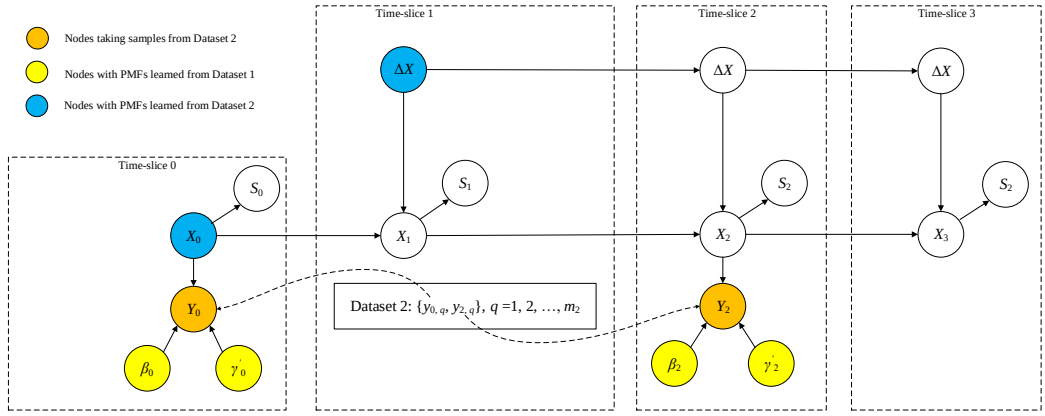
defect length. Since all the relevant random variables (such as σ_y, D, and o_p) other than X_j are assumed to have no observations for updating, they are treated as continuous random variables and incorporated in the simple Monte Carlo simulation to develop the CPT of S_j (Straub and Der Kiureghian, 2010); in other words, they are not explicitly considered in the DBN growth model.

The PMFs of β_0, γ_0', β_2, γ_2', X_0 and ΔX are developed by applying the parameter learning technique based on two distinct datasets of corrosion defects, referred to as Datasets 1 and 2 respectively, consistent with the typical pipeline corrosion management practice. In practice, once an ILI is conducted, pipeline engineers usually select a set of pipe joints to be excavated almost immediately after the ILI (a pipeline consists of many pipe joints welded together with each pipe joint about 12 – 20 m long) for the purpose of verifying the accuracy of the ILI data as well as repairing those pipe joints containing critical defects, i.e. defects with ILI-reported sizes exceeding the safety limit. The sizes of all the corrosion defects on the excavated pipe joints are measured in the ditch using laser scans. Since laser scans have negligible measurement errors (Al-Amin et al., 2012), the field-measured defect sizes can be assumed to equal the corresponding actual defect sizes. Furthermore, all the excavated pipe joints are repaired and recoated before reburied; therefore, the growth of these corrosion defects is arrested after repair. Such defects are referred to as static defects. The static defects will also be sized by ILIs conducted in the future. It follows that the field-measured and ILI-reported depths for the static defects establish a dataset (i.e. Dataset 1) that is used to quantify the measurement errors associated with multiple sets of ILI data. On the other hand, there are defects that have not been mitigated, referred to as active defects. The depths of active defects reported by ILIs conducted at different times establish Dataset 2, which is used to develop the growth model for the active defects. For the example shown in Fig. 2.3, Dataset 1 includes a total of m_1 static corrosion defects with the ILI-reported depths and actual (i.e. field-measured) depths at time-slices 0 and 2, and is used to learn the parameters of β_0, β_2, γ_0' and γ_2' using the EM algorithm, as illustrated in Fig. 2.4(a). Figure 2.4(a) indicates that, in the first step of the parameter learning, the green nodes (i.e. X_0, Y_0, X_2 and Y_2) in the DBN take samples from Dataset 1, and the parameters of the yellow nodes (i.e. β_0, β_2, γ_0' and γ_2') are learned. Dataset 2 contains the ILI-reported

defect depths at time-slices 0 and 2 for m_2 active corrosion defects, and is used to learn the parameters of X_0 and ΔX as illustrated in Fig. 2.4(b), given the parameters of β_0, β_2, γ_0' and γ_2' learned from Dataset 1. Figure 2.4(b) indicates that, in the second step of the parameter learning, the orange nodes (i.e. Y_0 and Y_2) take samples from Dataset 2, and the parameters of the blue nodes (i.e. X_0 and ΔX) are learned.



(a) Step 1: Learn the parameters of β_i and γ_i' using Dataset 1



(b) Step 2: Learn the parameters of X_0 and ΔX using Dataset 2

Figure 2.4 Illustration of parameter learning in two sequential steps using Datasets 1 and 2

2.4 Illustrative examples and model validation

This section first describes a numerical example involving simulated corrosion data to demonstrate the effectiveness of the proposed DBN model and parameter learning technique by the means of comparing the learned parameters with the corresponding values prescribed in the data simulation. Then, the DBN growth model is developed and validated using real corrosion data.

2.4.1 Example 1: simulated corrosion data

This example considers a pipeline inspected by three ILI tools denoted by I-0, I-3 and I-6, at years 0, 3 and 6, respectively. Each of Datasets 1 and 2 contains 100 defects. The actual defect depths in Dataset 1 are generated as independent samples of a Weibull distribution with the corresponding mean and coefficient of variation (COV) equal to $0.4w_{tn}$ and 30%, respectively, where w_{tn} is the nominal pipe wall thickness. The actual depths in Dataset 2 at year 0 are independent samples of a Weibull distribution with the corresponding mean and COV equal to $0.3w_{tn}$ and 30%, respectively. The annual depth growths of the defects in Dataset 2 are further generated as independent samples of a gamma distribution with the corresponding mean and COV equal to $0.04w_{tn}$ and 50%, respectively. The prescribed parameters characterizing the measurement errors associated with the three ILI tools are shown in Table 2.2. Note that truncations, if necessary, are performed in the process of simulation such that the simulated actual and ILI-reported depths are within the range of 0 to w_{tn}.

The DBN growth model consists of 7 time-slices, i.e. years 0 through 6 (Fig. 2.5) and is implemented using the commercial BN software Netica®. The time-slices corresponding to years 0, 3 and 6 contain nodes representing measurement errors and ILI-reported depths. The failure probability of the defect is not evaluated for this example as the focus is on validating the DBN model in terms of quantifying the ILI measurement errors and defect growth rate; therefore, the failure probability node S_j (j = 0, 1, …, 6) is not included in the model. The discretization schemes adopted for the random variables involved in the DBN are summarized in Table 2.1. The lower and upper bounds of the random variables are selected based on a combination of physical limits and subjective considerations. For

example, the lower and upper bounds of the defect depth (X_j, j = 0, 1, …, 6) and ILI-reported defect depth (Y_i, i = 0, 3, 6) are 0 and 100%w_{tn}, respectively, based entirely on the physical limits. For the same reason, the lower and upper bounds of the additive measurement error associated with the ILI tool (γ_i', i = 0, 3, 6) are -100% and 100%w_{tn}. Fuller (1987) indicates that the multiplicative bias is generally less than 2. Therefore, the range for β_i (i = 0, 3, 6) is selected to be between 0 to 2.5. The lower bound of ΔX must be zero, whereas its upper bound is subjectively defined to be 0.1w_{tn}, considered more than adequate for typical pipelines. To investigate the effect of the discretization scheme on the parameter learning, three different sets of interval sizes for the discretization are considered, as summarized in Table 2.1, with a smaller interval leading to a more refined discretization scheme. Scheme 1 is considered as the baseline case, whereas schemes 2 and 3 are more and less refined than scheme 1, respectively.

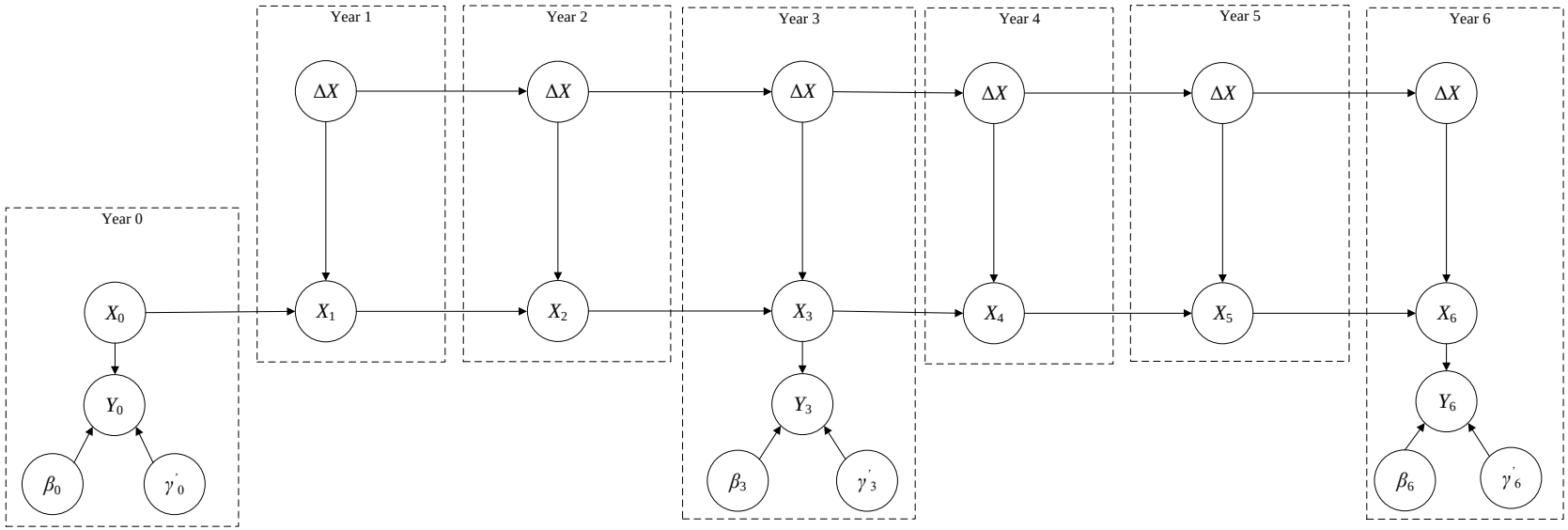


Figure 2.5 The DBN growth model for Example 1

Table 2.1 Summary of the probable range of values and discretization schemes for Example 1

Random variable	Range of values	Discretization interval		
		Scheme 1	Scheme 2	Scheme 3
ΔX	[0, 0.1] w_{tn}	0.005 w_{tn}	0.004 w_{tn}	0.01 w_{tn}
X_i	[0, 1] w_{tn}	0.05 w_{tn}	0.04 w_{tn}	0.1 w_{tn}
Y_i	[0, 1] w_{tn}	0.05 w_{tn}	0.04 w_{tn}	0.1 w_{tn}
β_i	[0, 2.5]	0.1	0.05	0.25
γ_i'	[-1, 1] w_{tn}	0.1 w_{tn}	0.05 w_{tn}	0.2 w_{tn}

The CPTs for X_j (j = 0, 1, …, 6) and Y_i (i = 0, 3, 6) are created using the MC simulation with 100,000 trials. The EM-based parameter learning is implemented in Netica® to learn the parameters of β_i and γ_i' (i = 0, 3, 6), X_0 and ΔX. The EM iteration is terminated if any of the following two conditions is met: the difference between the average log-likelihood per case in two consecutive iterations is less than 10^{-5}, and the maximum number of iterations reaches 1000.

Note that the variability in the simulated samples may introduce variability in results of the parameter learning. Therefore, the data simulation and parameter learning are repeated 10 times following the common practice (Zhou et al., 2016) of examining the accuracy of the parameter learning. The values of β_i, γ_i and σ_i (i = 0, 3, 6) are learned by considering the three discretization schemes, respectively. The mean value and standard deviation of the learned values of β_i, γ_i and σ_i (i = 0, 3, 6) calculated from the 10 trials are presented in Table 2.2. Note that in a given trial the learned value of β_i is taken as the mean of the corresponding learned (i.e. posterior) PMF, whereas the learned values of γ_i and σ_i are taken as the mean and standard deviation, respectively, of the learned PMF of γ_i'. The results indicate that discretization schemes 1 and 2 achieve better accuracy than discretization scheme 3. For β_i, γ_i and σ_i learned under the discretization schemes 1 and 2, the slight difference between the mean values of the 10 trials and prescribed values, together with the small standard deviations of the 10 trials suggest that the parameters learned in all the trails in general agree well with the prescribed values. While smaller discretization intervals are used in discretization scheme 2 than scheme 1, the improvement on the accuracy of the parameter learning is limited. On the other hand, the performance of the

parameter learning using discretization scheme 3 is unsatisfactory. Under the three discretization schemes, the parameters of X_0 and ΔX are learned (Table 2.3) using the EM algorithm based on Dataset 2, i.e. the ILI-reported depths of 100 active defects, and learned parameters of β_i and γ_i' (i = 0, 3, 6). Similar to the results in Table 2.2, Table 2.3 indicates that the learned means and COVs of X_0 and ΔX under the discretization schemes 1 and 2 agree well with the corresponding prescribed values, i.e. mean and COV of the Weibull distribution that is used to simulate X_0, and mean and COV of the gamma distribution that is used to simulate ΔX. However, in comparison with the prescribed values, the errors on the learned values under the discretization scheme 3 are relatively large. In summary, the above results suggest that discretization scheme 1 is adequate to achieve good accuracy for the parameter learning in the presented study, and the parameter learning technique can effectively infer the parameters involved in the corrosion growth model based on the field-measured and ILI-reported defect depths at different times. Further validation of the growth model by real-world data is presented in the following section.

Table 2.2 Prescribed and learned parameters for the ILI measurement errors in Example 1

		Prescribed values	Values from parameter learning		
			Scheme 1	Scheme 2	Scheme 3
I-0	β_0	1.1	1.11 ± 0.05[a]	1.10 ± 0.04	1.19 ± 0.03
	γ_0 (w_{tn})	-0.05	-0.059 ± 0.015	-0.052 ± 0.018	-0.098 ± 0.002
	σ_0 (w_{tn})	0.07	0.079 ± 0.009	0.084 ± 0.005	0.084 ± 0.002
I-3	β_3	0.8	0.83 ± 0.07	0.82 ± 0.08	0.90 ± 0.04
	γ_3 (w_{tn})	0.15	0.137 ± 0.029	0.141 ± 0.030	0.109 ± 0.011
	σ_3 (w_{tn})	0.1	0.099 ± 0.010	0.093 ± 0.011	0.095 ± 0.012
I-6	β_6	1.2	1.19 ± 0.06	1.20 ± 0.05	1.20 ± 0.03
	γ_6 (w_{tn})	-0.1	-0.094 ± 0.026	-0.101 ± 0.024	-0.101 ± 0.005
	σ_6 (w_{tn})	0.1	0.096 ± 0.011	0.096 ± 0.012	0.094 ± 0.008

[a] "1.11 ± 0.04" means that the mean and standard deviation of the learned values of β_0 from the ten trials are 1.11 and 0.04, respectively. The same explanation applies to the results of the parameter learning presented in Tables 2.2 through 2.4.

Table 2.3 Prescribed and learned parameters of ΔX and X_0 in Example 1

		ΔX		X_0	
		Mean value (w_{tn})	COV (%)	Mean value (w_{tn})	COV (%)
Prescribed values		0.04	50	0.3	30
Learned values	Scheme 1	0.038 ± 0.002	46.4 ± 7.4	0.305 ± 0.011	29.9 ± 2.8
	Scheme 2	0.038 ± 0.002	47.8 ± 8.0	0.302 ± 0.013	30.7 ± 3.5
	Scheme 3	0.032 ± 0.003	40.7 ± 10.3	0.320 ± 0.014	26.2 ± 2.4

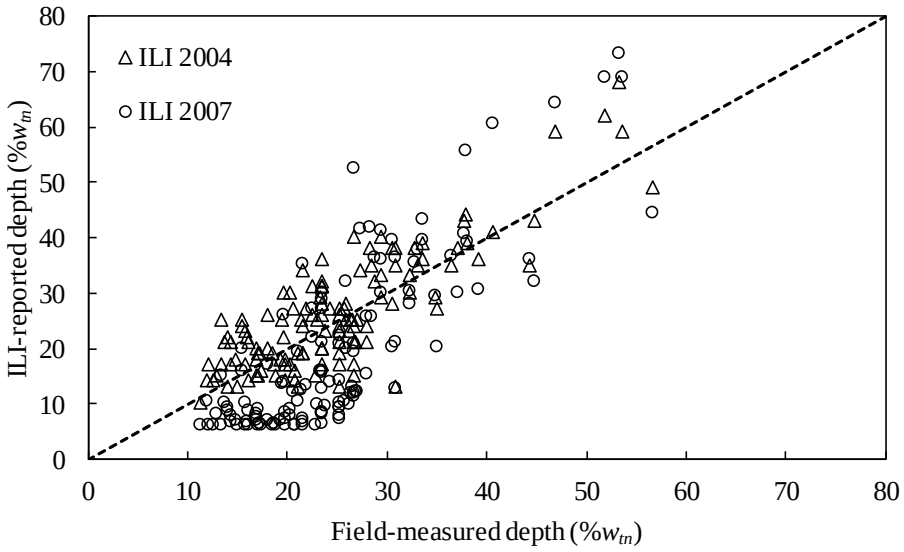
As defined before, Dataset 2 only contains ILI data. In practice, the population of corrosion defects on a given pipeline segment is generally large. Therefore, the sample size of Dataset 2 for a given pipeline segment is not of great concern. On the other hand, Dataset 1 includes field measurements. Since the cost of excavating a pipe joint is high (typically $200,000), the sample size of Dataset 1 may be limited. Therefore, it is valuable to investigate the sensitivity of the learning results for β_i, γ_i and σ_i (i = 0, 3, 6) to the sample size of Dataset 1. All else being equal, two additional sample sizes of Dataset 1 are considered, namely 50 and 150, respectively. The data simulation and parameter learning of β_i, γ_i and σ_i (i = 0, 3, 6) are repeated for these two sample sizes. Table 2.4 compares the results associated with the three different sample sizes, i.e. 50, 100 and 150. The results indicate slight differences in the parameters learned based on the sample sizes of 100 and 150. However, the accuracy of the parameter learning for the sample size of 50 is relatively poor. Therefore, to achieve relatively accurate quantification of the measurement errors in ILI data, it is recommended that the sample size of Dataset 1 be around 100 or greater.

Table 2.4 Prescribed and learned parameters for the ILI measurement errors under different sample sizes in Dataset 1 for Example 1

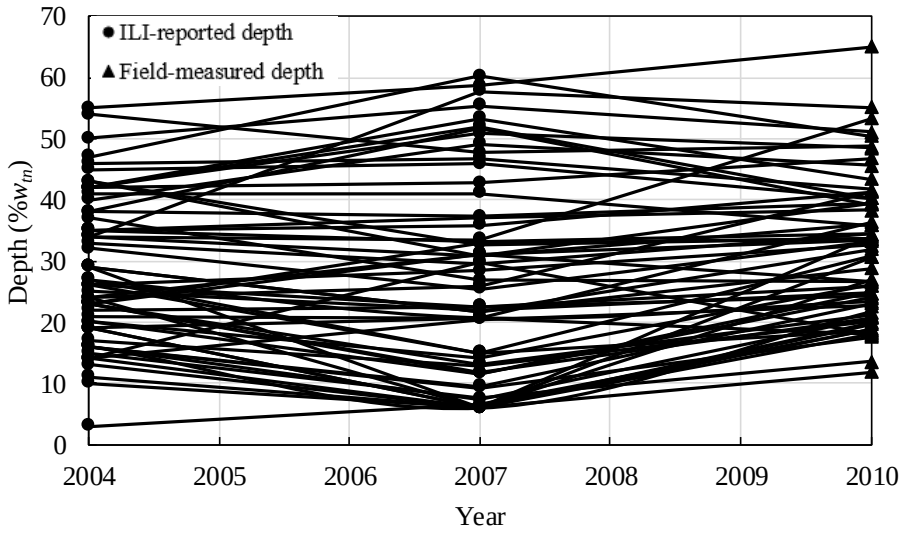
		Prescribed values	Values from parameter learning		
			100 samples	50 samples	150 samples
I-0	β_0	1.1	1.11 ± 0.05	1.15 ± 0.11	1.09 ± 0.04
	γ_0 (w_{tn})	-0.05	-0.059 ± 0.015	-0.075 ± 0.045	-0.045 ± 0.017
	σ_0 (w_{tn})	0.07	0.079 ± 0.009	0.099 ± 0.010	0.079 ± 0.006
I-3	β_3	0.8	0.83 ± 0.07	0.87 ± 0.08	0.83 ± 0.07
	γ_3 (w_{tn})	0.15	0.137 ± 0.029	0.125 ± 0.030	0.139 ± 0.026
	σ_3 (w_{tn})	0.1	0.099 ± 0.010	0.114 ± 0.010	0.102 ± 0.007
I-6	β_6	1.2	1.19 ± 0.06	1.07 ± 0.10	1.19 ± 0.05
	γ_6 (w_{tn})	-0.1	-0.094 ± 0.026	-0.054 ± 0.040	-0.093 ± 0.020
	σ_6 (w_{tn})	0.1	0.096 ± 0.011	0.109 ± 0.014	0.099 ± 0.009

2.4.2 Example 2: real corrosion data

In this section, the DBN is employed to quantify the growth of the defect depth with real ILI and field measurement data from a pipeline that was constructed in 1972 and is currently in service in Alberta, Canada. The pipeline has a nominal wall thickness of 5.56 mm and outside diameter of 508 mm, and is made of API 5L Grade X52 steel with a nominal yield strength of 359 MPa and a nominal operating pressure of 5.66 MPa. The pipeline was inspected by two different ILI tools in 2004 and 2007, respectively. A set of corroded pipe joints were excavated and recoated between 2002 and 2004, and the sizes of 128 corrosion defects on the excavated pipe joints were measured on the site. Therefore, the field-measured depths before 2004 and ILI-reported depths in 2004 and 2007 of the 128 static corrosion defects constitute Dataset 1 (Fig. 2.6(a)). Dataset 2 (Fig. 2.6(b)) contains the depths of 62 defects reported by the ILIs in 2004 and 2007, respectively. The defects in Dataset 2 were repaired in 2010, and their depths were measured on site during the repair. Dataset 2 is used to estimate the annual growth of the defect depth, and the field-measured depth in 2010 are used to validate the predictive accuracy of the growth model. Figure 2.6(b) also depicts the growth paths for the 62 defects by linking the ILI-reported depths in 2004 and 2007, and field-measured depth in 2010 belonging to the same defect. Note that corrosion growth is a monotonically increasing process. However, due to the measurement errors on the ILI-reported depths, the growth paths indicated in Fig. 2.6(b) do not necessarily increase monotonically over time. The two datasets have been used in a previous study (Al-Amin et al., 2012) employing Bayesian models and the MCMC technique to quantify the measurement errors associated with ILI tools and growth of defect depth.



(a) Dataset 1 used in Example 2



(b) Dataset 2 and field-measured depths used in Example 2

Figure 2.6 Data used for model development and validation for Example 2

A DBN-based growth model (Fig. 2.7) is developed and includes seven time-slices (i.e. years 2004 through 2010). The random variables are discretized following the discretization scheme 1 as presented in Table 2.1. Dataset 1 is used to learn the PMFs for β_{2004} and β_{2007} (Fig. 2.8(a)), γ'_{2004} and γ'_{2007} (Fig. 2.8(b)), associated with the ILI tools used

in 2004 and 2007, respectively, whereas Dataset 2 is used to learn the PMFs for ΔX (Fig. 2.8(c)) and X_{2004} (Fig. 2.8(d)). The mean and COV of ΔX corresponding to the learned PMF are $0.00618 w_{tn}$ and 178%, respectively. The mean and COV of X_{2004} corresponding to the learned PMF are $0.291 w_{tn}$ and 41%, respectively. Figure 2.9 depicts the average negative log-likelihood per case for Datasets 1 and 2 as a function of the number of iterations in the EM algorithm, which indicates that convergence is achieved typically after about 20 iterations. The learned values of β_{2004}, γ_{2004}, σ_{2004}, β_{2007}, γ_{2007} and σ_{2007} are compared with the results reported by Al-Amin et al. (2012) in Table 2.5. The results obtained in the present study and Al-Amin et al. (2012) are generally consistent; the difference between the results may be explained by the following two reasons. First, β_{2004}, γ_{2004}, σ_{2004}, β_{2007}, γ_{2007} and σ_{2007} are continuous random variables in the Bayesian model employed in Al-Amin et al. (2012), whereas β_{2004}, γ'_{2004}, β_{2007} and γ'_{2007} are discretized in the present study. Second, informative prior distributions were assigned to β_{2004} and β_{2007} in Al-Amin et al. (2012), whereas the parameter learning in the DBN is performed based on non-informative prior distributions.

Table 2.5 Comparison of the values of β_i, γ_i, σ_i obtained in the present study and Al-Amin et al. (2012) in Example 2

Parameter	Learned values in the present study	Posterior mean values reported in Al-Amin et al. (2012)
β_{2004}	0.85	0.97
$\gamma_{2004}(w_{tn})$	0.056	0.020
$\sigma_{2004}(w_{tn})$	0.063	0.060
β_{2007}	1.36	1.40
$\gamma_{2007}(w_{tn})$	-0.139	-0.153
$\sigma_{2007}(w_{tn})$	0.067	0.091

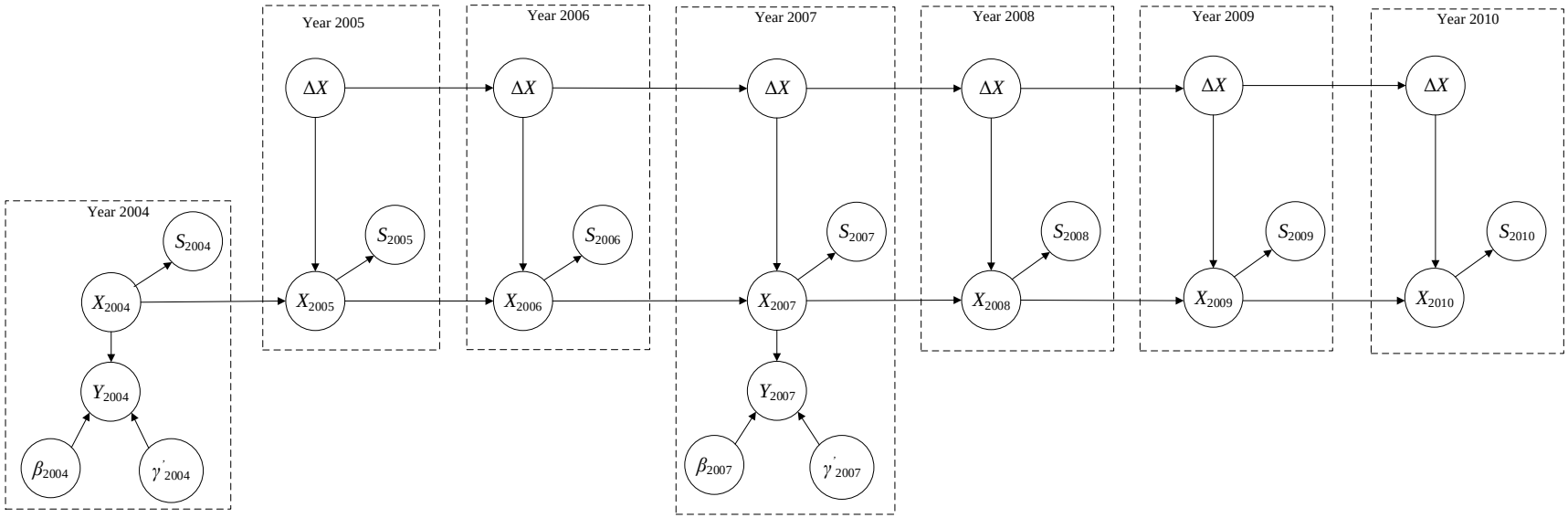
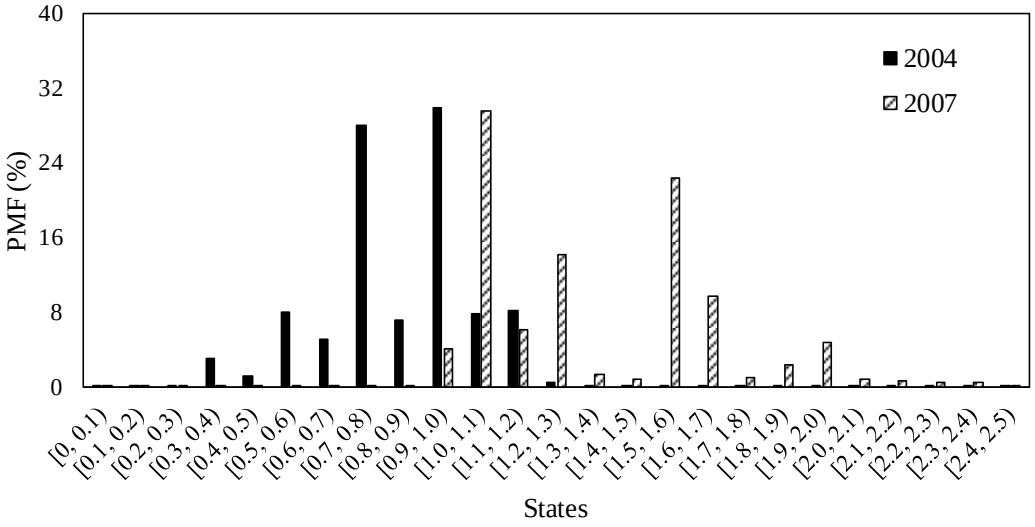
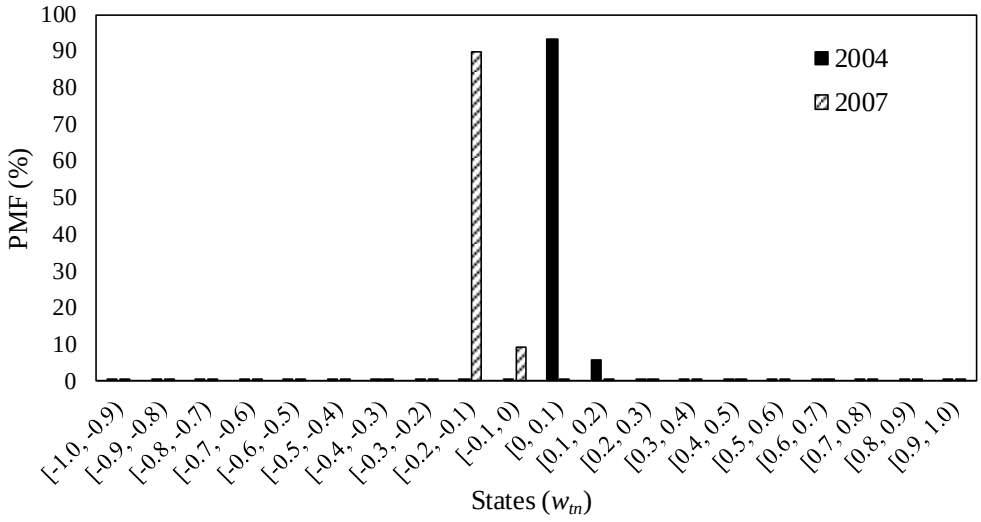


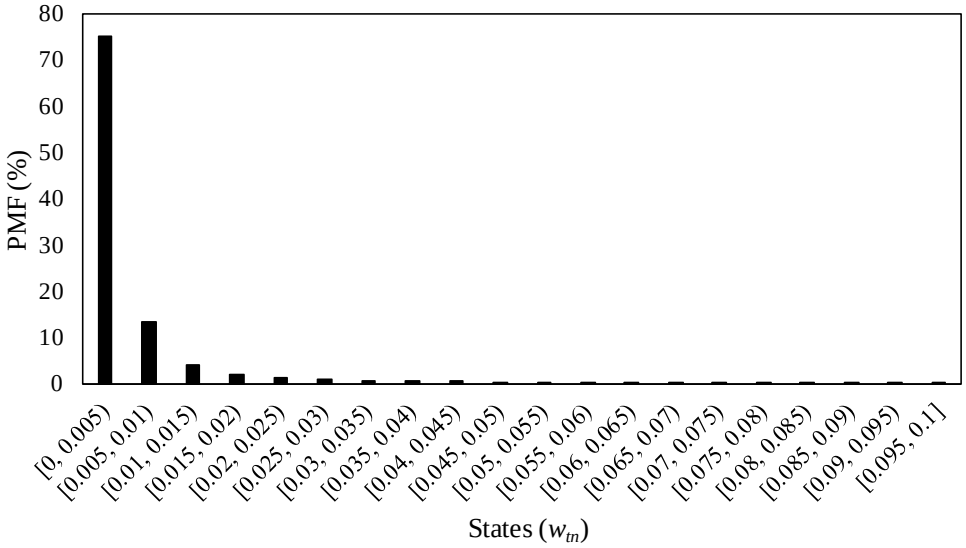
Figure 2.7 The DBN growth model developed for Example 2



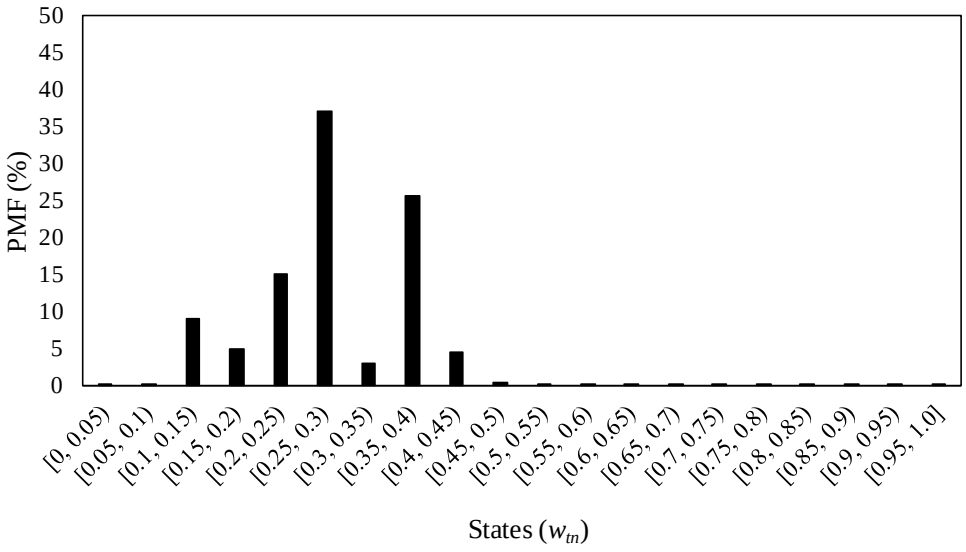
(a) Learned PMFs of β_{2004} and β_{2007}



(b) Learned PMFs of γ'_{2004} and γ'_{2007}



(c) Learned PMF of ΔX



(d) Learned PMF of X_{2004}

Figure 2.8 Results of parameter learning for Example 2

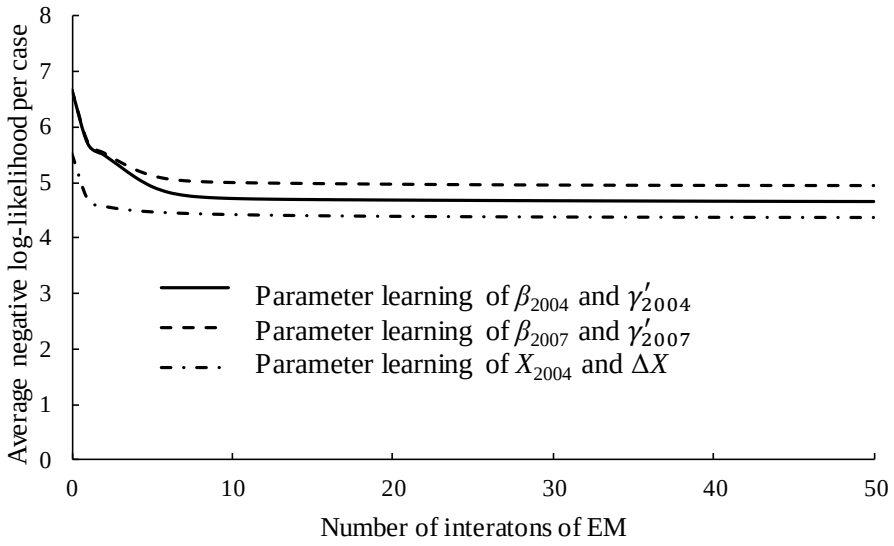
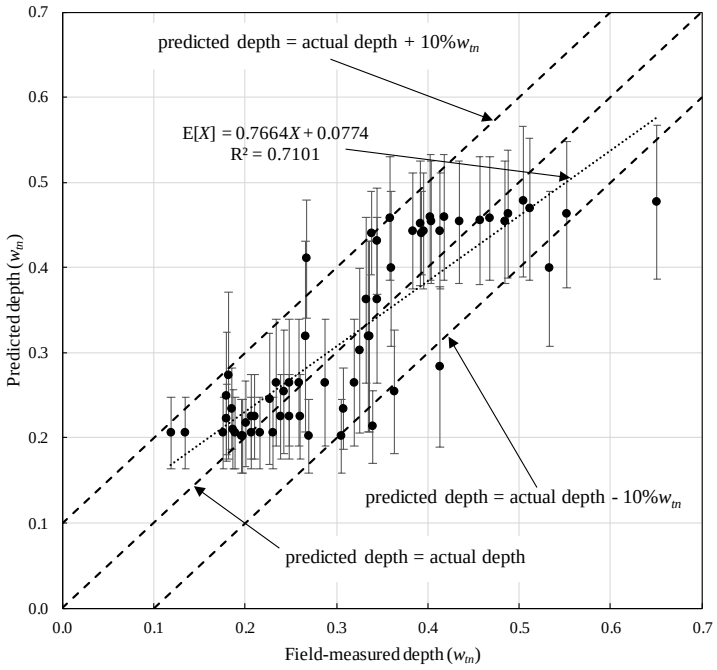


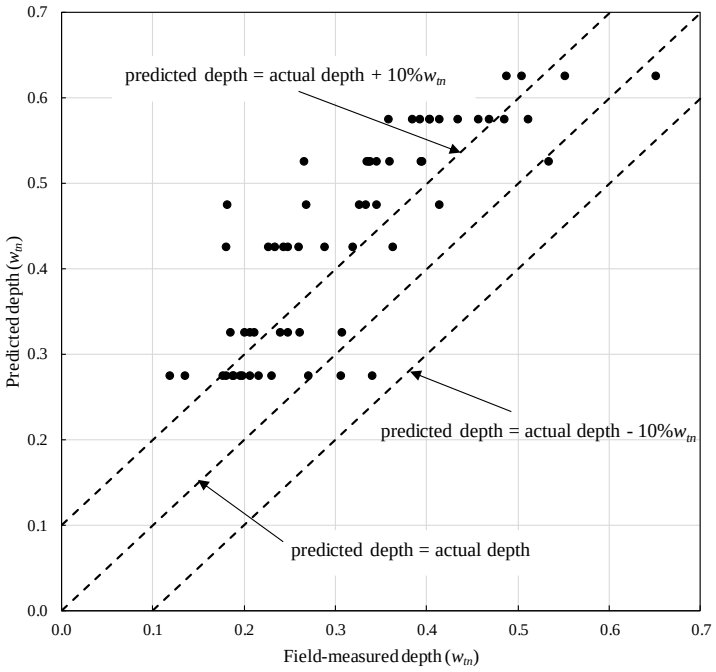
Figure 2.9 Convergence curve of parameter learning for Example 2

The developed DBN is used to predict the growth of the defect depth. The learned PMF for ΔX from Dataset 2 is considered representative of the common features shared by all the defects in the dataset and therefore used as the prior distribution of the growth rate for all the defects in Dataset 2. To predict the growth path of a specific defect in Dataset 2, the ILI-reported depths in 2004 and 2007 are used to instantiate the corresponding nodes to evaluate the posterior distributions of X_{2004}, X_{2007} and ΔX for the defect. The defect depth in years after 2007 can then be predicted from X_{2007} and ΔX based on the linear growth model. Figure 2.10(a) compares the posterior mean depths and corresponding field-measured depths in 2010 for the 62 defects in Dataset 2. One-standard-deviation intervals of model-predicted depths are also included to characterize the uncertainty associated with the predictions. The results show that the majority (i.e. 85%) of the predictions lie in the region bounded by the lines representing the prediction errors of $\pm 10\% w_{tn}$. The regression line between the mean predicted depth and field-measured depth is plotted in Fig. 2.10(a), where E[X], X and R^2 denote the mean predicted depth, field-measured depth and coefficient of determination of the regression line between E[X] and X, respectively. The value of R^2, i.e. 0.701, indicates a relatively strong correlation between E[X] and X. This good predictive accuracy validates the modeling and parameter learning of the DBN growth model. Figure 2.10(a) shows that the depths of a few deep defects are under-

predicted by the DBN model, which can lead to non-conservatism in the corrosion mitigation decision-making in practice. To address this issue, the 95-percentiles of the posterior defect depths (as opposed to the posterior mean depths) can be adopted as the predicted defect depths, as illustrated in Fig. 2.10(b). The figure indicates that the predictive accuracy for the deep defects is improved, albeit at a price of increased conservatism in the overall prediction.



(a) Mean value of predicted depth vs. field-measured depth



(b) 95-percentile of predicted depth vs. field-measured depth

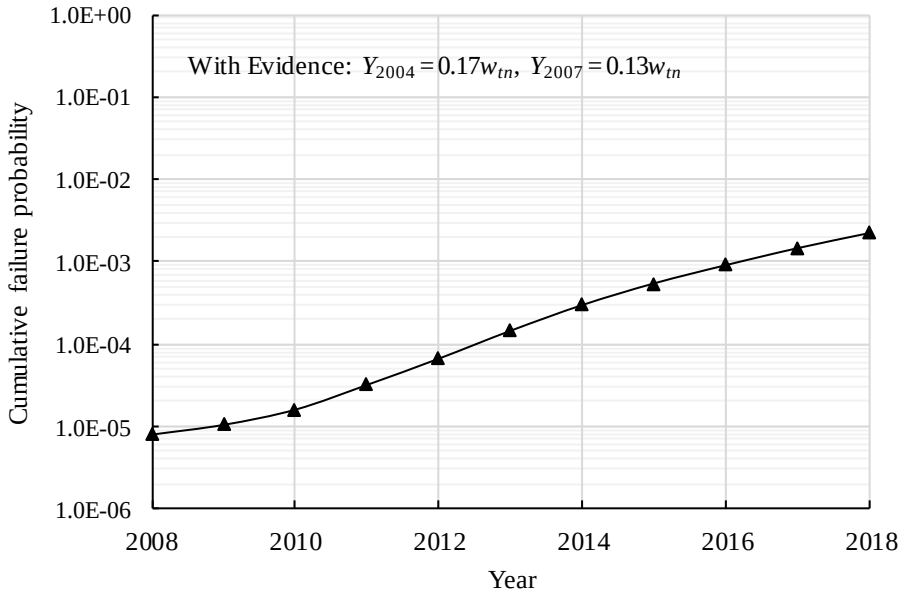
Figure 2.10 Predicted and actual defect depths in 2010 for Example 2

Table 2.6 Probabilistic characteristics of random variables of the pipeline (Zhou, 2010) in Example 2

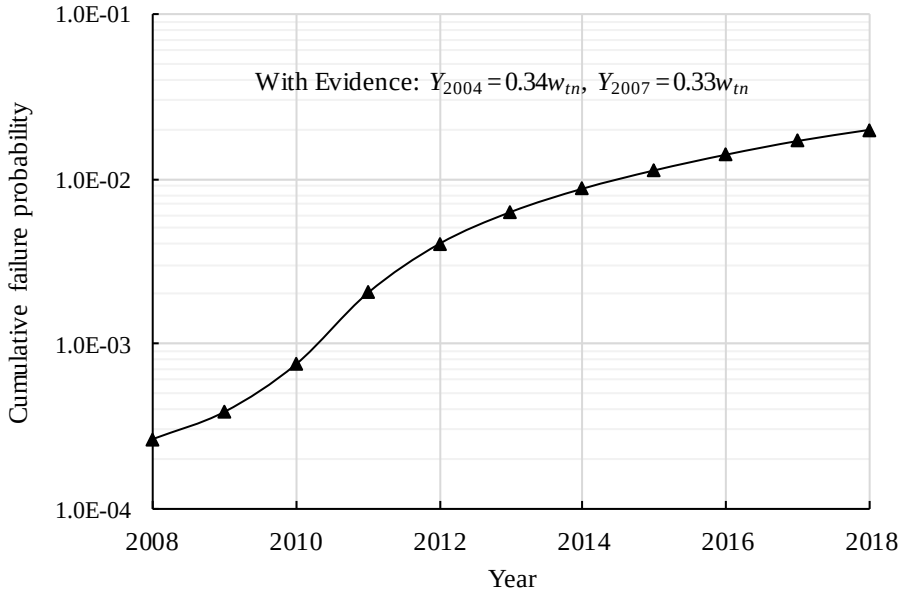
Variable	Distribution	Mean	COV (%)
l (mm)	Normal	75	20
D (mm)	Deterministic	508	-
w_t (mm)	Normal	5.56	1.5
σ_y (MPa)	Lognormal	395	3.5
o_p (MPa)	Gumbel	5.66	3.0
κ	Gumbel	1.2	20

The developed DBN model is further used to predict the time-dependent failure probability of the pipeline at a defect given the ILI-reported depths as evidence. The failure probability is output through the node S_i in the DBN. While modeling the growth of the defect length can be handled by the DBN in the same way as the defect depth, the growth of the defect length is ignored for simplicity in this example. The probabilistic characteristics of the variables that are used to develop the CPT for S_i are summarized in Table 2.6. Using ILI

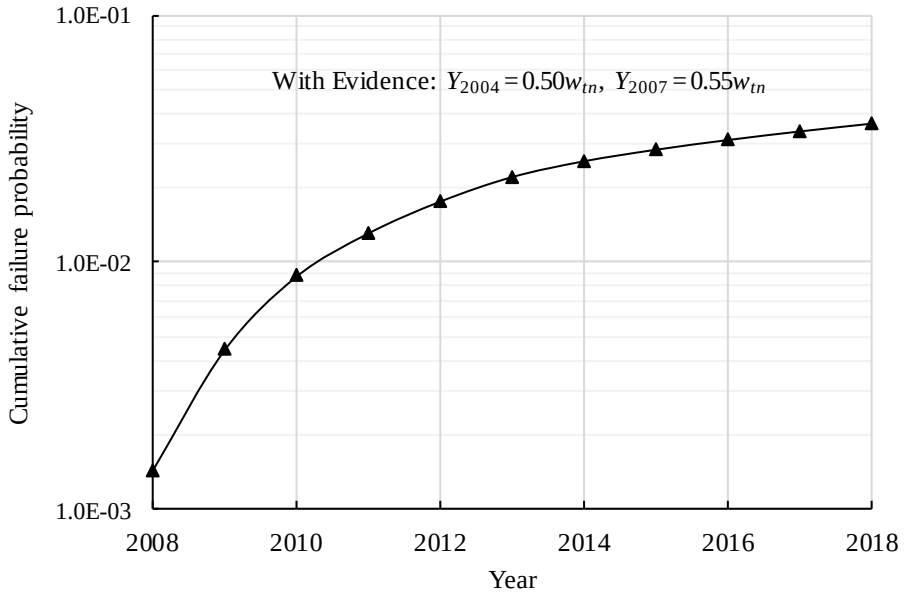
data to instantiate the nodes Y_{2004} and Y_{2007}, the cumulative failure probabilities of three representative defects in Dataset 2 evaluated by the DBN are shown in Figs. 2.11(a), 2.11(b) and 2.11(c), respectively.



(a) Failure probabilities over 2008 through 2018 for Defect 1



(b) Failure probabilities over 2008 through 2018 for Defect 2



(c) Failure probabilities over 2008 through 2018 for Defect 3

Figure 2.11 Failure probabilities calculated by the DBN model for three representative defects in Example 2

2.5 Conclusions

A DBN model is developed to quantify the growth of depths of corrosion defects on pipelines and predict the time-dependent failure probabilities of the pipeline at individual defects. The defect growth is assumed to be linear in time with a constant but uncertain growth rate. The EM algorithm in the context of the parameter learning technique is employed to evaluate parameters in the DBN based on the ILI-reported and field-measured defect depths. The failure probability of the defect at each time-slice is evaluated in the DBN to facilitate the updating of the failure probability based on the ILI data. The effectiveness of the parameter learning for the DBN model is validated by the numerical example. The application of the proposed model to real ILI and field-measured data shows that the predicted defect depths in general agree well with the field-measured depths, and the time-dependent failure probability can be evaluated effectively and efficiently by using ILI data to instantiate corresponding nodes in the model. The developed model is advantageous in the following three respects. First, the defect growth modeling,

quantification of measurement errors associated with ILI tools and failure probability evaluation are integrated into a single model. Second, the parameter learning technique allows parameters of the DBN model to be quantified in an automated and objective manner. Third, the efficient inference algorithm of DBN enables the model updating to be completed highly efficiently. These advantages make the model more accessible to non-specialists in Bayesian data analysis and facilitate the reliability-based corrosion management of oil and gas pipelines.

Sensitivity analyses suggest that the sample size of Dataset 1 should be around 100 or greater to ensure the accuracy of the parameter learning results with respect to the ILI measurement errors. This condition may not be easily met if the pipeline contains a small number of critical defects that have been excavated for mitigation. Analysis results for Example 2 indicate that posterior mean depths of the DBN model tend to under-predict the depths of some deep defects. This issue can be addressed by using the 95-percentile of the posterior defect depth as the predicted depth, although with increased conservatism in the overall prediction. Finally, a simple linear corrosion growth model is adopted in the present study. More sophisticated growth models such as the power-law and gamma process-based models can be incorporated into the DBN without much difficulty.

References

Ahammed, M. (1998). Probabilistic estimation of remaining life of a pipeline in the presence of active corrosion defects. *International Journal of Pressure Vessels and Piping*, 75, 321-329.

Al-Amin, M., and Zhou, W. (2014). Evaluating the system reliability of corroding pipelines based on inspection data. *Structure and Infrastructure Engineering*, 10(9), 1161-1175.

Al-Amin, M., Zhou, W., Zhang, S., Kariyawasam, S., and Wang, H. (2012). Bayesian model for calibration of ILI tools. In: *Proceedings of the 9th International Pipeline Conference* (pp. 201-208), Calgary, Alberta, Canada.

Amirat, A, Mohamed-Chateauneuf, A., and Chaoui, K. (2010). Reliability assessment of underground pipelines under the combined effect of active corrosion and residual stress. *International Journal of Pressure Vessels and Piping*, 83, 107-117.

Canadian Energy Pipeline Association (CEPA). (2015). *Committed to safety, committed to Canadians: 2015 pipeline performance report*. Calgary, Alberta: Canadian Energy Pipeline Association.

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 (Statistical Methodology)*, 39(1), 1-38.

Fuller, W.A. (1987). *Measurement error models*. John Wiley and Sons, Inc., New York, NY, USA

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In: Jordan M. (Eds.), *Learning in graphical models* (pp. 301-354), Springer, Dordrecht.

Hong, H-P. (1999) Application of the stochastic process to pitting corrosion. *Corrosion*, 55(1):10-16.

Kariyawasam, S., and Peterson, W. (2010). Effective improvements to reliability based corrosion management. In: *Proceedings of the 8th International Pipeline Conference* (pp. 603-615), Calgary, Alberta, Canada.

Kiefner, J.F., and Veith P.H. (1989). *A modified criterion for evaluating the remaining strength of corroded pipe*. United States, (Report No. PR-3-805). Battelle Columbus Div., OH, USA.

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

Luque, J., and Straub, D. (2016). Reliability analysis and updating of deterioration systems with dynamic Bayesian networks. *Structural Safety*, 62, 34-46.

Mahadevan, S., Zhang, R., and Smith, N. (2001). Bayesian networks for system reliability reassessment. *Structural Safety*, 23, 231-251.

Marquez, D., Martin, N., and Fenton, N. (2010). Improved reliability modeling using Bayesian networks and dynamic discretization, *Reliability Engineering and System Safety*, 95(4), 412-425.

Masegosa, A., Feelders, A., and van der Gaag, L. (2016). Learning from incomplete data in Bayesian networks with qualitative influences. *International Journal of Approximate Reasoning*, 69, 18-34.

Murphy, K. (2002). *Dynamic Bayesian networks: representation, inference and learning* (Doctoral dissertation). Department of Computer Science, UC Berkeley, California.

Nessim, M., Dawson, J., Mora, R., and Hassanein, S. (2008). Obtaining corrosion growth rates from repeat in-line inspection runs and dealing with the measurement uncertainties. In: *Proceedings of the 7th International Pipeline Conference* (pp. 593-600), Calgary, Alberta, Canada.

Nielsen, T., and Jensen, F. (2009). *Bayesian networks and decision graphs*. New York, NY: Springer Science and Business Media.

Pandey, M.D., Yuan, X.-X., and van Noortwijk, J.M. (2009). The influence of temporal uncertainty of deterioration on life-cycle management of structures. *Structure and Infrastructure Engineering*, 5(2), 145-156.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Rafiq, M, Chryssanthopoulos, M., and Sathananthan, S. (2010). Bridge condition modelling and prediction using dynamic Bayesian belief networks. *Structure and Infrastructure Engineering*, 11(1), 38-50.

Spiegelhalter, D., Dawid, P., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8(3), 219-247.

Straub, D. (2009). Stochastic modeling of deterioration processes through dynamic Bayesian networks. *Journal of Engineering Mechanics*, 135(10), 1089-1099.

Straub, D., and Der Kiureghian, A. (2010). Bayesian network enhanced with structural reliability methods: Methodology. *Journal of Engineering Mechanics*, 136(10), 1248-1258.

Valor, A., Caleyo, F., Alfonso, L., Rivas, D., and Hallen, J. M. (2007). Stochastic modeling of pitting corrosion: a new model for initiation and growth of multiple corrosion pits. *Corrosion Science*, 49(2), 559-579.

Zhou, W. (2010). System reliability of corroding pipelines. *International Journal of Pressure Vessels and Piping*, 87(10), 587-595.

Zhou, W., Xiang, W., and Hong H-P. (2017). Sensitivity of system reliability of corroding pipelines to modeling of stochastic growth of corrosion defects. *Reliability Engineering and System Safety*, 167,428-438.

Zhou, Y., Fenton, N., and Zhu, C. (2016). An empirical study of Bayesian network parameter learning with monotonic influence constraints. *Decision Support System*, 87, 69-79.

Zweig, G. (1996). A forward-backward algorithm for inference in Bayesian networks and an empirical comparison with HMMs (Master's thesis). Department of Computer Science, UC Berkeley, California.

3 Bayesian network model for predicting probability of third-party damage to underground pipelines and learning model parameters from incomplete datasets

3.1 Introduction

The historical pipeline incident data indicate that the mechanical damage from excavations by third parties is one of the leading threats to the structural integrity of buried pipelines (Lam and Zhou, 2016; EGIG, 2018). A third party is neither a pipeline operator nor a contractor hired by the operator to service the pipeline; in other words, a third party is an individual or organization unrelated to pipeline assets. About 26% of the pipe-related incidents on onshore gas transmission pipelines in the United States resulted from third-party excavations between 2002 and 2013, almost equal to the number of incidents caused by external and internal corrosions combined (Lam and Zhou, 2016); the third-party damage is the leading threat to gas transmission pipelines in Europe and accounted for 28.4% of all the gas pipeline incidents between 1970 and 2016 (EGIG, 2018). Therefore, the pipeline industry and regulatory agencies are devoting significant efforts to preventing pipelines from being damaged by third-party excavations. Commonly used preventative measures for the third-party damage (TPD) include, for example, the one-call system (third parties notify the pipeline operators through one-call centers before excavations), warning signs along the pipeline right-of-way (ROW), regular patrol of ROW, and supervision of excavations by personnel of pipeline operators. Protective measures for TPD include the burial depth of pipelines and physical protection such as concrete slabs buried above the pipeline alignment. The Pipeline and Hazardous Materials Safety Administration (PHMSA) of the US Department of Transportation and common ground alliance (CGA) have been using the damage information reporting tool (DIRT) to collect data regarding the damage of underground utilities including pipelines to facilitate the analysis of the effectiveness of preventative and protective measures against TPD.

The reliability-based pipeline integrity management program with respect to TPD is being increasingly adopted by pipeline operators to deal with uncertainties associated with the occurrence of TPD events (Koduru and Nessim, 2017). A key task in such a program is to estimate the hit rate due to third-party excavations, which is the product of the rate of

excavation activities (typically expressed in terms of per year per kilometer of pipeline) and probability of hit given a third-party activity (Chen and Nessim, 1999; Chen et al., 2006; Koduru and Lu, 2016; Lu and Stephen, 2016). The activity rate is estimated from the observed third-party activities occurred in the vicinity of the pipeline alignment. A fault tree model developed by Chen and Nessim (1999) has been widely employed by the pipeline industry to estimate the probability of hit. The fault tree models a pipeline being hit by a third-party excavation as the result of failures of all the preventative and protective measures such as the third party failing to notify the pipeline operator before the excavation, excavation undetected by the ROW patrol and excavation depth exceeding the burial depth of the pipeline. Various improvements of the original fault tree developed by Chen and Nessim (1999) have been proposed since its development. Chen et al. (2006) enhanced the fault tree model by taking into account a broader range of preventative and protective measures typically used in the pipeline industry. Lu and Stephens (2016) classified third-party activities into authorized activities (AAs) and unauthorized activities (UAs) based on whether or not the pipeline operator's permission has been obtained prior to the start of the excavation. They then developed a hierarchical fault tree model to evaluate the probability of hit as the weighted sum of the probabilities of hit due to authorized and unauthorized activities.

The failures of individual preventative and protective measures are the basic events of the fault tree models reported in the literature (Chen and Nessim, 1999; Lu and Stephens, 2016). Chen and Nessim (1999) carried out an industry-wide survey to estimate probabilities of basic events, generally as functions of relevant pipeline attributes (e.g. patrol frequency, pipeline burial depth, dig notification response time). In the practice of TPD management over the past few decades, pipeline operators have collected a substantial amount of TPD related data such as the individual TPD activities including the information of pipeline attributes, prevention measures and consequences of the TPD activities, and it is highly desirable to use the collected data to estimate the probabilities of basic events. However, the nature of the fault tree analysis, i.e. top-down deduction, and the fact that the collected TPD data generally contain missing information, i.e. the so-called incomplete data, present significant challenges to the probability updating within the fault tree framework.

Fault tree models can be straightforwardly mapped to corresponding Bayesian Networks (BNs) (Bobbio et al., 2001; Khakzad et al., 2011), which are well suited for inference and probability updating based on observed data. However, there is limited literature on the use of BNs to evaluate the probability of hit. Koduru and Lu (2016) developed a BN model to evaluate the probability of hit based on the fault tree model reported in Chen et al. (2006). They used the information in the DIRT report to evaluate probabilities of basic events in the BN model. However, since participants of the DIRT program report the TPD data only if third-party incidents are detected, the TPD data in the DIRT report are conditional on the occurrence of pipelines being hit. To estimate the unconditional probability of a basic event, Koduru and Lu (2016) manually adjusted its probability iteratively until the probability of the event conditional on a hit equals the probability estimated from the DIRT report. Such an approach for evaluating the probability of the basic event is highly inefficient. Furthermore, it is very difficult, if possible at all, to estimate the probabilities of multiple basic events simultaneously using this approach.

Extensive studies in the area of artificial intelligence have demonstrated that the parameter learning technique associated with BNs provides an automated and objective means to estimate a large number of parameters of BNs from observed data, particularly incomplete data (Heckerman, 1998; Liao and Ji, 2010; Masegosa et al., 2016; Zhou et al., 2016). The TPD-related data (i.e. individual cases of third-party activities) collected by the pipeline industry generally contain incomplete information for estimating the failure probabilities of preventative and protective measures against third-party excavations. The present study considers two typical incomplete datasets that consist of individual third-party activities and proposes to employ the parameter learning technique of BN to learn the probabilities mentioned above. To this end, a BN model for evaluating the probability of hit given a third-party activity is first developed based on the fault tree commonly used by the pipeline industry (Chen and Nessim, 1999; Lu and Stephens, 2016), whereby the probabilities to be learned are converted to the parameters of the BN model. The Expectation-Maximization (EM) algorithm in the context of parameter learning is then employed to learn the parameters of the BN from two TPD datasets.

The remainder of this chapter is organized as follows. Section 3.2 describes the fault tree model widely used to evaluate the probability of hit given a third-party activity. Section 3.3 presents the development of the BN based on the fault tree described in Section 3.2, incomplete datasets provided by the pipeline industry, and EM algorithm for the parameter learning. Section 3.4 demonstrates the effectiveness of the parameter learning through a numerical example involving simulated TPD data and an application using real-world TPD datasets, followed by conclusions in Section 3.5.

3.2 Fault tree model for evaluating the probability of hit

A fault tree is a top-down deductive tool to evaluate the probability of failure of a system that is attributed to failures of multiple components of the system (Mearns, 1965). In a fault tree, the system failure is the top event; events that result from occurrences of other events are called intermediate events, and events that cannot be broken down into other events are called basic events. The relationship between higher-level and lower-level events is characterized by Boolean logic, i.e. the "or" and "and" gates. The higher-level and lower-level events associated with a gate are called the output and input events of the gate, respectively. For the "or" gate, the output event occurs if any of the input events occurs; for the "and" gate, the output event occurs only if all of the input events occur. Once the probabilities of basic events are input into the fault tree, the probability of the top event can be evaluated by transmitting the probabilities through the gates using the following two rules,

$$p_{\text{and}} = \prod_{i=1}^{n} p_i \tag{3.1}$$

$$p_{\text{or}} = 1 - \prod_{i=1}^{n} (1 - p_i) \tag{3.2}$$

where p_{and} and p_{or} are the probabilities of the output event of the "and" and "or" gates, respectively; p_i is the probability of the i-th input event of the gate, and n is the total number of input events of the gate.

A fault tree for calculating the probability of hit (P_h) given a third-party activity (Fig. 3.1) is adapted from the fault tree model developed by Chen and Nessim (1999). The fault tree model is developed through a top-down process as follows. The top event (T_0) represents

a pipeline being hit by a third-party activity, which is connected to the ineffectiveness of all the preventative measures (E_7) and the only protective measure considered in the fault tree, i.e. excavation depth exceeding the cover depth of the pipeline (B_9). As E_7 and B_9 must both occur to result in the pipeline hit, they are connected to T_0 via the "and" gate. The event E_7 is linked via the "or" gate to event B_8 (accidental hit due to insufficient excavation practice for the correctly located pipeline) and event E_6 (the pipeline incorrectly located and marked by the operator). The event E_6 is linked via the "or" gate to event B_7 (the operator aware of the activity but failing to locate the pipeline correctly) and event E_5 (the operator unaware of the activity). The "and" gate links E_5 to event E_4 (the excavation being unauthorized) and event B_6 (the UA undetected by the ROW patrol). A UA results either from the operator not notified by the one-call system (E_3), or from the third-party starting the excavation prior to the operator's response to the one-call notification (B_5). The "or" gate links E_3 to B_4 (the one-call center failing to notify the operator when contacted by the third-party) and E_2 (the one-call center not contacted by the third-party). The "and" gate links E_2 to E_1 (one-call not made while the excavator on ROW) and B_1 (one-call not made before the operator mobilizing to ROW), and finally E_1 is linked via the "or" gate to B_2 (ROW signs not recognized by the excavator) and B_3 (ROW signs ignored by the excavator). It follows that the fault tree model contains nine basic events, i.e. B_1 through B_9, and seven intermediate events, i.e. E_1 through E_7.

Once the probabilities B_1 through B_9 are input into the fault tree, the probabilities of E_1 through E_7 as well as the top event T_0 are evaluated by transmitting probabilities based on the rules given by Eqs. (3.1) and (3.2). The probabilities of basic events are defined as functions of pipeline attributes such as the burial depth, ROW patrol frequency, and public awareness of the one-call system (Chen and Nessim, 1999) as summarized in Table 3.1. The values of pipeline attributes denoted by A_1 through A_9 are summarized in Table 3.2.

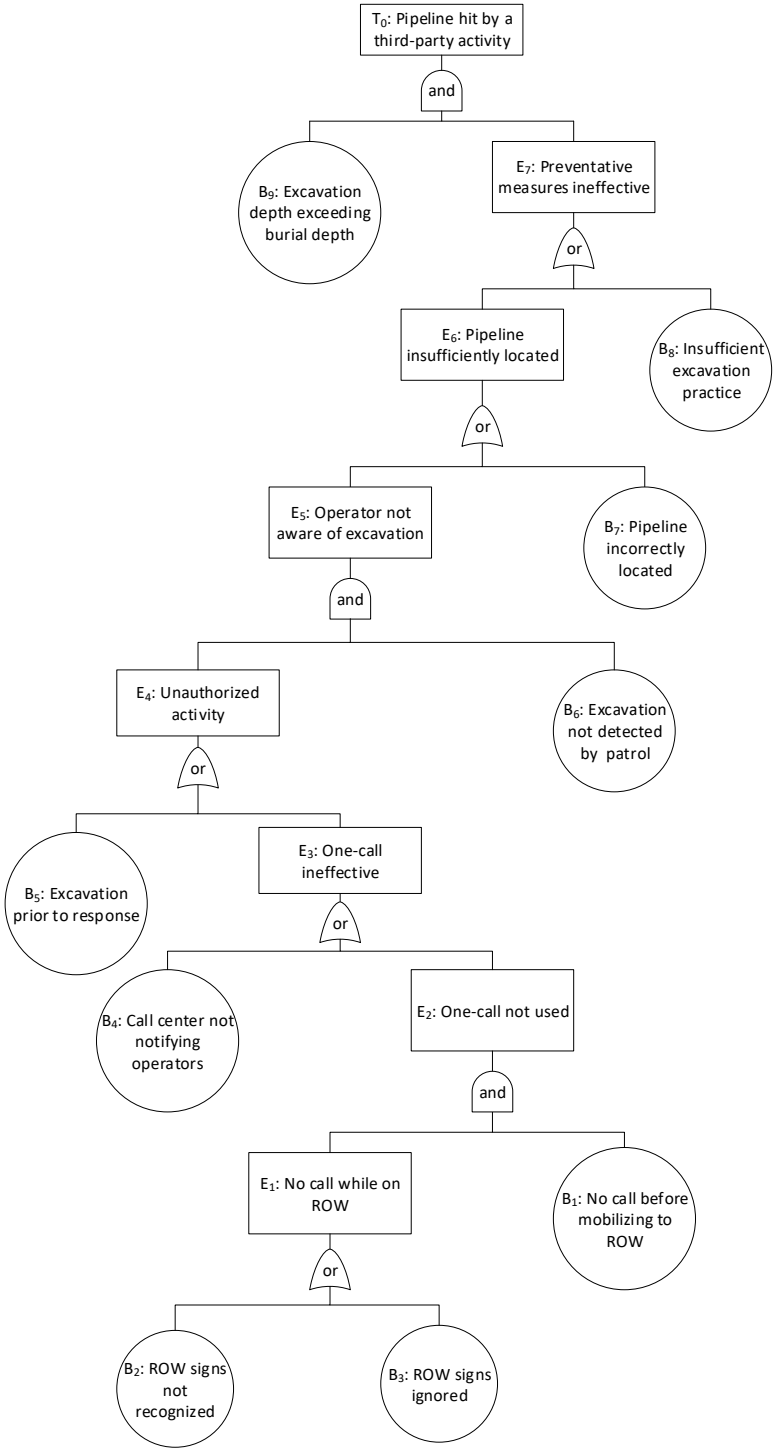


Figure 3.1 Fault tree model to evaluate P_h given a third-party activity

Table 3.1 Description of the dependence of basic events on pipeline attributes

Basic event	Pipeline attributes influencing the probability of the basic event
B_1: No call before mobilizing to ROW	A_1: Dig notification requirement; A_2: Public awareness level of one-call; A_4: One-call type
B_2: ROW signs not recognized	A_2: Public awareness level of one-call; A_3: ROW spacing
B_3: ROW signs ignored	A_2: Public awareness of one-call; A_4: One-call type
B_4: Call center not notifying operators	A_4: One-call type
B_5: Excavation prior to response	A_5: Response time to dig notification
B_6: Excavation not detected by patrol	A_6: Patrol frequency
B_7: Pipeline incorrectly located	A_7: Locating method
B_8: Insufficient excavation practice	A_8: Response method to notification
B_9: Excavation depth exceeding burial depth	A_9: Burial depth

Table 3.2 Values of pipeline attributes

Pipeline attributes	Values
A_1: Dig notification requirement	Not required; Required but not enforced; Required and enforced
A_2: Public awareness level of one-call	Below average; Average; Above average
A_3: ROW spacing	Intermittent and/or very limited indication; Continuous but limited indication; Continuous and highly indicative
A_4: One-call type	Multiple systems; Unified system to minimum standard; Unified system
A_5: Response time to dig notification	One day; Two days; Three days
A_6: Patrol frequency	Twice daily; Daily; Three times per week; Twice per week; Weekly; Bi-weekly; Monthly; Quarterly; Three times per year; Semi-annually
A_7: Locating method	Company records; Magnetic techniques
A_8: Response method to notification	Provide location information only; Locate/mark/site supervision
A_9: Burial depth	0.6 m, 0.7 m, …, 2.0 m

3.3 BN modeling, TPD datasets and parameter learning

3.3.1 BN modeling based on the fault tree

A BN is a directed graphical model representing the joint probabilistic distribution of a set of random variables that are symbolized by nodes. The dependence between nodes is symbolized by directed arcs and quantified by conditional probability tables (CPTs) attached to nodes. The entries in the CPT are called parameters of the corresponding node. The assignment of observed values to the corresponding nodes is called the instantiation of the nodes, which can lead to the Bayesian updating of the nodes that are dependent on the instantiated nodes. As an example, consider the BN model shown in Fig. 3.2. The nodes X_1 and X_2 with arcs pointing to node X_3 are the parents of X_3, and X_3 is the child node of X_1 and X_2. The nodes X_1 and X_2 are called the root nodes, as they do not have parent nodes. The CPT of the root node coincides with its probability mass function (PMF). Details of the BN modeling and efficient inference algorithms such as the junction tree algorithm for BNs are described in many textbooks, e.g. Nielsen and Jensen (2009) and Pearl (2004).

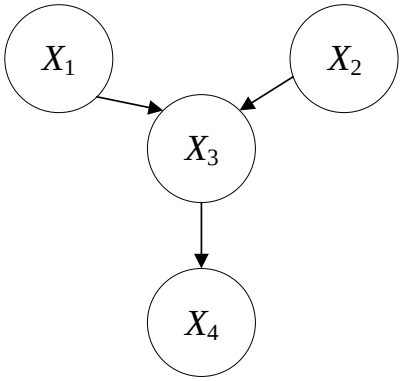


Figure 3.2 An example BN

Figure 3.3 shows the BN model that is developed in the commercial software Netica® based on the fault tree model in Fig. 3.1. Note that the fault tree model does not include the pipeline attributes A_n (n = 1, 2, …, 9), whereas these attributes are explicitly modeled as the parent nodes of basic events B_i (i = 1, 2, …, 9) in the BN model (gray nodes in Fig. 3.3). The conditional probabilities of the basic events are entries in the CPTs attached to

corresponding nodes. Nodes B_i (i = 1, 2, …, 9), E_m (m = 1, 2, …, 7) and T_0 have binary states "Yes" and "No", and the marginal probability associated with the state "Yes" represents the probability of the corresponding event. The following three examples are used to illustrate the BN modeling of three types of dependences involved in the fault tree, respectively. Figure 3.4 illustrates the BN modeling of the dependence of basic events B_i (i = 1, 2, …, 9) on pipeline attributes A_n (n = 1, 2, …, 9). The conditional probabilities of B_7 given A_7 are the parameters associated with the state "Yes" in the CPT attached to B_7 as shown in Table 3.3. The BN models equivalent to the "or" and "and" gates of the fault tree are shown in Figs. 3.5 and 3.6, respectively, and the corresponding CPTs attached to the output events of the "or" and "and" gates are shown in Tables 3.4 and 3.5, respectively. Uniform PMFs are assigned to the nodes A_n (n = 1, 2, …, 9) to represent the noninformative prior distributions for pipeline attributes before any information is available.

The BN model is a more flexible tool than the fault tree to predict P_h given a third-party activity under different scenarios of available information. To predict P_h given a third-party activity with an unknown authorization status, one instantiates nodes A_1 through A_9 by the given pipeline attributes and obtains the probability associated with the state "Yes" of the node T_0. In practice, P_h values corresponding to authorized and unauthorized activities respectively are often of interest (Lu and Stephen, 2016). To predict P_h given an authorized activity, nodes A_6 through A_9 (as opposed to A_1 through A_9) are instantiated, and node E_4 is instantiated by the state "No"; to predict P_h given an unauthorized activity, nodes A_6 through A_9 are instantiated, and node E_4 is instantiated by the state "Yes".

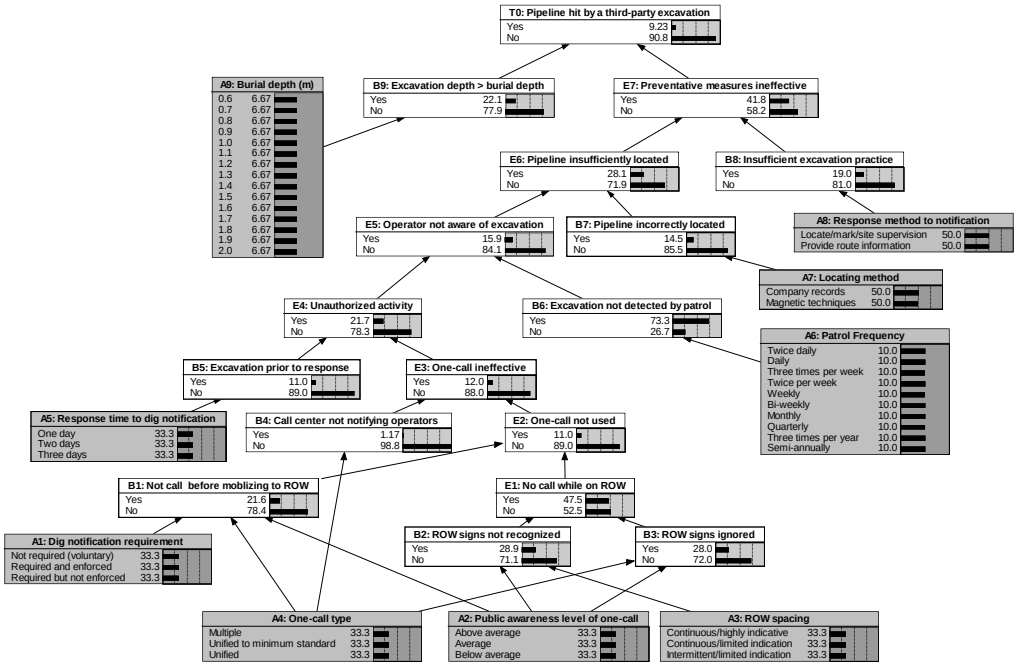


Figure 3.3 BN for evaluating P_h given a third-party activity

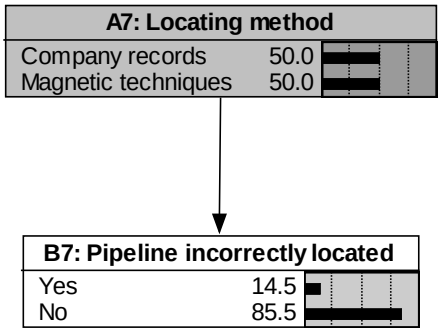


Figure 3.4 BN modeling the dependence of B_7 on A_7

Table 3.3 CPT of node B_7 in Fig. 4 based on the data in Chen and Nessim (1999)

Conditions	Conditional probabilities of node B_7	
States of A_7	State = "Yes"	State = "No"
Company records	0.2	0.8
Magnetic techniques	0.09	0.91

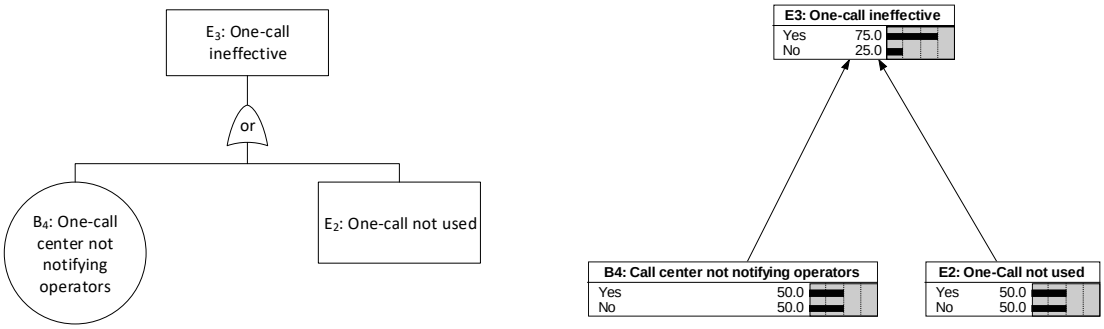


Figure 3.5 BN modeling of the "or" gate

Table 3.4 CPT of node E_3 in Fig. 3.5

Conditions		Conditional probabilities of node E_3	
States of B_4	States of E_2	State = "Yes"	State = "No"
Yes	Yes	1	0
Yes	No	1	0
No	Yes	1	0
No	No	0	1

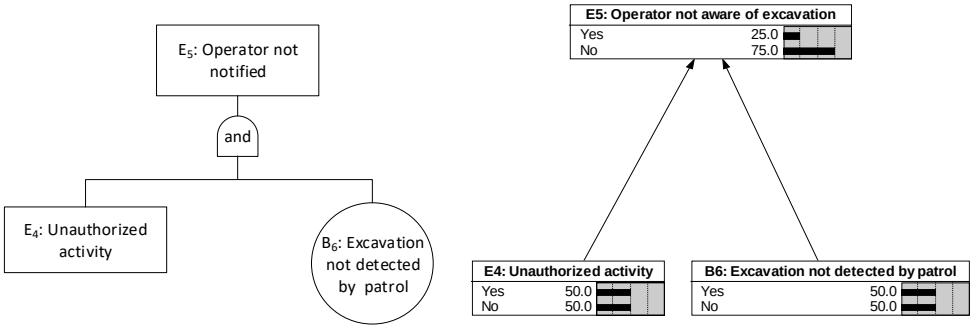


Figure 3.6 BN modeling of the "and" gate

Table 3.5 CPT of node E_5 in Fig. 3.6

Conditions		Conditional probabilities of node E_5	
States of B_6	States of E_4	State = "Yes"	State = "No"
Yes	Yes	1	0
Yes	No	0	1
No	Yes	0	1
No	No	0	1

3.3.2 TPD datasets for parameter learning

The pipeline company that provided the TPD data to the present study owns and operates an extensive network of transmission pipelines in Canada, and has been applying the fault-tree model to manage the TPD threat in the past decade. The company groups its pipeline assets into seven TPD regions based on the geographic location of the pipeline. The pipeline attributes denoted by nodes A_1 through A_9 in the BN model are the same for all the pipelines within the same TPD region. The company has been keeping records of third-party activities that were either notified by one-call systems, detected by ROW patrols, reported by landowners, or reported by the company employees since 2010. The records provided to the present study cover the period from 2010 to 2016. A recorded third-party activity is classified as unauthorized if one-call is not made or the excavation commences prior to the company's response to the one-call, which is consistent with the logic of the fault tree (see the "or" gate involving E_4, E_3, and B_5). Two datasets, referred to as Datasets 1 and 2, are extracted from these records for parameter learning. Dataset 1 consists of individual cases of third-party activities and the information that each individual case is classified as authorized or unauthorized activity, i.e. the values of A_1 through A_5 as well as E_4 in the BN model. As this dataset contains the information regarding the effectiveness of one-call systems, it is used to learn the parameters of basic events B_1 through B_5. Dataset 2 consists of individual cases of UAs and the outcome of given UAs (i.e. pipeline hit or not). That is, each case of Dataset 2 contains the values of A_6 through A_9, E_3, and T_0. Note that, since the third-party activities in Dataset 2 are known as UAs, the value of E_3 for every case is "Yes". As this dataset contains the information regarding the effectiveness of preventative and protective measures against UA activities, they are used to learn the parameters of basic events B_6 through B_9. It is noted that there is missing information in both Datasets 1 and 2, specifically, the information about events B_i (i = 1, 2, …, 9), which presents significant challenges to estimating the conditional probabilities of B_i (i = 1, 2, …, 9) given the values of A_n (n = 1, 2, …, 9). The EM algorithm presented in the next section is employed to learn these conditional probabilities from Datasets 1 and 2.

3.3.3 Parameter learning based on EM algorithm

Given the BN model described in Section 3.3.1, the task of estimating probabilities of the basic events in the fault tree model is now a problem of learning parameters of the BN with a known structure from incomplete datasets, more specifically, learning the parameters of nodes B_i (i = 1, 2, …, 9) from two incomplete TPD datasets. The parameter learning is performed in two steps. In the first step, the parameters of nodes B_1 through B_5 are learned from Dataset 1; in the second step, the parameters of nodes B_6 through B_9 are learned from Dataset 2.

As an example, the parameter learning of B_1 through B_5 is formulated as follows. While the parameter learning in the present study is focused on the incomplete datasets, the parameter learning based on the complete dataset is described first to improve the clarity of the formulation. Let $\theta_{i,j,k}$ (i = 1, 2, …,5; j = 1, 2, …, r_i; k = 1, 2) denote the parameters of the node B_i, i.e. the probability of the k-th state (i.e. Yes or No) under the j-th parent configuration, where r_i is the total number of parent configurations of B_i. For a given parent configuration j of B_i, $\theta_{i,j,k}$ (k = 1, 2) are considered as a vector of two random variables following the Dirichlet distribution with hyperparameters $\alpha_{i,j,1}$ and $\alpha_{i,j,2}$. The term hyperparameter is used to distinguish $\alpha_{i,j,1}$ and $\alpha_{i,j,2}$ from the parameters of the BN model. Before observations are obtained, the estimated value of $\theta_{i,j,k}$, denoted by $\hat{\theta}_{i,j,k}$, can be set to the corresponding mean values of the Dirichlet distribution:

$$\hat{\theta}_{i,j,k} = \frac{\alpha_{i,j,k}}{\alpha_{i,j,0}} \tag{3.3}$$

where $\alpha_{i,j,0} = \alpha_{i,j,1} + \alpha_{i,j,2}$ is known as the equivalent sample size of the Dirichlet distribution (Heckerman, 1998). Once a set of observations are obtained, the Bayesian updating of the distribution of $\theta_{i,j,k}$ (j = 1, 2, …, r_i) is carried out. Assume that there are a total of n sets of observations (i.e. n cases), each of which contains the complete information, i.e. values of A_n (n = 1, 2, …, 5) and B_i (i = 1, 2, …, 5). Let $n_{i,j,1}$ and $n_{i,j,2}$ denote numbers of observations of B_i in the state of “Yes” and “No”, respectively, under the j-th parent configuration; $n_{i,j,k}$ (k = 1, 2) are considered drawn from a multinomial distribution, of which the hyperparameters are $\theta_{i,j,1}$ and $\theta_{i,j,2}$. Given the Dirichlet-

multinomial conjugate pair, the posterior distribution of $\theta_{i,j,k}$ is also a Dirichlet distribution (Heckerman, 1998) with parameters $\alpha_{i,j,1} + n_{i,j,1}$ and $\alpha_{i,j,2} + n_{i,j,2}$. With these observations, $\hat{\theta}_{i,j,k}$ can be set to the mean of the posterior Dirichlet distribution, i.e.

$$\hat{\theta}_{i,j,k} = \frac{\alpha_{i,j,k}+n_{i,j,k}}{\alpha_{i,j,0}+n_{i,j,0}} \tag{3.4}$$

where $n_{i,j,0} = n_{i,j,1} + n_{i,j,2}$. This completes the parameter learning for B_i (i = 1, 2, …, 5) under the complete data scenario.

Now consider the scenario of incomplete or missing data, i.e. parameter learning from Dataset 1 described in Section 3.3.2. Assume that Dataset 1 contains a total of n cases. The EM algorithm (Dempster et al., 1977) is commonly employed to learn the parameters with incomplete data. To this end, the posterior distribution of $\theta_{i,j,k}$ is a Dirichlet distribution with parameters $\alpha_{i,j,1}$ + E[$n_{i,j,1}$] and $\alpha_{i,j,2}$ + E[$n_{i,j,2}$], where E[$n_{i,j,1}$] and E[$n_{i,j,2}$] are the expected numbers of observations of B_i in the state of "Yes" and "No", respectively, under the j-th parent configuration. The value of E[$n_{i,j,k}$] (k = 1 and 2) is calculated as follows,

$$\mathrm{E}\left[n_{i,j,k}\right] = \sum_{l=1}^{n} p\left(b_{i,k}, \mathrm{pa}_{i,j} | O_l\right) \tag{3.5}$$

where $b_{i,k}$ and $\mathrm{pa}_{i,j}$ are the k-th state and j-th parent configuration of B_i, respectively, and $p\left(b_{i,k}, \mathrm{pa}_{i,j} | O_l\right)$ denotes the joint probability of $b_{i,k}$ and $\mathrm{pa}_{i,j}$ given the l-th case (O_l) and can be obtained from the Bayesian updating once the BN is instantiated by the evidence in O_l, i.e. corresponding values of A_n (n = 1, 2, …, 5) and E_3. The value of $\hat{\theta}_{i,j,k}$ is now given by,

$$\hat{\theta}_{i,j,k} = \frac{\alpha_{i,j,k}+\mathrm{E}\left[n_{i,j,k}\right]}{\alpha_{i,j,0}+\sum_{j=1}^{r_c} \mathrm{E}\left[n_{i,j,k}\right]} \tag{3.6}$$

It follows that the evaluation of Eqs. (3.5) and (3.6) is an iterative process, as $\hat{\theta}_{i,j,k}$ obtained in the current iteration is used to estimate $\mathrm{E}\left[n_{i,j,k}\right]$ and thus leads to a new $\hat{\theta}_{i,j,k}$ in the next iteration. The iteration is terminated once the log-likelihood of the observations converges to a local maximum, and this completes the parameter learning for B_i (i = 1, 2, …, 5) under

the incomplete data scenario. The above formulation of EM algorithm applies equally to the parameter learning for nodes B_6 through B_9 from Dataset 2.

3.4 Numerical example and case study

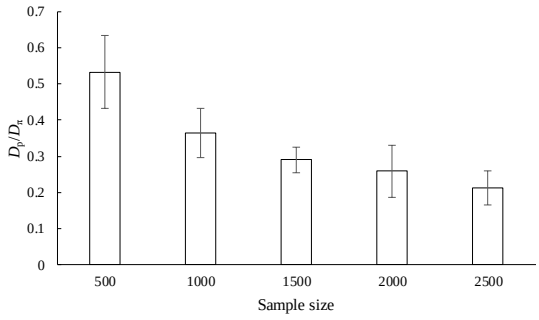
This section first uses a numerical example involving simulated TPD data to demonstrate the effectiveness of the parameter learning for this specific BN model. Then, a case study involving real-world TPD datasets, i.e. Datasets 1 and 2 described in Section 3.3.2, is presented.

3.4.1 Numerical example involving simulated TPD data

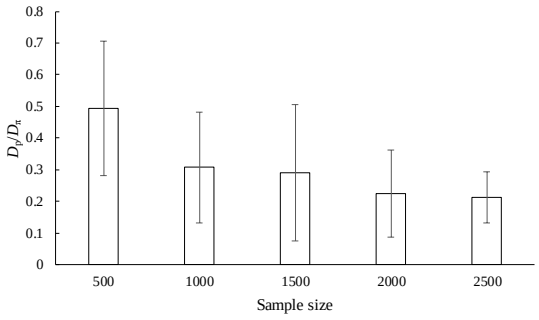
This example is introduced following the common practice of examining the effectiveness of parameter learning using simulated data in the literature (Masegosa et al., 2016; Liao and Ji, 2010; Zhou et al., 2016). First, a baseline BN is developed as described in Section 3.3.1, for which the CPTs of nodes B_1 through B_9 are created based on the data in the literature (Chen and Nessim, 1999; TransCanada Corporation, 2017). The parameters of B_1 through B_9 of the baseline BN are considered the true parameters. Then, two datasets, Datasets I and II consisting of individual cases of third-party activities are simulated using the baseline BN as follows. Note that Datasets I and II are missing the same information as that in Datasets 1 and 2, respectively, described in Section 3.3.2. To simulate Dataset I, a prescribed number of individual cases of third-party activities are drawn from the baseline BN model using the forward algorithm (Henrion, 1988), which is the standard sampling algorithm implemented in the software Netica®. For each simulated case of third-party activity in Dataset I, the values of A_1 through A_5 as well as E_4 are kept, whereas the values associated with the other nodes in the BN are removed, thus creating an incomplete dataset. The individual cases of third-party activities in Dataset II are simulated by first instantiating the state of node E_4 as "Yes" (i.e. unauthorized activities only). Then, for each simulated case, the values of A_6 through A_9, E_4 and T_0 are kept, whereas the values of the other nodes are removed. For both Datasets I and II, five sample sizes are considered: 500, 1000, 1500, 2000 and 2500.

The parameter learning is carried out on a prior BN model, of which the structure and the parameters associated with A_n (n = 1, 2, …, 9), E_m (m = 1, 2, …, 7) and T_0 are the same as

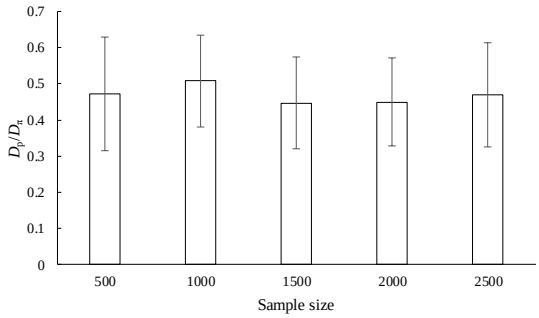
the baseline model. However, the symmetric Dirichlet distribution with equivalent sample size of unity is assigned to the parameters of B_i (i = 1, 2, …, 9) for given parent configuration, that is, $\alpha_{i,j,1} = \alpha_{i,j,2} = 0.5$. This prior assumption is corresponding to the least-informative Jeffreys prior (Kelly and Atwood, 2011). The EM algorithm is employed to learn the parameters of B_1 through B_5, and B_6 through B_9 from Datasets I and II, respectively. For a given node B_i, the Kullback-Leibler (KL) divergence between the CPT of the learned BN and CPT of the baseline BN is evaluated as a measure of the difference between the two CPTs (Kullback and Leibler, 1951). The smaller is the KL-divergence, the closer is the learned CPT to the true CPT, indicating more effective parameter learning. To facilitate the observation, the normalized KL-divergence, D_p/D_π, is used to express the parameter learning results, where D_π denotes the KL-divergence between the CPTs of the prior BN and baseline BN, and D_p denotes the KL-divergence between the CPT of the learned BN and baseline BN. It follows that D_p/D_π less than unity is desirable. As the variability in the simulated samples may introduce variability in the results of the parameter learning, the simulation of the TPD dataset and corresponding parameter learning for a given sample size are repeated 10 times. The mean value (vertical bar) and one-standard-deviation interval (error line on the vertical bar) of D_p/D_π of the 10 trials for nodes B_1 through B_9 are depicted in Figs. 3.7(a) through 3.7(i), respectively. These figures indicate that the values of D_p/D_π associated with all the nodes are less than unity, demonstrating the effectiveness of the parameter learning. In general, as the sample size increases, the mean value and standard deviation of D_p/D_π decrease, which indicates that the performance of parameter learning is improved as the sample size increases.



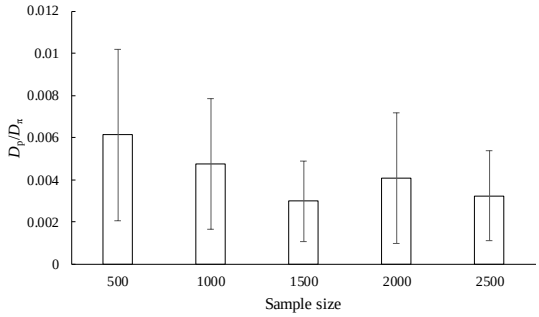
(a) Node B_1



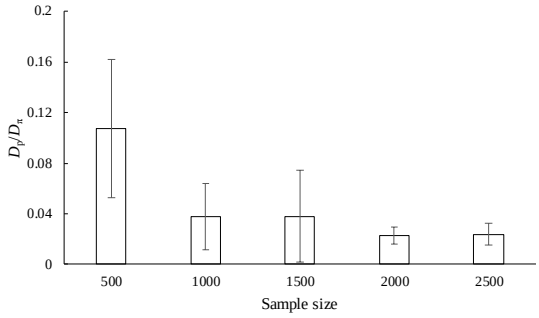
(b) Node B_2



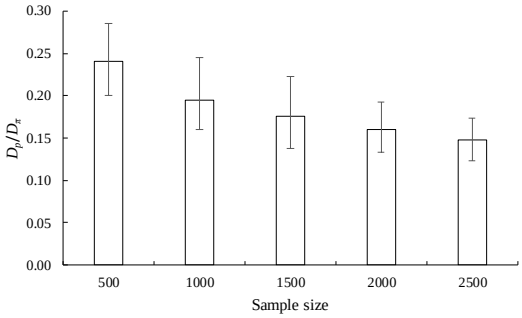
(c) Node B_3



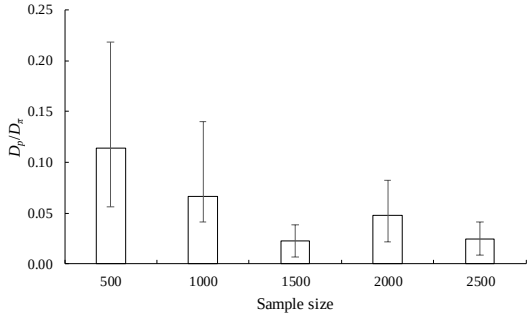
(d) Node B_4



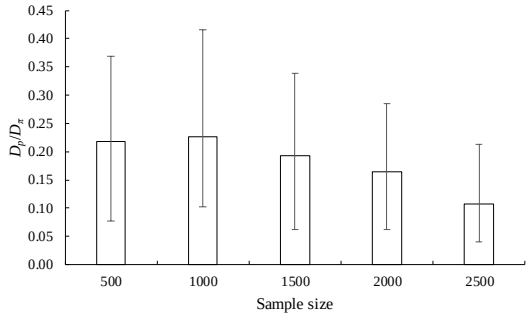
(e) Node B_5



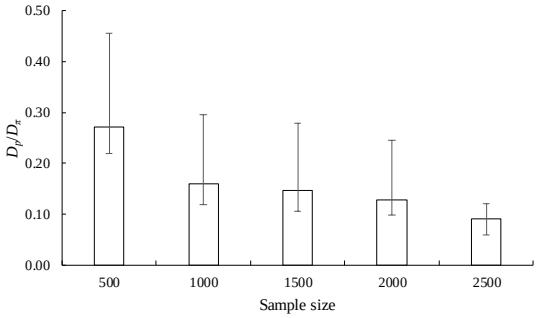
(f) Node B_6



(g) Node B_7



(h) Node B_8

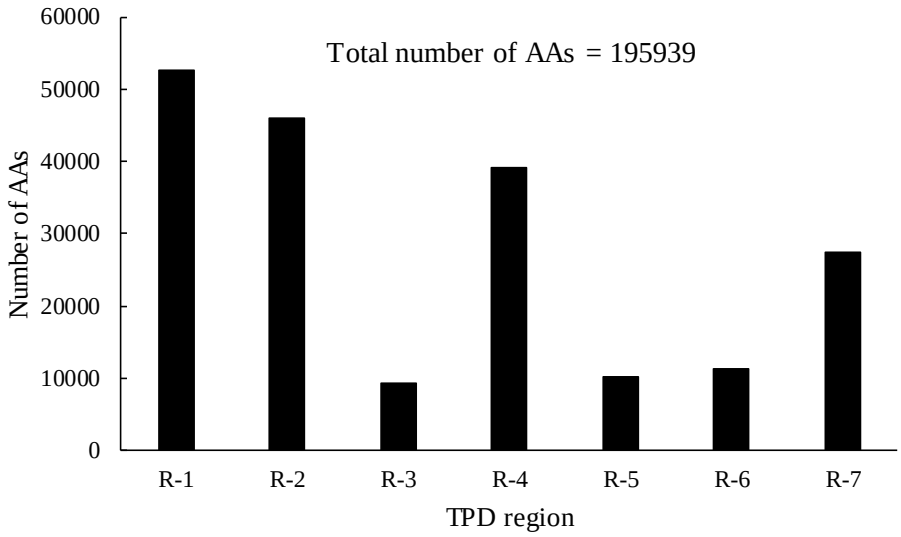


(i) Node B_9

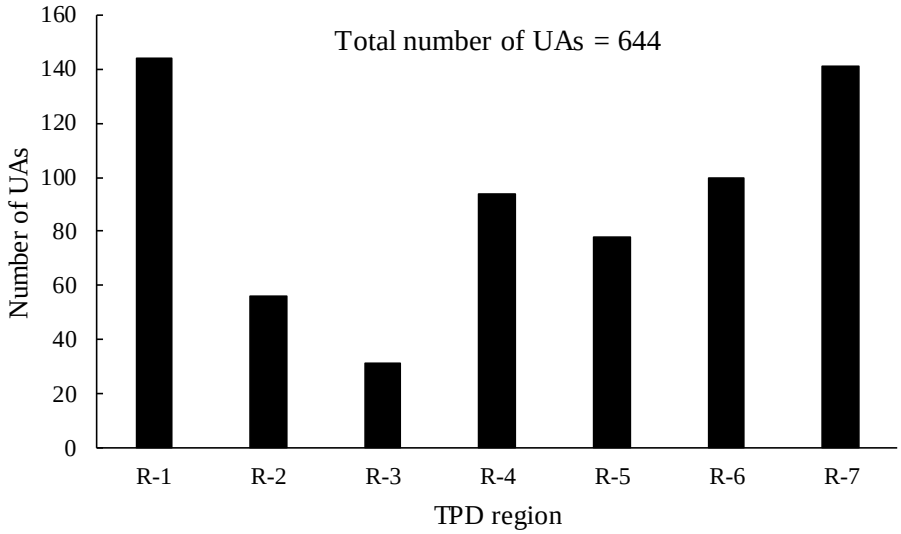
Figure 3.7 KL-divergence associated with nodes B_1 through B_9 in the numerical example

3.4.2 Case study using real TPD data

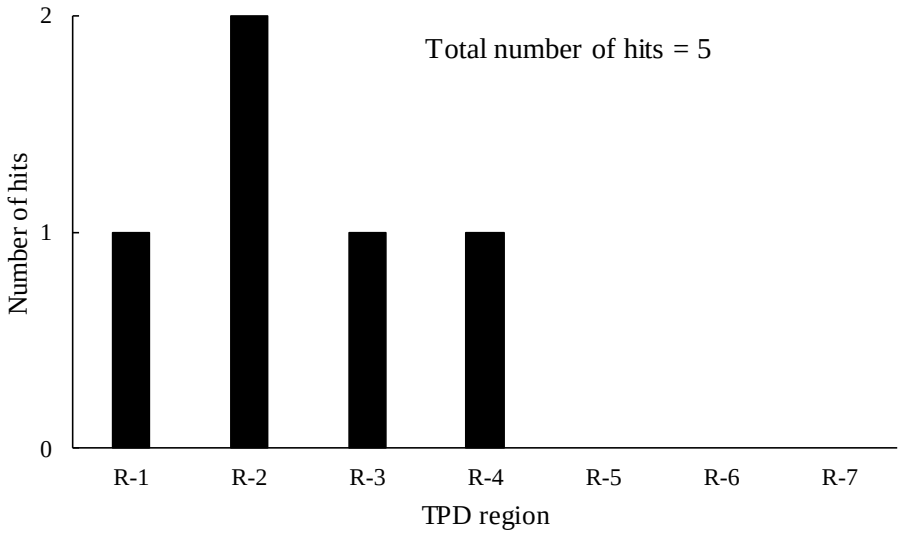
In this section, parameter learning is applied to the prior BN to learn the parameters of node B_1 through B_9 using Datasets 1 and 2 described in Section 3.3.2. The datasets are the third-party activities occurring on pipelines in seven TPD regions in Canada between 2010 and 2016. The TPD regions are denoted as R-1 through R-7, of which the pipeline attributes are given in Appendix A. The number of AAs, UAs, and pipeline hits resulting from UAs within R-1 through R-7 are shown in Figs. 3.8(a) through 3.8(c), respectively. Note that there were no pipeline hits in R-5, R-6 and R-7 between 2010 and 2016. The TPD data associated with Figs. 3.8(a) and 3.8(b) constitute Dataset 1, and TPD data associated with Figs. 3.8(b) and 3.8(c) constitute Dataset 2.



(a) Number of AAs per TPD region



(b) Number of UAs per TPD region



(c) Number of pipeline hits caused by UAs per TPD region

Figure 3.8 Number of pipeline hits caused by UAs per TPD region

The EM algorithm is implemented on a prior model that is the same as described in Section 3.4.1 to learn the parameters of B_1 through B_5 (i.e. nodes related to the effectiveness of one-call) from Dataset 1, and B_6 through B_9 (i.e. nodes related to the effectiveness of preventative and protective measures against UAs) from Dataset 2. Since the true parameters corresponding to the real-world datasets are unknown, the performance of the parameter learning is examined indirectly by comparing the model-predicted probabilities with corresponding empirical probabilities as follows. To examine the performance of the parameter learning with respect to nodes B_1 through B_5, we compare the empirical and model-predicted probabilities of a third-party activity being unauthorized. For a given TPD region, the model-predicted value is the probability associated with the state "Yes" of node E_4 by instantiating nodes A_1 through A_5 with the corresponding pipeline attributes shown in Appendix A. The empirical value is evaluated as the ratio of the number of UAs to the total number of third-party activities (i.e. the sum of the numbers of UAs and AAs) associated with the TPD region. Figure 3.9 depicts the empirical probability and probabilities predicted by the baseline BN and learned BN (i.e. parameters of B_1 through B_5 obtained from the parameter learning) for the seven TPD regions. The figure indicates that the probabilities predicted by the baseline BN model are in general two orders of

magnitude higher than the empirical probabilities, whereas the probabilities predicted by the learned BN model agree well with the empirical probabilities. Therefore, the parameter learning has effectively elicited the parameters of nodes B_1 through B_5 from Dataset 1.

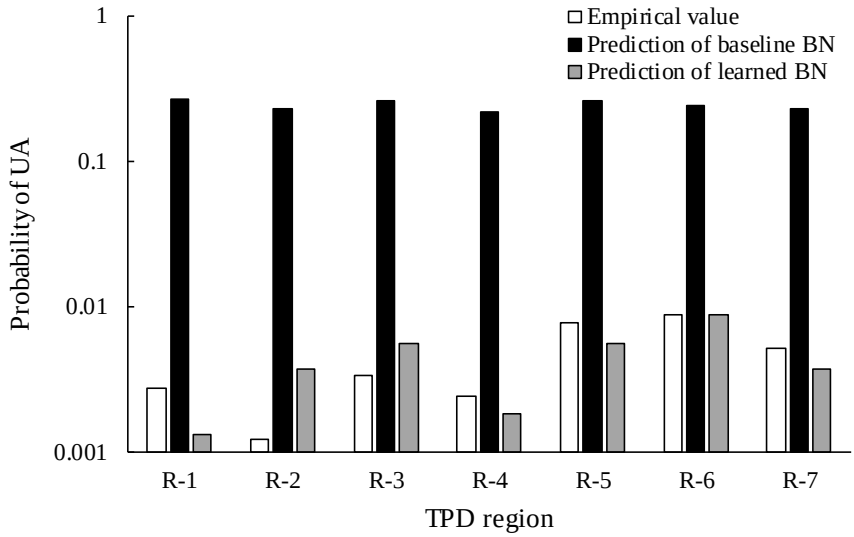


Figure 3.9 Comparison of the empirical and model-predicted probability of a third-party activity being unauthorized

The second quantity used to examine the performance of the parameter learning with respect to nodes B_6 through B_9 is P_h given a UA. As shown in Fig. 3.8(b), pipeline hits caused by UAs are only observed on pipelines in R-1 through R-4; the empirical P_h given a UA for each of these four TPD regions is evaluated as the ratio of the corresponding number of hits to the total number of UAs observed. Since no pipeline hits are observed for regions R-5 through R-7, the empirical P_h given a UA for these TPD regions is zero. For a given TPD region, the model-predicted P_h given a UA is the probability associated with the state "Yes" of node T_0 by instantiating nodes A_6 through A_9 by the corresponding pipeline attributes shown in Appendix A, and node E_4 by the state "Yes". The comparison of the empirical probability and probabilities predicted by the baseline BN and learned BN is shown in Fig. 3.10. This figure indicates that, for TPD regions R-1 through R-4, the probabilities predicted by the baseline BN are generally one order of magnitude higher than the corresponding empirical values, whereas the probabilities predicted by the learned BN (i.e. parameters of B_6 through B_9 obtained from the parameter learning) agree well with the

empirical values. For TPD regions R-5 through R-7, where no pipeline hits are observed, the probabilities predicted by the baseline BN are comparable to those for regions R-1 through R-4, which is overly conservative. On the other hand, the probabilities predicted by the learned BN for R-5 through R-7 are significantly lower than those for R-1 through R-4, therefore more reflective of the reality.

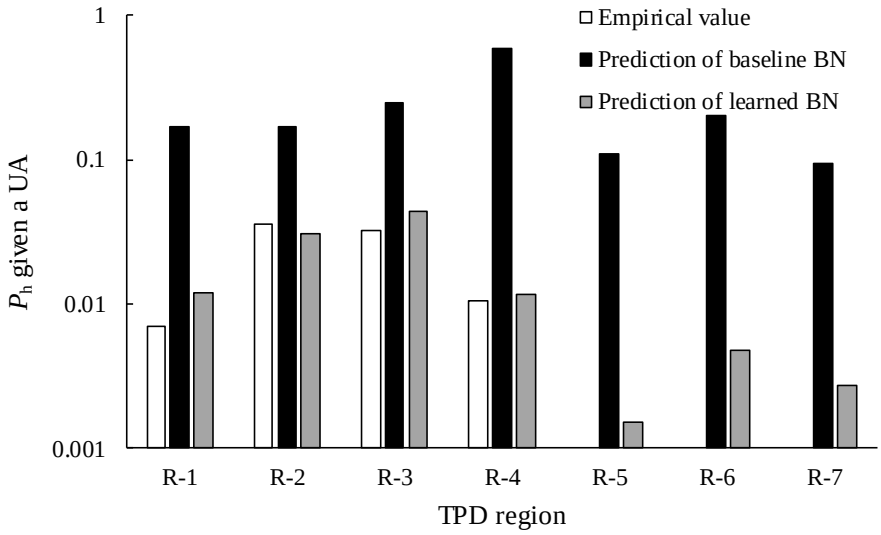


Figure 3.10 Comparison of the model-predicted and empirical P_h given a UA

3.5 Conclusions

The present study proposes a BN model to evaluate the probability of pipelines being hit by third-party excavation activities and apply the parameter learning technique to learn CPTs of the BN model from TPD datasets. The BN model is developed based on a fault tree model commonly used in the pipeline industry. The EM algorithm in the context of parameter learning is employed to learn CPTs of the BN model from two incomplete datasets consisting of individual cases of third-party activities. The effectiveness of the parameter learning is first demonstrated by a numerical example involving simulated TPD datasets, where the KL-divergence between the learned CPT and true CPT is adopted as the metric. The effectiveness of the parameter learning is further demonstrated by using two real-world TPD datasets collected by a Canadian pipeline operator between 2010 and 2016. The performance of the parameter learning is examined by comparing the empirical

value and model-predicted value of two quantities: the probability of a third-party activity being unauthorized and the probability of hit given a UA. The results indicate that the probabilities predicted by the BN with the parameters obtained from the parameter learning agree well with the corresponding empirical values. Therefore, the techniques of BN modeling and parameter learning provide an effective and efficient means to exploit the historical TPD datasets collected by pipeline operators to improve the pipeline integrity management practice with respect to TPD.

References

Bobbio, A., Portinale, L., Minichino, M., and Ciancamerla, E. (2001). Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability Engineering and System Safety*, 71(3), 249-260.

Chen, Q., Davis, K., and Parker, C. (2006). Modeling damage prevention effectiveness based on industry practices and regulatory framework. In: *Proceedings of the 9th International Pipeline Conference* (pp. 635-645), Calgary, Alberta, Canada.

Chen, Q., and Nessim, M. A. (1999). Reliability-based prevention of mechanical damage to pipelines. *Pipeline Research Council International, Project PR-244-9729*.

Common Ground Alliance. (2016). DIRT analysis and recommendation, Volume 11.

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 (Statistical Methodology)*, 39(1), 1-38.

European Gas Pipeline Incident Data Group (EGIG). (2018), 10th EGIG Report (period 1970-2016), Document number: VA 17.R.0395.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. *Learning in graphical models* (pp. 301-354). Springer, Dordrecht.

Henrion, M. (1988). Propagating uncertainty in Bayesian networks by probabilistic logic sampling. *Machine Intelligence and Pattern Recognition*, 5, 149-163.

Kelly, D., and Atwood, C. (2011). Finding a minimally informative Dirichlet prior distribution using least squares. *Reliability Engineering and System Safety*, 96(3), 398-402.

Khakzad, N., Khan, F., and Amyotte, P. (2011). Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. *Reliability Engineering and System Safety*, 96(8), 925-932.

Koduru, S. D., and Lu, D. (2016). Equipment Impact Rate Assessment Using Bayesian Networks. In *Proceedings of the 11th International Pipeline Conference* (pp. V002T07A013-V002T07A013), Calgary, Alberta, Canada.

Koduru, S. D., and Nessim, M. A. (2017). Review of Quantitative Reliability Methods for Onshore Oil and Gas Pipelines. In *Risk and Reliability Analysis: Theory and Applications* (pp. 67-95). Springer.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79-86.

Lam, C., and Zhou, W. (2016). Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database. *International Journal of Pressure Vessels and Piping*, 145, 29-40.

Liao, W., and Ji, Q. (2010). Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 42(11), 3046-3056.

Lu, D., and Stephens, M. (2016). Analyzing the Effectiveness of Prevention Measures for Third-Party Damage to Underground Pipelines Using a Hierarchical Fault Tree Model. In: *Proceedings of the 11th International Pipeline Conference* (pp. V002T07A022-V002T07A022), Calgary, Alberta, Canada.

Masegosa, A. R., Feelders, A. J., and van der Gaag, L. C. (2016). Learning from incomplete data in Bayesian networks with qualitative influences. *International Journal of Approximate Reasoning*, 69, 18-34.

Mearns, A. B. (1965). Fault tree analysis- the study of unlikely events in complex systems (Fault tree analysis as tool to identify component failure as probable cause of undesired event in complex system). In: *System Safety Symposium*, Seattle, Washington.

Nielsen, T., and Jensen, F. (2009). *Bayesian networks and decision graphs*. New York, NY: Springer Science and Business Media.

Pearl, J. (2004). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

TransCanada Corporation. (2017). System-wide risk assessment (SWRA) - Appendix D: Third-party external interference. Calgary, Canada.

Zhou, Y., Fenton, N., and Zhu, C. (2016). An empirical study of Bayesian network parameter learning with monotonic influence constraints. *Decision Support Systems*, 87, 69-79.

4 A non-parametric Bayesian network model for predicting corrosion depth on buried pipelines

4.1 Introduction

Historical failure data indicate that metal-loss corrosion is one of the major threats to the structural integrity of underground oil and gas pipelines (CEPA, 2015; Lam and Zhou, 2016). Since the corrosion deterioration on buried pipelines is greatly influenced by the corrosive properties of surrounding soils, characterizing the correlation of corrosion sizes with relevant local soil parameters has received a great deal of attention in the research community (Caleyo et al, 2009; Jyrkama et al., 2016; Melchers and Petersen, 2018; Ricker, 2010; Velázquez et al., 2009). Predicting the corrosion depth based on soil properties is of great practical value to the corrosion assessment of buried pipelines to which in-line inspection (ILI) technique is infeasible, i.e. unpiggable pipelines (Beauregard et al., 2018).

Many models to predict corrosion depths using soil parameters as predictors have been reported in the literature during the past several decades (Romanoff, 1957; Velázquez et al., 2009; Caleyo et al, 2009; Ricker, 2010; Alamilla et al., 2009; Yajima et al., 2015; Wang et al., 2016). Most of these models were developed based on multivariate regression analyses of corrosion datasets reported in the open literature such as the National Bureau of Standards (NBS) dataset (Romanoff, 1957) and dataset reported by Velázquez et al. (2010) (referred to as Velázquez's dataset hereafter). The NBS dataset was collected from extensive field studies of corrosion on a variety of ferrous specimens including pipelines buried in 128 sites with representative soils across the United States for up to 17 years (Romanoff, 1957). This dataset contains the measurements of the deepest corrosion depths on experimental specimens together with a group of local soil parameters. The analysis of the corrosion data indicated that the growth path of the corrosion depth follows a power-law function of exposure time with the exponent parameter less than unity (Romanoff, 1957). Since then, extensive studies have been performed based on the NBS dataset to investigate the correlation between the measured soil parameters and develop regression models for predicting the corrosion depth (Jyrkama et al., 2016; Romanoff, 1956; Ricker, 2010; Schwerdtfeger, 1966). However, the analyses indicated a lack of strong correlation between the corrosion depth and soil parameters (Jyrkama et al., 2016; Ricker, 2010).

Ricker (2010) concluded that due to a lack of statistical considerations in the process of designing the experiment, the multivariate regression analysis was not suitable to develop predictive models using the NBS dataset. Velázquez's corrosion dataset was collected from 259 excavation sites of underground energy pipelines in southern Mexico, of which each individual sample consists of the maximum corrosion depth in the excavation site (i.e. the maximum corrosion depth on the exposed pipeline segment), age of the pipeline, and local soil parameters (Velázquez et al., 2010). The relatively large sample size of Velázquez's dataset makes it more suitable to use for statistical and probabilistic analysis than the NBS dataset. Moreover, since Velázquez's dataset was collected from real pipelines instead of experimental specimens, it is considered more reflective of characteristics of pipeline corrosion in reality than the NBS dataset. Using this dataset, Velázquez et al. (2009) employed the power-law function to characterize the growth of corrosion depth. The proportionality and exponent parameters of the power-law function were assumed to be linear functions of soil parameters and determined by the multivariate regression analysis.

The predictive regression models developed based on the corrosion datasets have a few drawbacks. First, the functional forms of model parameters, such as the proportionality and exponent parameters of the power-law function, in terms of soil parameters are usually decided based on assumptions, which brings marked subjectivity into the developed regression model. Second, due to the interaction of different soil parameters, analyses of corrosion data often indicate a lack of strong correlation between the corrosion depth and individual soil parameters (Jyrkama et al., 2016; Ricker, 2010; Velázquez et al., 2009). This implies that deterministic models such as regression models are not appropriate to characterize the relationship between corrosion depths and soil parameters. The inherent spatial and temporal variability associated with the soil parameters and corrosion depths further suggest that it is more appropriate and objective to characterize the relationship between corrosion depth and soil parameters probabilistically than deterministically.

In the present study, the non-parametric Bayesian network (NPBN) technique (Kurowicha and Cooke, 2005; Hanea et al., 2006) is employed to develop a probabilistic predictive model for the corrosion depth based on Velázquez's dataset. An NPBN is a directed acyclic

graph (DAG) with nodes and arcs symbolizing a set of continuous random variables and dependence between them, respectively. Due to the intuitive graphical nature and ability to efficiently deal with continuous random variables, NPBNs have become increasingly popular for the high dimensional dependence modeling and risk analysis (Zilko et al., 2016; Morales-Napoles and Steenbergen, 2014; Hanea et al., 2015; Morale-Napoles et al., 2014; Hanea et al., 2013; Lee and Pan, 2018; Wang et al., 2019). The model mining method, established by Hanea et al. (2010) to facilitate the development of NPBN based on multivariate datasets, is employed in the present study to develop an NPBN model, which involves the corrosion depth and ten predictors including the pipeline age and local soil parameters. Once the nodes representing predictor variables are instantiated, the developed NPBN can infer the probabilistic distribution of the corrosion depth.

The remainder of this chapter is organized as follows. Section 4.2 presents a brief introduction to the theory of NPBN and mining method of developing NPBN from a multivariate dataset. Section 4.3 formulates the NPBN model for predicting the corrosion depth based on the Gaussian copula. An overview of Velázquez's dataset is provided in Section 4.4. Section 4.5 develops the NPBN predictive model using the Velázquez's dataset and validates the model by the means of 5-fold cross-validation, followed by conclusions in Section 4.6.

4.2 Non-parametric Bayesian network and model mining method

4.2.1 Bayesian network, copula and non-parametric Bayesian network

A Bayesian network (BN) is a DAG of the joint probability distribution of a set of random variables (Nielsen and Jensen, 2009). A BN consists of nodes representing the random variables and directed arcs representing causal (i.e. parent-child) relationships between the nodes. Through the conditional independence statements encoded in the graph, a high-dimensional joint probability distribution can be represented as a factorization of a series of conditional probability distributions, thus simplifying the computation. Given data or evidence observed on a subset of the nodes in a BN, the joint probability distribution of the rest of the nodes in the BN can be updated through Bayes' theorem. This is the so-called inference, the most important application of BNs. Various exact and approximate

inference algorithms are described in many textbooks (e.g. Nielsen and Jensen, 2009; Pearl, 2014). BNs are generally applicable to discrete random variables (Langseth et al., 2009): the marginal and conditional distributions are defined through the probability mass functions and conditional probability tables, respectively. Continuous random variables are generally discretized to be included in a BN. If a significant number of continuous random variables are however included in a BN, each of them discretized by a sufficiently large number of states to ensure the computational accuracy, the efforts for specifying the conditional probability tables can become prohibitively burdensome. The discretization can be avoided if the continuous random variables are assumed to follow a jointly normal distribution (Hanea et al., 2006); however, the joint normality assumption may not be justified by reality.

NPBN is developed to overcome the above-described drawbacks of BN in dealing with continuous random variables. Introduced by Kurowicka and Cooke (2005) and extended by Hanea et al. (2006), an NPBN is a DAG with nodes and arcs symbolizing a set of continuous random variables and dependence between them, respectively. The term "non-parametric" reflects the fact that copulas are used to couple marginal distributions of random variables in NPBN, therefore eliminating the need to assume their joint probability distribution. The dependence between any two nodes is quantified by the (conditional) Spearman's rank correlation, which is the correlation coefficient between ranks, i.e. cumulative distribution functions (CDFs), of the two variables.

Since the copula concept is central to NPBN, a brief description of copula is presented in the following. A copula, $C(u_1, u_2, \ldots, u_n) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_n \leq u_n)$, is a joint probability distribution of standard uniformly distributed random variates U_i ($i = 1, 2, \ldots, n$). Sklar (1959) showed that any n-variate probability distribution function, $F(x_1, x_2, \ldots, x_n)$, can be written as the following copula form:

$$C\left(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(x_n)\right) = F(x_1, x_2, \ldots, x_n) \tag{4.1}$$

where $F_{X_i}(x_i)$ is the marginal CDF of random variable X_i ($i = 1, 2, \ldots, n$) evaluated at the value x_i, and $C(\bullet)$ is the copula. Many copula functions have been developed, e.g. the

Frechet, Clayton and Gaussian copulas (Nelsen, 2007). While any copula function can be used in NPBN, the Gaussian copula is of particular importance to NPBN mainly because it allows analytical inferences, which greatly improves the computational efficiency of NPBN. The Gaussian copula is given by,

$$C\left(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(x_n)\right) = \Phi_n\left(\Phi^{-1}\left(F_{X_1}(x_1)\right), \Phi^{-1}\left(F_{X_2}(x_2)\right), \ldots., \Phi^{-1}\left(F_{X_n}(x_n)\right)\right) \quad (4.2)$$

where $\Phi_n(\bullet)$ is the n-variate normal distribution function with the $(n \times n)$-dimensional linear correlation matrix Σ, and $\Phi^{-1}(\bullet)$ is the inverse of the standard univariate normal distribution function. Let ρ_{ij} ($i, j = 1, 2, \ldots, n$) denote the elements of Σ, i.e. the linear correlation coefficient between U_i and U_j (for brevity the term "linear" is omitted thereafter), where U_i (U_j) corresponds to X_i (X_j) through the inverse normal transformation, and let r_{ij} denote the rank correlation coefficient between U_i and U_j. Then ρ_{ij} is related to r_{ij} through the following equation (Pearson, 1907):

$$\rho_{ij} = 2 \sin\left(\frac{\pi}{6} r_{ij}\right) \ (i, j = 1, 2, \ldots, n) \quad (4.3)$$

Consider the simple example of a DAG consisting of four nodes (Fig. 4.1) as described in Hanea et al. (2006). The four nodes represent four continuous random variables, respectively, with the corresponding invertible marginal distributions. Note that if data are available, the marginal distribution can be straightforwardly defined, e.g. using the empirical CDF or parametric CDF obtained from distribution fitting techniques. Four rank correlation coefficients are then assigned, respectively, to the four directed arcs in Fig. 4.1. To this end, r_{13} and r_{24} define the unconditional rank correlation coefficients between nodes 1 and 3, and nodes 2 and 4, respectively. Since node 4 has two parents (2 and 3), the conditional rank correlation concept is employed: r_{24} defines the unconditional rank correlation coefficient between 2 and 4, whereas $r_{34|2}$ defines the conditional rank correlation coefficient between 3 and 4 given 2. It follows that the order of the factorization is not unique, i.e. r_{34} and $r_{24|3}$ being also valid specifications. The theorem developed by Hanea et al. (2006) ensures that the joint distribution of the four random variables in Fig.

4.1 is uniquely defined given the DAG and specifications of marginal distributions, (conditional) rank correlation coefficients, and the copula function to link the marginal distributions. The (conditional) rank correlation coefficients defined as such are algebraically independent, i.e. any numbers in (-1, 1) and consistent.

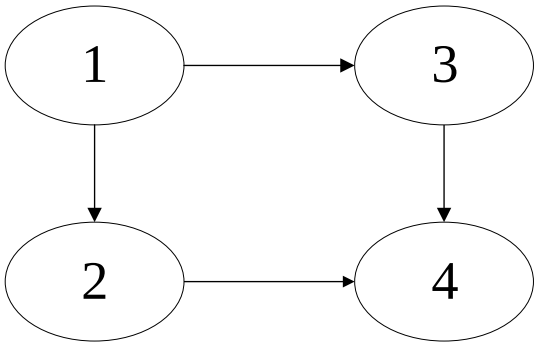


Figure 4.1 NPBN with four nodes and four arcs

If the Gaussian copula is used, then the (conditional) correlation coefficient can be evaluated from the (conditional) rank correlation coefficient using Eq. (4.3). Furthermore, the conditional correlation coefficient equals the partial correlation coefficient for the Gaussian copula. For a set of n random variables X_1, X_2, …, X_n, the partial correlation coefficient between X_1 and X_2 based on X_3, …, X_n, denoted by $\rho_{12;3,\ldots,n}$, is geometrically interpreted as the correlation between the projections of X_1 and X_2 on the plane orthogonal to the space spanned by X_3, …, X_n (Hanea, 2008; Zeng et al. 2017). Partial correlation coefficients can be recursively computed from the correlation coefficients as follows:

$$\rho_{12;3,\ldots,n} = \frac{\rho_{12;4,\ldots,n} - \rho_{13;4,\ldots,n} \cdot \rho_{23;4,\ldots,n}}{\sqrt{(1-\rho^2_{13;4,\ldots,n})(1-\rho^2_{23;4,\ldots,n})}} \tag{4.4}$$

Examples of using Eqs. (4.3) and (4.4) to evaluate the conditional rank correlations given the unconditional rank correlations, and inversely evaluate the unconditional rank correlations based on the conditional rank correlations attached to the arcs of the NPBN are included in Appendix B. Note that while the NPBN method is similar to the joint normal transform, the NPBN method is advantageous in that (conditional) rank correlations specified in an NPBN need not satisfy the algebraic constraint of positive definiteness as the elements in the correlation matrix do (Kurowicka and Cooke, 2006).

The above description suggests the following advantages of using the Gaussian copula in NPBN. First, correlation coefficients can be determined from the (conditional) rank correlation coefficients using Eqs. (4.3) and (4.4) as well as the fact that the partial correlation coefficient equals the conditional correlation coefficient for the Gaussian copula. Second, zero correlation (i.e. no arc) between two nodes is equivalent to (conditional) independence between the variables for the Gaussian copula. Third, the correlation matrix Σ is uniquely defined because the (conditional) rank correlation coefficients are algebraically independent (i.e. consistent). Finally, analytical updating given evidence is available because conditional distributions arising from a joint normal distribution are also normal.

4.2.2 Method for mining an NPBN from a multivariate dataset

Mining an NPBN from a given multivariate dataset that contains *k* sets of samples of *n* random variables involves evaluating the marginal distribution associated with each node and determining the dependence structure of the NPBN. The method proposed by Hanea et al. (2010) is employed in the present study. Evaluating empirical marginal distributions for the random variables is straightforward. The method of developing and validating the dependence structure of an NPBN involves, firstly validating the assumption that the multivariate dataset is drawn from a Gaussian copula, and secondly demonstrating that the developed NPBN has captured the significant dependences implicated in the multivariate dataset. The validation is carried out based on three correlation matrices: 1) the empirical rank correlation matrix, Σ_E, that is evaluated using the original samples of the variables in the dataset; 2) the empirical normal rank correlation matrix, Σ_N, that is evaluated by transforming original samples to standard normal variates, then transforming the linear correlation between the standard normal variates to rank correlations using Eq. (4.3); 3) the rank correlation matrix associated with the NPBN model, Σ_M, that is determined based on the (conditional) rank correlations attached to the arcs of the NPBN using Eqs. (4.3) and (4.4).

Let $\det(\Sigma_E)$, $\det(\Sigma_N)$ and $\det(\Sigma_M)$ denote the determinants of Σ_E, Σ_N and Σ_M, respectively. To validate the assumption that the multivariate dataset is from a Gaussian copula, a statistical test described as follows is carried out.

1.1) Evaluate Σ_E and $\det(\Sigma_E)$ using the original samples in the dataset.

1.2) Transform original samples to standard normal variates and evaluate the linear correlation matrix using the standard normal variates.

1.3) Generate *k* sets of samples from the *n*-variate normal distribution with zero mean values and the linear correlation matrix evaluated from step 1.2); use these samples to evaluate the rank correlation matrix and its determinant.

1.4) Repeat step 1.3) for 1000 times and thus generate 1000 samples of the determinant of the rank correlation matrix.

1.5) If $\det(\Sigma_E)$ is within the 5-95 percentile range of the samples generated in step 1.4), it is valid to assume the multivariate dataset being from a Gaussian copula; otherwise, it is not appropriate to use the Gaussian copula thus NPBN to model the multivariate dataset.

The rank correlation matrix associated with a saturated NPBN (i.e. an NPBN in which each node is connected with all the other nodes) coincides with Σ_N. However, the arcs corresponding to correlations of small magnitudes are considered to reflect the sampling jitter and should be eliminated from the NPBN. To model the multivariate dataset parsimoniously, one develops NPBN by adding arcs between random variables such that the rank correlations between them are the greatest among the elements in Σ_N. The rank correlation matrix associated with the developed NPBN model is denoted by Σ_M. To validate that the developed NPBN has captured the significant dependences implicated in the multivariate dataset, a statistical test described as follows is carried out.

2.1) Build a skeletal NPBN, which involves only the arcs representing causal relationships between nodes; the elements in the linear correlation matrix evaluated in step 1.2) are used to compute the partial correlations associated with the arcs as per Eq. (4.4); the evaluated partial correlations are then transformed to conditional rank correlations attached to the arcs of the NPBN using Eq. (4.3).

2.2) The *n*-variate normal distribution corresponding to the NPBN has zero mean values and linear correlation matrix that is evaluated from the non-zero partial correlations associated with the arcs and zero partial correlations implied by the missing arcs using Eq. (4.4) (see the example in Appendix B).

2.3) Generate *k* sets of samples from the *n*-variate normal distribution described in step 2.2); use these samples to evaluate the rank correlation matrix and its determinant.

2.4) Repeat step 2.3) 1000 times and thus generate 1000 samples of the determinant of the rank correlation matrix Σ_M.

2.5) If det(Σ_N) is within the 5-95 percentile range of the samples generated in step 2.4), the current NPBN is accepted; otherwise, go to step 2.6).

2.6) Find a pair of variables between which there is no arc present in the current NPBN and the corresponding rank correlation (i.e. the elements in Σ_N) is greater than any other pairs not present in the current NPBN; add the corresponding arc in the current NPBN, and repeat steps 2.2) through 2.5) until a satisfactory NPBN is found.

4.3 Formulation of the NPBN model to predict the corrosion depth

Let X_d and $\mathbf{X} = [X_1, X_2, \dots, X_s]^{\mathrm{T}}$ denote the corrosion depth and a vector containing a total of *s* predictor variables (i.e. pipeline age and soil parameters), respectively. The NPBN characterizes the cumulative distribution function (CDF) of X_d and **X** using the Gaussian copula as follows (Sklar, 1959),

$$C\left(F_{X_d}(x_d), F_{X_1}(x_1), \dots, F_{X_s}(x_s)\right) = \Phi_{\Sigma}\left(\Phi^{-1}\left(F_{X_d}(x_d)\right), \Phi^{-1}\left(F_{X_1}(x_1)\right), \dots, \Phi^{-1}\left(F_{X_s}(x_s)\right)\right) \quad (4.5)$$

The distribution of X_d conditional on observations of predictor variables X_i (i = 1, 2, …, s) can be derived by employing the property of the multivariate normal distribution as follows. The correlation matrix of the (s+1)-variate normal distribution is partitioned as follows,

$$\Sigma = \begin{bmatrix} 1 & \Sigma_{U_d,\mathbf{U}} \\ \Sigma_{U_d,\mathbf{U}}^{\mathrm{T}} & \Sigma_{\mathbf{U},\mathbf{U}} \end{bmatrix} \tag{4.6}$$

where U_d and $\mathbf{U}$ corresponds to X_d and $\mathbf{X}$, respectively, through the inverse normal transformation; $\Sigma_{U_d,\mathbf{U}}$ denotes the correlation between U_d and $\mathbf{U}$, and $\Sigma_{\mathbf{U},\mathbf{U}}$ denotes the correlation matrix of $\mathbf{U}$.

Let $\mathbf{x}_{\mathrm{e}} = \left[x_{1,\mathrm{e}}, \ldots, x_{s,\mathrm{e}}\right]^{\mathrm{T}}$ denote the evidence for the model updating (i.e. observations of the predictor variables $\mathbf{X}$). The normal variates transformed from $\mathbf{x}_{\mathrm{e}}$ are denoted by $\mathbf{u}_{\mathbf{e}} = \left[\Phi^{-1}\left(F_{X_1}\left(x_{1,\mathrm{e}}\right)\right), \ldots, \Phi^{-1}\left(F_{X_s}\left(x_{s,\mathrm{e}}\right)\right)\right]^{\mathrm{T}}$. The distribution of U_d conditional on the observation $\mathbf{u}_{\mathrm{e}}$ is a normal distribution with the mean value $\bar{m}$ and standard deviation $\bar{\sigma}$, denoted by $(U_d|\mathbf{u}_{\mathrm{e}}) \sim N(\bar{m}, \bar{\sigma})$, where

$$\bar{m} = \Sigma_{U_d,\mathbf{U}} \Sigma_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{u}_{\mathrm{e}} \tag{4.7}$$

$$\bar{\sigma} = 1 - \Sigma_{U_d,\mathbf{U}} \Sigma_{\mathbf{U},\mathbf{U}}^{-1} \Sigma_{U_d,\mathbf{U}}^{T} \tag{4.8}$$

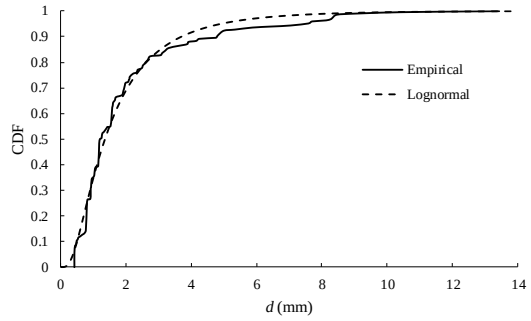
Then, the updated CDF of X_d is as follows,

$$F_{X_d}(x_d|\mathbf{x}_{\mathbf{e}}) = \Phi\left(\frac{\Phi^{-1}\left(F_{X_d}(x_d)\right) - \bar{m}}{\bar{\sigma}}\right) \tag{4.9}$$

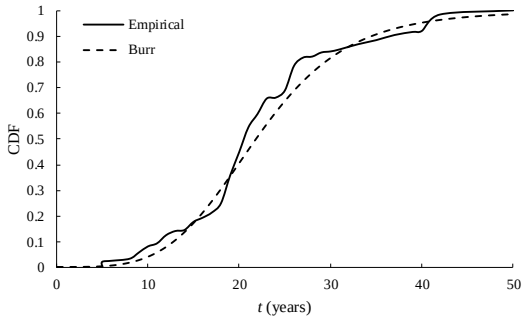
4.4 Overview of Velázquez's dataset

Velázquez's dataset consists of 259 samples of the maximum corrosion depth (*d*) together with the age of pipeline (*t*) and local soil parameters collected by excavating buried onshore pipelines in southern Mexico and carrying out field measurements (Velázquez et al., 2010). The maximum corrosion depth is the deepest corrosion-caused metal loss on the pipeline segment exposed in the excavation site (Velázquez et al., 2010), which will be simply called the corrosion depth hereafter. Detailed information about the data collection process (e.g. the length of each excavation site and number of measurements of the corrosion depth at each site) is however unavailable. Each sample consists of values of nine soil parameters including pH value (pH), pipe-to-soil potential (*pp*), soil resistivity (*re*), water content (*wc*),

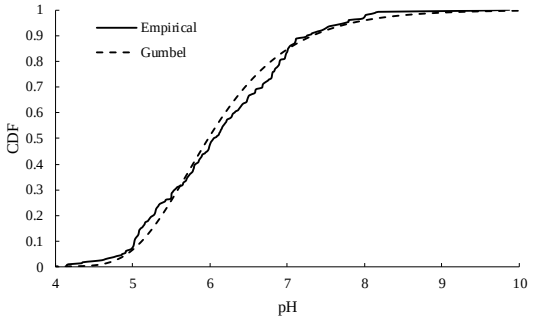
bulk density (*bd*), dissolved chloride (*cc*), bicarbonate (*bc*), sulfate ion concentrations (*sc*) and redox potential (*rp*). Velázquez et al. (2009) indicated that nine samples in the dataset are outliers with respect to the overall pattern of the data distribution. After the removal of these outliers, 250 samples are used in the present study, which belong to six soil types: namely clay (107 samples), sandy clay loam (75 samples), clay loam (59 samples), silty clay loam (6 samples), silty clay (2 samples) and silt loam (1 sample). Figure 4.2 depicts the empirical CDFs and CDFs of best-fit parametric distributions for the corrosion depth and predictor variables (i.e. pipeline age and nine soil parameters). The CDF and probability density function (PDF) of the Burr distribution shown in Fig. 4.2(b) is given in Appendix C. The statistics of *d* and predictor variables based on samples of the entire dataset and three soil types with reasonably large sample sizes (i.e. clay, sandy clay loam, and clay loam) are shown in Table 4.1.



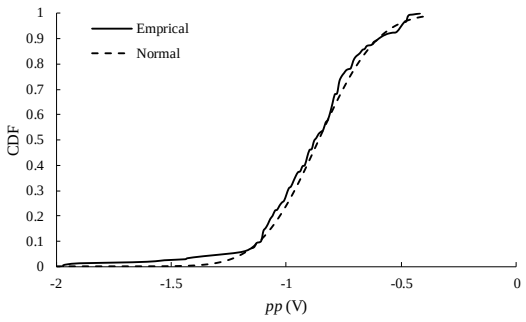
(a) *d*



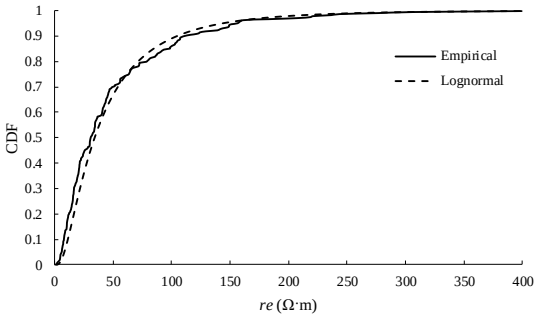
(b) *t*



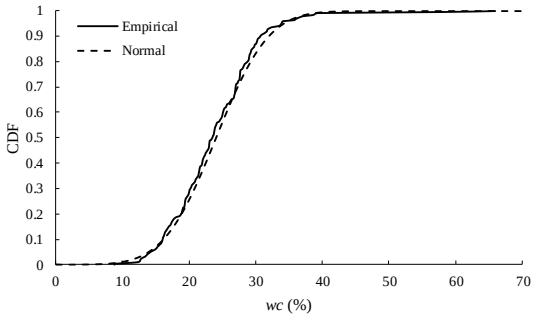
(c) pH



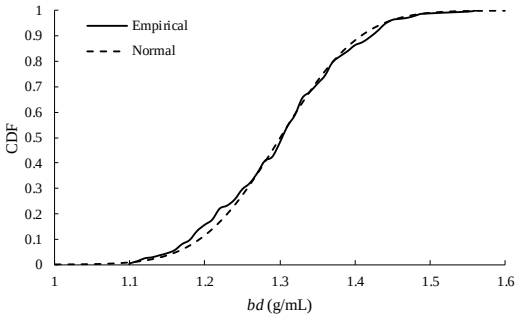
(d) *pp*



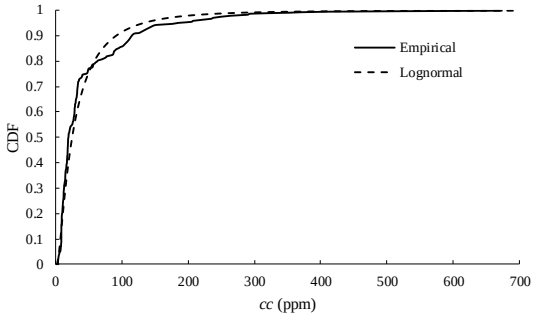
(e) *re*



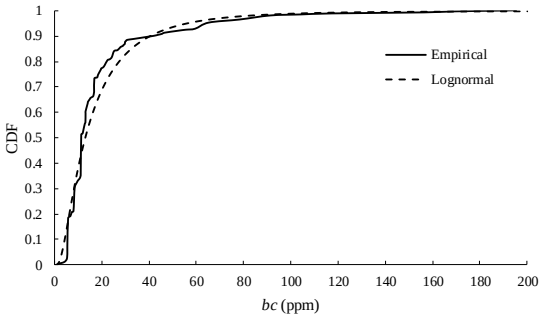
(f) *wc*



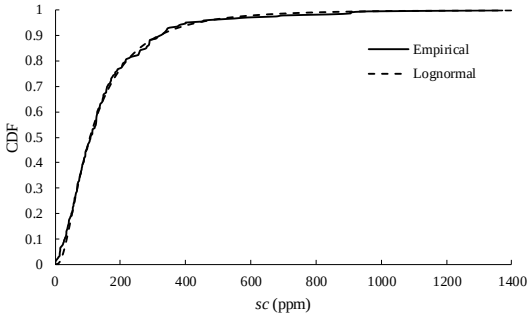
(g) *bd*



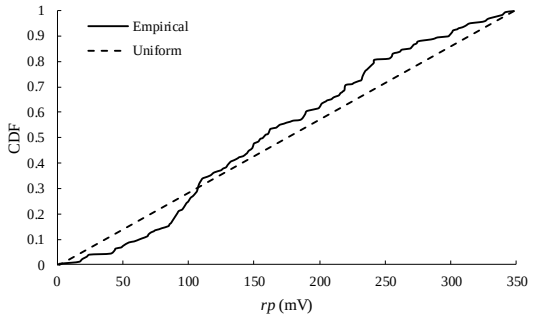
(h) *cc*



(i) *bc*



(j) *sc*



(k) *rp*

Figure 4.2 CDFs of the random variables in Velázquez's dataset

Table 4.1 Statistics of variables involved in Velázquez's dataset

Variable	Entire dataset		Clay		Sandy Clay Loam		Clay Loam	
	Mean	COV (%)	Mean	COV (%)	Mean	COV (%)	Mean	COV (%)
d (mm)	1.92	95	2.34	88	1.25	80	2.03	100
t (years)	23.01	39	24.45	35	18.91	36	24.63	43
pH	6.11	15	5.93	17	6.24	13	6.34	14
pp (mV)	-0.87	27	-0.86	28	-0.95	24	-0.82	26
re (Ω·m)	49.81	109	61.10	107	49.24	99	28.17	84
wc (%)	23.69	26	24.06	28	22.42	26	24.80	21
bd (g/mL)	1.30	6.6	1.23	4.2	1.40	3.4	1.32	1.7
cc (ppm)	41.91	139	53.09	128	21.82	108	44.61	121
bc (ppm)	18.25	115	19.26	130	13.77	44	22.85	102
sc (ppm)	148.70	106	129.33	87	143.76	69	205.15	124
rp (mV)	168.39	51	177.45	50	169.48	56	158	44

The empirical rank correlation matrix, Σ_E, and empirical normal rank correlation matrix, Σ_N associated with the dataset are shown in Tables 4.2 and 4.3, respectively. These tables indicate that the empirical rank correlations between *d* and predictor variables range from 0.07 to 0.41, which represents weak to moderate correlations. Among all the predictor variables, the pH value, dissolved chloride, pipeline age, bulk density, pipe-to-soil potential and water content have relatively strong correlations with the corrosion depth. While weak correlations between the corrosion depth and predictors such as the resistivity, sulfate content and redox potential suggest that direct influences of these predictors on the

corrosion depth are insignificant, the relatively strong correlations between these predictors and other predictors (i.e. *pp* and *re*, *wc* and *rp*) imply their indirect influences on the corrosion depth.

Table 4.2 Empirical rank correlation matrix (i.e. Σ_E) of variables involved in Velázquez's dataset

	d	*t*	pH	*pp*	*re*	*wc*	*bd*	*cc*	*bc*	*sc*	*rp*
d	1.00	0.39	-0.41	0.33	-0.15	0.30	-0.37	0.40	-0.28	-0.14	-0.07
t		1.00	0.15	0.32	-0.20	0.21	-0.36	0.19	0.03	-0.06	-0.06
pH			1.00	0.04	-0.29	0.16	0.21	-0.06	0.45	0.15	-0.25
pp				1.00	-0.36	0.27	-0.14	0.31	0.04	0.14	-0.25
re					1.00	-0.54	-0.08	-0.15	-0.19	-0.15	0.35
wc			Symmetric			1.00	-0.16	0.21	0.11	0.17	-0.29
bd							1.00	-0.23	0.17	0.17	-0.04
cc								1.00	0.10	0.00	-0.27
bc									1.00	0.27	-0.12
sc										1.00	-0.12
rp											1.00

Table 4.3 Empirical normal rank correlation matrix (i.e. Σ_N) of variables involved in Velázquez's dataset

	d	*t*	pH	*pp*	*re*	*wc*	*bd*	*cc*	*bc*	*sc*	*rp*
d	1.00	0.37	-0.38	0.33	-0.15	0.31	-0.37	0.36	-0.26	-0.13	-0.05
t		1.00	0.16	0.29	-0.11	0.22	-0.36	0.14	0.05	-0.03	-0.05
pH			1.00	0.03	-0.29	0.16	0.17	-0.01	0.42	0.17	-0.25
pp				1.00	-0.35	0.24	-0.13	0.27	0.05	0.15	-0.22
re					1.00	-0.54	-0.08	-0.13	-0.19	-0.19	0.33
wc			Symmetric			1.00	-0.17	0.18	0.11	0.16	-0.25
bd							1.00	-0.23	0.16	0.19	-0.07
cc								1.00	0.18	0.00	-0.23
bc									1.00	0.28	-0.10
sc										1.00	-0.11
rp											1.00

4.5 NPBN model development and validation using Velázquez's dataset

4.5.1 Model development

The model mining method described in Section 4.2.2 is implemented in the software UNINET® (Hanea, 2008) to develop an NPBN predictive model using Velázquez's dataset. First, the empirical marginal distributions for all the variables, empirical rank correlation matrix (i.e. Σ_E) and empirical normal rank correlation matrix (i.e. Σ_N) are evaluated. Since $\det(\Sigma_E) = 0.069$ is within the 5-95 percentile range of the generated samples of the determinant of the rank correlation matrix (see step 1.5) in Section 4.2.2), i.e. [0.045, 0.096], it is valid to assume that Velázquez's dataset is from a Gaussian copula and can be modeled by an NPBN.

The NPBN that models the dependence structure parsimoniously is developed as follows. Since the arcs representing the correlations between *d* and predictor variables are essential for the predictive model, the corresponding arcs are first added to develop the skeletal NPBN as shown in Fig. 4.3. Note that the correlations present on the arcs of the NPBN are the (conditional) rank correlations in the normal space. Since $\det(\Sigma_N) = 0.081$ is outside the 5-95 percentile range of generated samples of the determinant of the rank correlation matrix associated with the skeletal NPBN (see step 2.4) in Section 4.2.2), i.e. [0.31, 0.44], the skeletal NPBN shown in Fig. 4.3 is rejected. This suggests that the correlations between predictor variables should not be completely ignored; in other words, arcs between some of the predictor variables should be added to the NPBN. By following the procedure described in step 2.6) of Section 4.2.2, the NPBN shown in Fig. 4.4 is developed through a few iterations. Since $\det(\Sigma_N) = 0.081$ is within 5-95 percentile range of samples of the determinant of the rank correlation matrix associated with the NPBN in Fig. 4.4, it is considered a satisfactory model to represent Velázquez's dataset. The histogram characterizing the empirical marginal distribution, mean value and standard deviation associated with each node (expressed as mean ± standard deviation) are also shown in Fig. 4.4. In general, there is no best NPBN for modeling a multivariate dataset. The model developed in Fig. 4.4 only represents one valid NPBN to model Velázquez's dataset. The rank correlation matrix, Σ_M, associated with the NPBN in Fig. 4.4 is shown in Table 4.4.

Note that, in comparison to the regression model developed by Velázquez et al. (2009), the NPBN model takes into account the correlations between soil parameters, which is meaningful for predicting the corrosion depth under the missing information scenario, i.e. the values of part of the soil parameters are missing. The missing information scenario is however not considered in the present study.

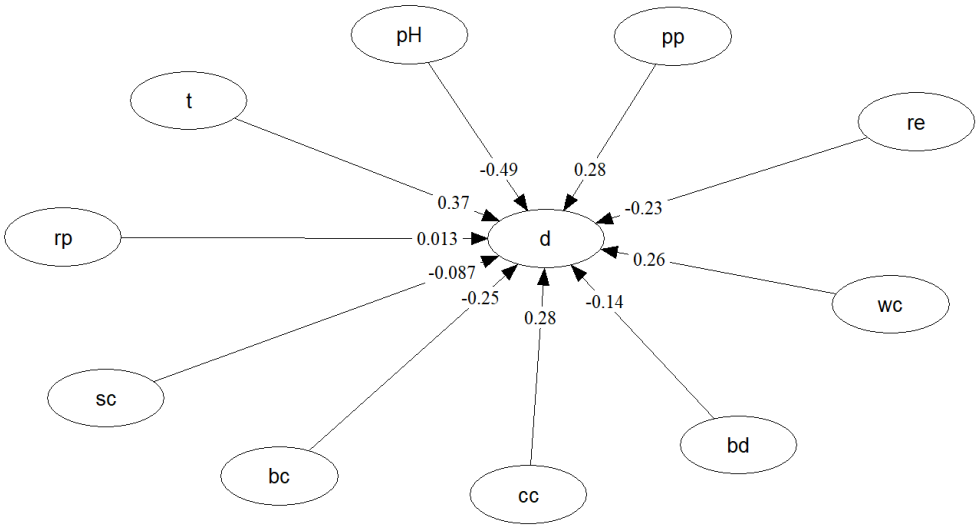


Figure 4.3 The skeletal NPBN

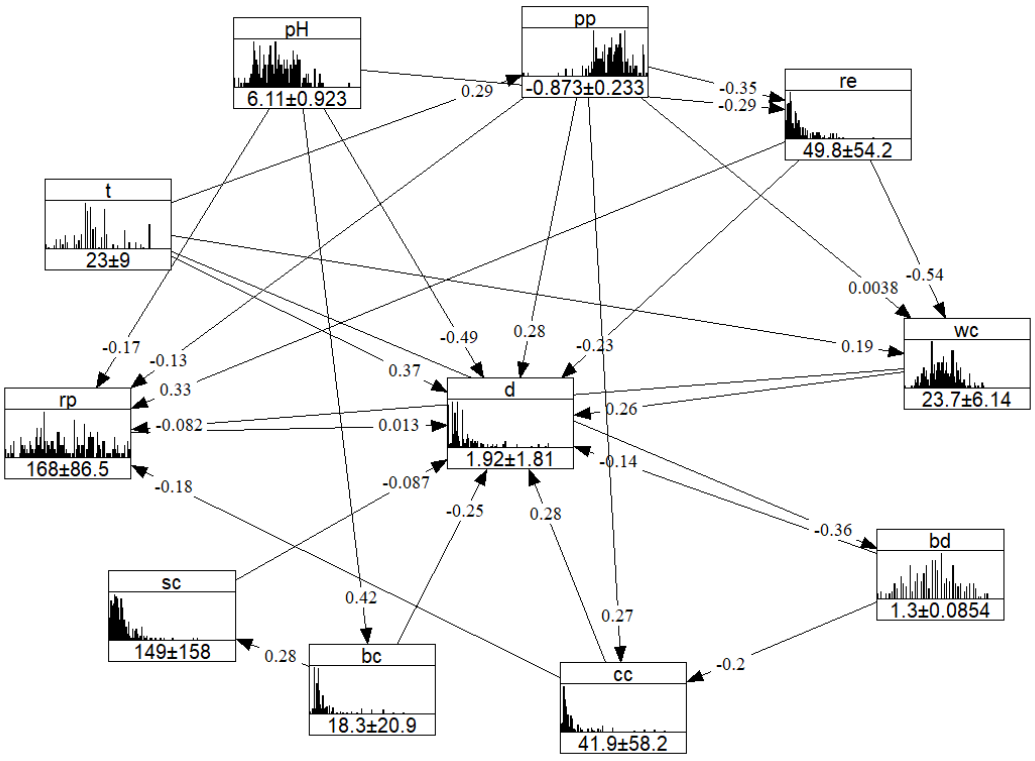


Figure 4.4 Final NPBN developed based on Velázquez's dataset

Table 4.4 Rank correlation matrix (i.e. Σ_M) associated with the NPBN in Fig. 4.4

	d	*t*	pH	*pp*	*re*	*wc*	*bd*	*cc*	*bc*	*sc*	*rp*
d	1.00	0.37	-0.45	0.33	-0.15	0.30	-0.24	0.32	-0.34	-0.09	-0.05
t		1.00	0	0.29	-0.10	0.22	-0.37	0.15	0	0	-0.09
pH			1.00	0	-0.27	0.15	0	0	0.42	0.12	-0.25
pp				1.00	-0.35	0.24	-0.11	0.27	0	0	-0.22
re					1.00	-0.54	0.04	-0.10	-0.12	-0.03	0.33
wc						1.00	-0.08	0.08	0.06	0.02	-0.25
bd			Symmetric				1.00	-0.22	0	0	0.06
cc								1.00	0	0	-0.22
bc									1.00	0.28	-0.11
sc										1.00	-0.03
rp											1.00

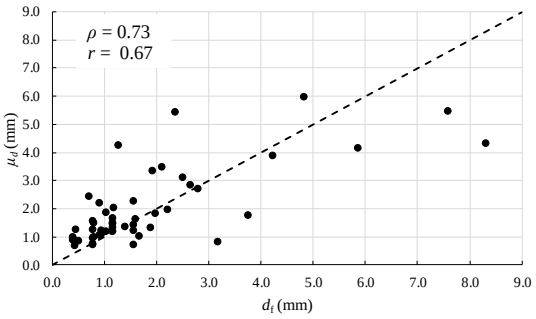
4.5.2 Model validation

The exhaustive 5-fold cross-validation (Kuhn and Johnson, 2013) is performed to examine the predictive capability of the developed NPBN. The entire dataset is divided into five sub-datasets of equal sample size, i.e. 50. The validation process includes five rounds. In

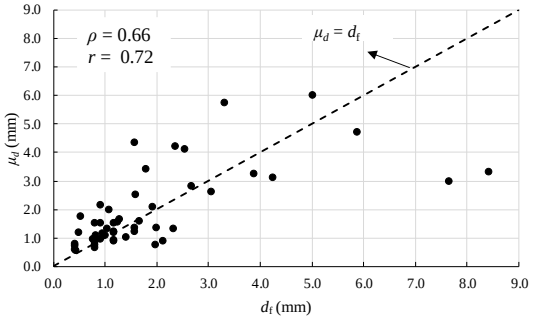
each round, four sub-datasets are assembled to be a training dataset (i.e. 200 samples) which is used to develop the NPBN. The remaining sub-dataset, referred to as the validation dataset, is used to examine the predictive capability of the developed NPBN. This approach ensures that the NPBN is developed and validated by two independent datasets. Moreover, since every sample in Velázquez's dataset is used both for training and validating the model, bias will be avoided in the predictive results. The arcs present in the NPBN developed in each round are the same as those present in the NPBN in Fig. 4.4, whereas the marginal distributions and correlation matrix vary slightly with different training datasets. For the developed NPBN, once the nodes denoting soil parameters and pipeline age are instantiated, the probabilistic distribution, mean value and standard deviation of the corrosion depth are inferred. Let μ_d and d_f denote the predicted mean value and field-measurement of the corrosion depth, respectively. μ_d and d_f associated with the samples in the five validation datasets are plotted in Figs. 4.5(a) through 4.5(e), respectively. The linear correlation, ρ, and rank correlation, r, between μ_d and d_f are also included in these figures.

Figures 4.5(a) through 4.5(e) indicate that the results associated with the five validation sets are similar. Most of the points distribute close to the line representing $\mu_d = d_f$, in particular, for relatively shallow corrosion, say less than 2 mm. As the corrosion depth increases, the scattering in the points increases. The better predictive accuracy for small corrosion depths may be explained by the fact that the majority (i.e. more than 80%) of corrosion depths in the entire Velazquez's dataset are less than 3 mm. While the prediction errors for some samples are large, the relatively strong correlation between μ_d and d_f indicates that the predicted mean values of the corrosion depths in general agree well with the corresponding field-measured values. The scattering in the predictions in Fig. 4.5 may be attributed to the following reasons. First, Velázquez's dataset does not capture the spatial variability associated with soil properties. Soil properties are in general heterogeneous over the length of an excavation site (Ricker, 2010). However, the soil properties of an excavation site are characterized by a single set of parameters in Velázquez's dataset. Therefore, differences may exist between the recorded soil parameters and those of the soil to which the field-measured corrosion depth is exposed. Second, the temporal variability of soil parameters is not considered. The soil parameters

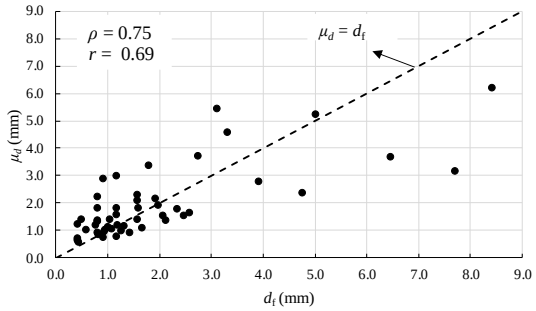
recorded in the dataset only reflect the soil properties at the time of the field survey. However, some soil parameters could change over time, e.g. the water content and pipe-to-soil potential.



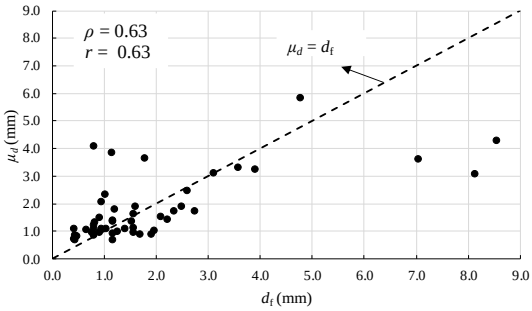
(a) Validation dataset 1



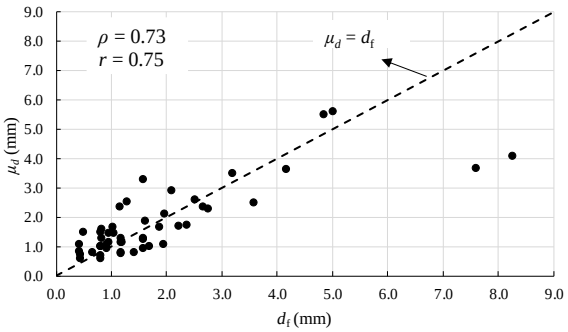
(b) Validation dataset 2



(c) Validation dataset 3



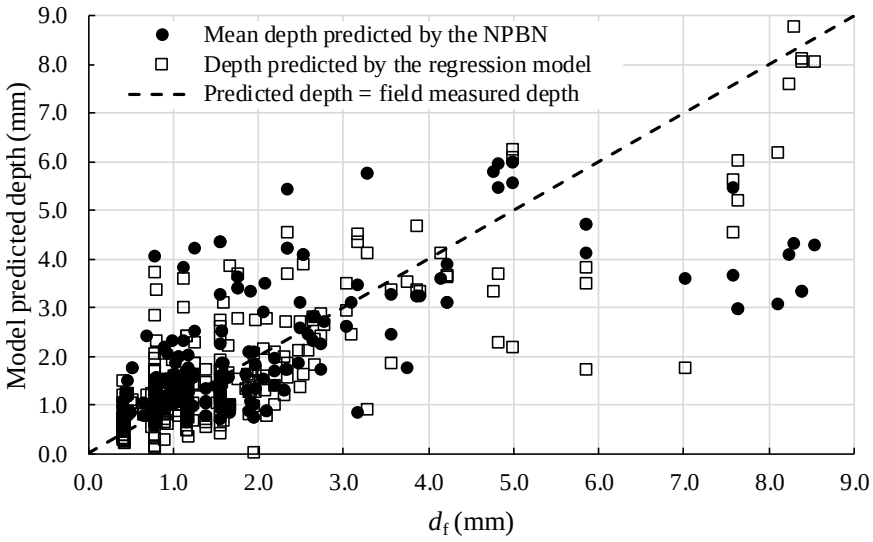
(d) Validation dataset 4



(e) Validation dataset 5

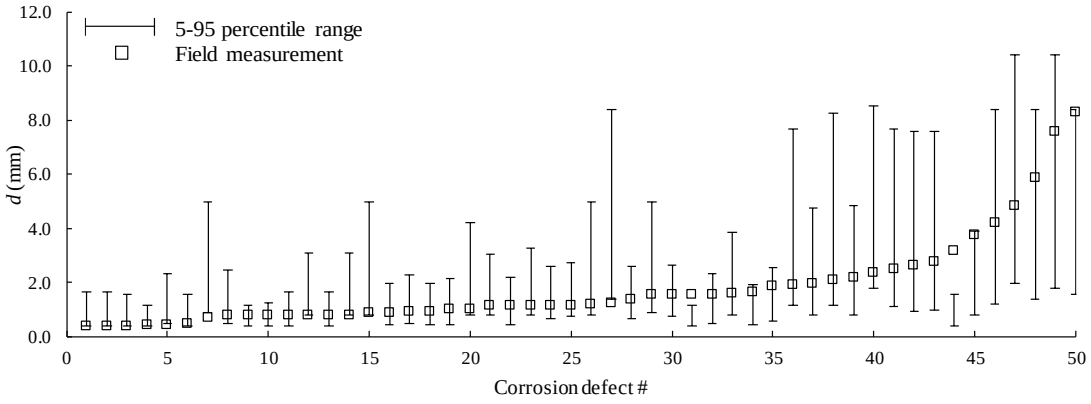
Figure 4.5 Predicted mean values and field-measurements of corrosion depth in Velázquez's dataset in the 5-fold cross-validation

The predictive accuracy of the NPBN is compared with the regression model developed by Velázquez et al. (2009). Figure 4.6 depicts the field-measured depths, mean depths predicted by the NPBN model (i.e. results in Fig. 4.5), and corrosion depths predicted by the regression model developed by Velázquez et al. (2009) using the entire Velázquez's dataset. The figure indicates that differences in the predictive accuracies of the two models are slight for corrosion depths less than 6 mm, whereas the regression model outperforms the NPBN model for extremely deep corrosion defects, say corrosion depths greater than 7 mm. However, the NPBN predictive model is advantageous over the regression models in that the probabilistic distribution of the corrosion depth can be predicted. The point estimate (i.e. predicted mean value) together with the 5-95 percentile range can characterize the uncertainty associated with the prediction. Figures 4.7(a) through 4.7(e) depict the field-measurements and 5-95 percentile ranges for the samples in the five validation datasets, respectively. These figures indicate that more than 95% of the field-measured corrosion depths fall in the 5-95 percentile range of the predictions. To be conservative, appropriate percentile values of the prediction may be used as the point estimate of the corrosion depth.

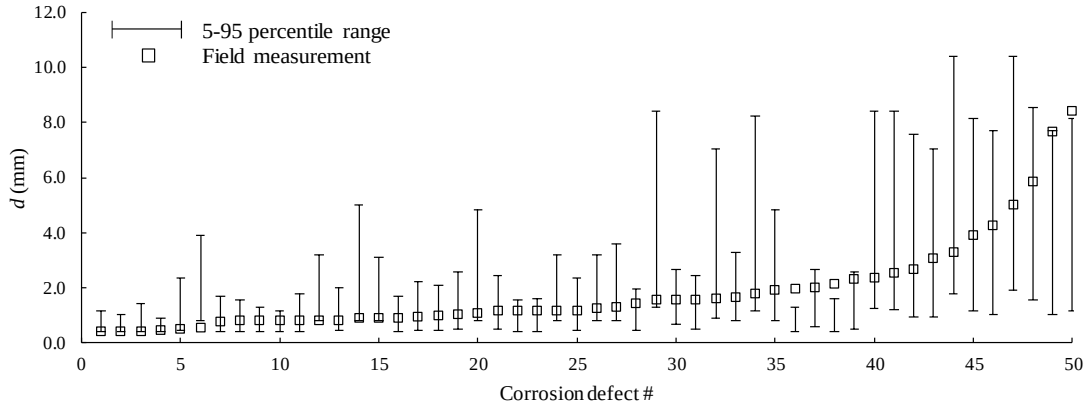


.

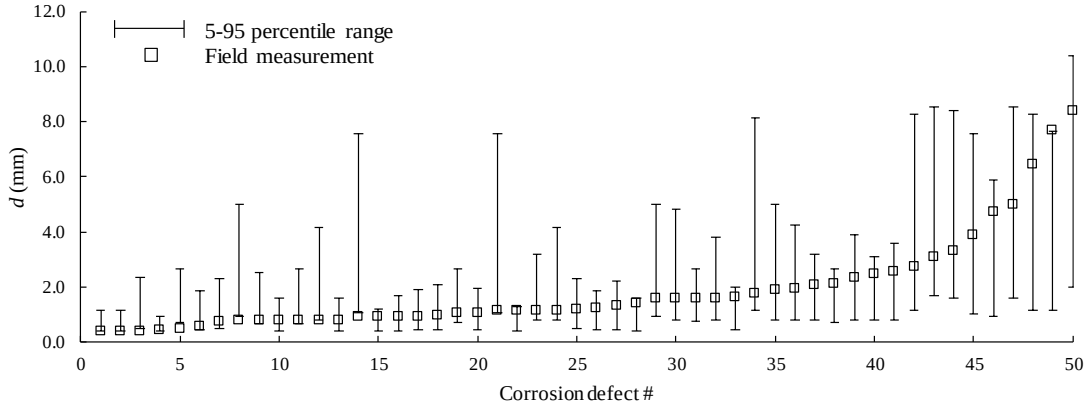
Figure 4.6 Comparison of predictions by the NPBN model and regression model developed by Velázquez et al. (2009) based on the entire Velázquez's dataset



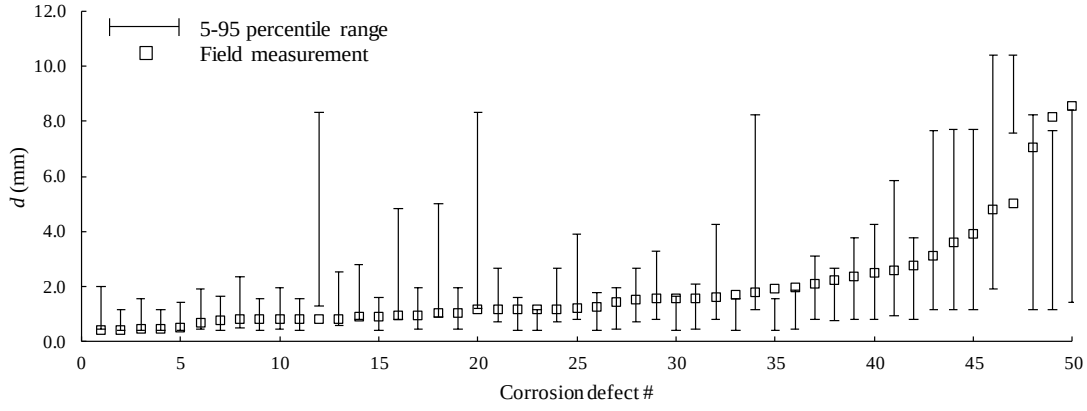
(a) Validation dataset 1



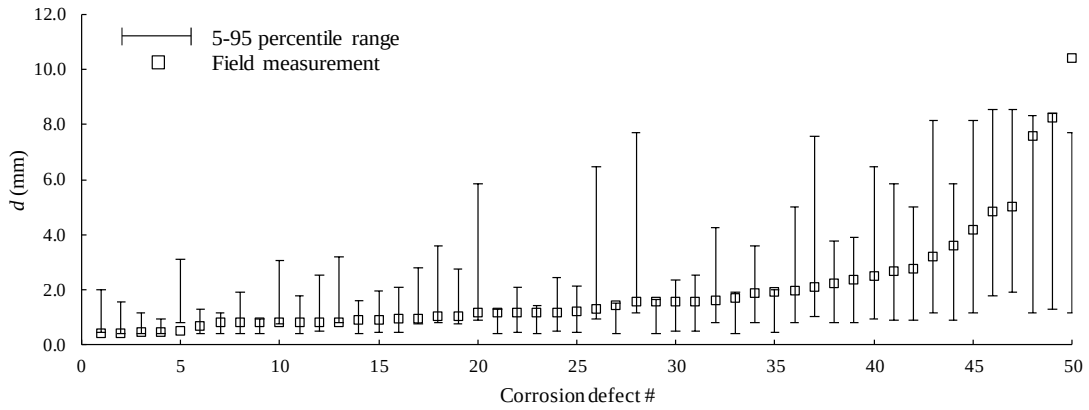
(b) Validation dataset 2



(c) Validation dataset 3



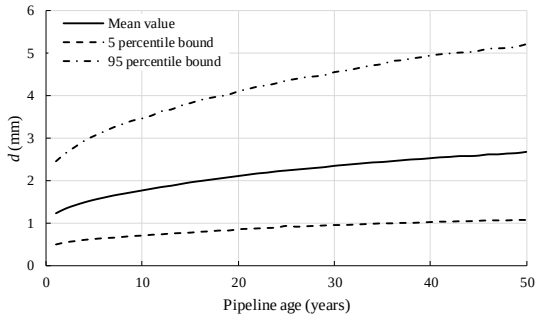
(d) Validation dataset 4



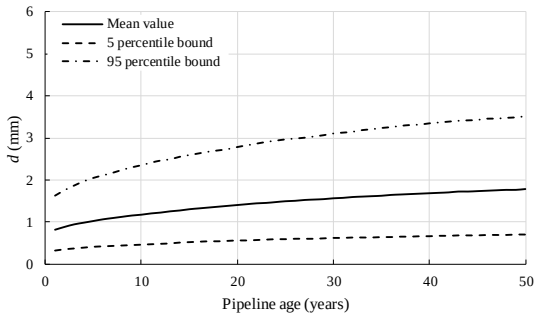
(e) Validation dataset 5

Figure 4.7 5-95 percentile ranges of predicted corrosion depths and field measurements

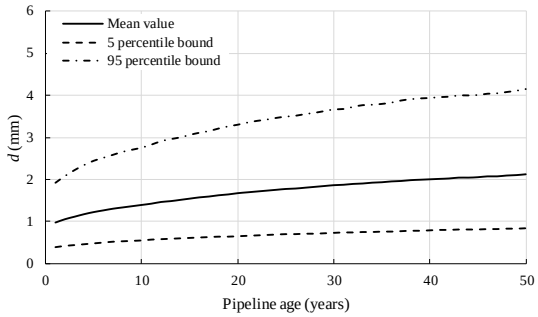
The NPBN model can be used to predict the corrosion depth on pipelines buried in different types of soil, which provides the basis to compare the corrosivity of different soil types. To create smooth corrosion growth paths, the parametric marginal distributions shown in Fig. 4.2 are used to replace the empirical marginals in the NPBN, whereas the dependence structure established in Section 4.5.1 remains. Consider the three representative soil types with reasonably large sample sizes involved in Velázquez's dataset (i.e. clay, sandy clay loam and clay loam), and use the mean values of corresponding soil parameters given by Table 4.1 to instantiate the NPBN. Figures 4.8(a) through 4.8(c) depict the predicted mean values and 5-95 percentile ranges of the predicted corrosion depths over a 50-year period. These figures indicate that the corrosivity of clay is the highest, followed by that of clay loam and sandy clay loam. This is consistent with the observation in the literature (Jyrkama et al., 2016; Velázquez et al., 2009).



(a) Clay



(b) Sandy clay loam



(c) Clay loam

Figure 4.8 Predicted corrosion depths for clay, sandy clay loam and clay loam using NPBN with parametric marginal distributions

4.6 Conclusions

The present study employs the NPBN technique to develop a predictive model for the corrosion depth on underground pipelines based on Velázquez's dataset, which consists of values of the corrosion depth, pipeline age and nine parameters of surrounding soils from 250 excavation sites in southern Mexico. While the empirical rank correlations indicate that only pH value, dissolved chloride, bulk density, water content and pipe-to-soil potential have relatively strong correlations with the corrosion depth, the nine soil parameters are all involved in the NPBN as predictors due to the correlations between the soil parameters themselves. Taking into account the correlations between predictors enables the NPBN model to predict the corrosion depth under missing information scenario, i.e. the values of part of the soil parameters are missing. In comparison with the

regression models, the NPBN can quantify the probabilistic distribution of the corrosion depth.

The results of the 5-fold cross-validation indicate that the predicted mean corrosion depths in general agree well with the field measurements, and more than 95% field measurements fall in the 5-95 percentile range of the predicted distributions. Moreover, the analysis based on the NPBN model indicates that, among the three representative soil types in Velázquez's dataset, the corrosivity of clay is the highest followed by that of clay loam and sandy clay loam. The present study demonstrates that the NPBN and associated model mining method provide an effective means of developing probabilistic predictive models for the corrosion depth using soil parameters as predictors. This has significant practical implications in terms of the integrity management of unpiggable pipelines with respect to corrosion.

References

Alamilla, J. L., Espinosa-Medina, M. A., and Sosa, E. (2009). Modelling steel corrosion damage in soil environment. *Corrosion Science*, 51(11), 2628-2638.

Beauregard, Y., Woo, A., and Huang, T. (2018). Application of In-Line Inspection and Failure Data to Reduce Subjectivity of Risk Model Scores for Uninspected Pipelines. In: *Proceedings of the 12th International Pipeline Conference* (pp. V002T07A028-V002T07A028), Calgary, Alberta, Canada.

Caleyo, F., Velázquez, J. C., Valor, A., and Hallen, J. M. (2009). Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study. *Corrosion Science*, 51(9), 1925-1934.

Canadian Energy Pipeline Association (CEPA). (2015). *Committed to safety, committed to Canadians: 2015 pipeline performance report*. Calgary, Alberta: Canadian Energy Pipeline Association.

Hanea, A. M. (2008). Algorithms for non-parametric Bayesian belief nets (Doctoral dissertation). Delft University of Technology, Delft, Netherlands.

Hanea, A. M., Gheorghe, M., Hanea, R., and Ababei, D. (2013). Non-parametric Bayesian networks for parameter estimation in reservoir simulation: a graphical take on the ensemble Kalman filter (part I). *Computational geosciences*, 17(6), 929-949.

Hanea, A. M., Kurowicka, D., and Cooke, R. M. (2006). Hybrid method for quantifying and analyzing Bayesian belief nets. *Quality and Reliability Engineering International*, 22(6), 709-729.

Hanea, A. M., Kurowicka, D., Cooke, R. M., and Ababei, D. A. (2010). Mining and visualising ordinal data with non-parametric continuous BBNs. *Computational Statistics and Data Analysis*, 54(3), 668-687.

Hanea, A., Napoles, O. M., and Ababei, D. (2015). Non-parametric Bayesian networks: Improving theory and reviewing applications. *Reliability Engineering and System Safety*, 144, 265-284.

Jyrkama, M., Pandey, M., Angell, P., and Munson, D. (2016). Estimating External Corrosion Rates for Buried Carbon Steel Piping in Different Soil Conditions. *CNL Nuclear Review*, 7(1), 85-94.

Kuhn, M., and Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York: Springer.

Kurowicka, D., and Cooke, R. M. (2005). Distribution-free continuous Bayesian belief. *Modern statistical and mathematical methods in reliability*, 10, 309.

Kurowicka, D., and Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley and Sons.

Lam, C., and Zhou, W. (2016). Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database. *International Journal of Pressure Vessels and Piping*, 145, 29-40.

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

Lee, D., and Pan, R. (2018). A nonparametric Bayesian network approach to assessing system reliability at early design stages. *Reliability Engineering and System Safety*, 171, 57-66.

Melchers, R. E., and Petersen, R. B. (2018). A reinterpretation of the Romanoff NBS data for corrosion of steels in soils. *Corrosion Engineering, Science and Technology*, 53(2), 131-140.

Morales-Nápoles, O., Delgado-Hernández, D. J., De-León-Escobedo, D., and Arteaga-Arcos, J. C. (2014). A continuous Bayesian network for earth dams' risk assessment: methodology and quantification. *Structure and Infrastructure Engineering*, 10(5), 589-603.

Morales-Nápoles, O., and Steenbergen, R. D. (2014). Large-scale hybrid Bayesian network for traffic load modeling from weigh-in-motion system data. *Journal of Bridge Engineering*, 20(1), 04014059.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science and Business Media.

Nielsen, T., and Jensen, F. (2009). *Bayesian networks and decision graphs*. New York, NY: Springer Science and Business Media.

Pearl, J. (2014). Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier.

Pearson, K. (1907). Mathematical contributions to the theory of evolution. *Philosophical Transactions of the Royal Society of London.* 187, 253-318D.

Ricker, R. E. (2010). Analysis of pipeline steel corrosion data from NBS (NIST) studies conducted between 1922–1940 and relevance to pipeline management. *Journal of research of the National Institute of Standards and Technology*, 115(5), 373.

Romanoff, M. (1957). *Underground corrosion*. Washington (DC: US Government Printing Office.

Schwerdtfeger, W. J. (1966). Soil resistivity as related to underground corrosion and cathodic protection. *Highway Research Record*, (110).

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8, 229-231.

Velázquez, J. C., Caleyo, F., Valor, A., and Hallen, J. M. (2009). Predictive model for pitting corrosion in buried oil and gas pipelines. *Corrosion*, 65(5), 332-342.

Velázquez, J. C., Caleyo, F., Valor, A., and Hallen, J. M. (2010). Field Study—Pitting Corrosion of Underground Pipelines Related to Local Soil and Pipe Characteristics. *Corrosion*, 66(1), 016001-016001.

Wang, F., Li, H., Dong, C., and Ding, L. (2019). Knowledge representation using non-parametric Bayesian networks for tunneling risk analysis. *Reliability Engineering and System Safety*, 106529.

Wang, H., Yajima, A., Liang, R. Y., and Castaneda, H. (2016). Reliability-based temporal and spatial maintenance strategy for integrity management of corroded underground pipelines. *Structure and Infrastructure Engineering*, 12(10), 1281-1294.

Yajima, A., Wang, H., Liang, R. Y., and Castaneda, H. (2015). A clustering based method to evaluate soil corrosivity for pipeline external integrity management. *International Journal of Pressure Vessels and Piping*, 126, 37-47.

Zeng, B., Chen, K., and Wang, C. (2017). Geometric views of partial correlation coefficient in regression analysis. *International Journal of Statistics and Probabilities*, 6(3), 51-60

Zilko, A. A., Kurowicka, D., and Goverde, R. M. (2016). Modeling railway disruption lengths with Copula Bayesian Networks. *Transportation Research Part C: Emerging Technologies*, 68, 350-368.

5 Optimal sample size determination based on Bayesian reliability and value of information

5.1 Introduction

The structural reliability analysis of engineering structures generally involves estimating the failure probability, P_f, as follows,

$$P_f = \int_{\Omega_f} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \tag{5.1}$$

where $f_{\mathbf{X}}(\mathbf{x})$ denotes the joint probability density function (PDF) of a vector of basic random variables $\mathbf{X}$ such as dimensions of the structural members, material properties and magnitudes of loads, and Ω_f denotes the failure domain that is typically defined through one or more so-called limit state functions. The integral in Eq. (5.1) can be evaluated using, for example, the simple Monte Carlo (MC) simulation (Melchers and Beck, 2018), important sampling-based MC simulation (Melchers and Beck, 2018) and first-order reliability method (FORM) (Der Kiureghian, 2005; Zhou et al., 2017). Since $f_{\mathbf{X}}(\mathbf{x})$ is often elicited from imperfect information such as expert opinions and databases with limited sample sizes, there are epistemic uncertainties associated with $f_{\mathbf{X}}(\mathbf{x})$. The epistemic uncertainties can be taken into account in the analysis by considering the distribution parameters of basic random variables to be uncertain (Der Kiureghian 1989; Der Kiureghian 2008; Der Kiureghian and Ditlevsen 2009; Hong 1996). This introduces uncertainty in P_f, which may affect the decision making based on P_f. It is therefore desirable to gather sufficient samples of $\mathbf{X}$ to reduce the uncertainties in $f_{\mathbf{X}}(\mathbf{x})$. The determination of appropriate sample sizes for $\mathbf{X}$ is a challenging yet often-encountered task in the design and assessment of engineering structures; for instance, gathering soil property data in the design of foundations (Goldswarthy, 2007), proof-load testing quasi-identical multi-components structural systems (Nishijima and Faber, 2007; Shafieezadeh and Ellingwood, 2012), collecting corrosion defect data for the integrity management of buried oil and gas pipelines (Caleyo et al., 2015) and measuring the wall thickness of deteriorating piping systems in nuclear reactors (Higo and Pandey, 2016). Since the cost of sampling is in general high, the sample size should be determined by balancing the cost and associated benefit. This is known as the problem of the sample size determination (SSD).

The Bayesian pre-posterior analysis (Raiffa and Schlaifer, 1961) is a viable approach to deal with SSD. Pham and Turkkan (1992) employed the pre-posterior analysis to study SSD for the parameter of the binomial distribution. Assuming the parameter to have a beta prior distribution and exploiting the conjugacy of the beta-binomial pair, the authors derived analytical expressions for the expectations of the posterior mean and variance of the binomial parameter with respect to the outcome of sampling with given sample size. The appropriate sample size can then be determined by using one of three criteria: limiting the maximum posterior variance and Bayes risk to pre-determined allowable values, respectively, and maximizing the expected net gain of sampling (ENGS). Adcock (1992) extended Pham and Turkkan's approach to investigating SSD for parameters of the multinomial distribution by assuming the prior distribution of the parameters to be the Dirichlet distribution and utilizing the conjugacy of the Dirichlet-multinomial pair. Based on the pre-posterior analysis and value of information (VoI) concept, Higo and Pandey (2016) derived an analytical expression for the optimal number of wall thickness measurements for nuclear piping systems by assuming the wall thickness to follow a normal distribution. The aforementioned studies address SSD for parameters of specific distributions; however, there is a lack of a general framework that can deal with SSD for a wide range of probability distributions by considering the impact of uncertainties in $f_{\mathbf{X}}(\mathbf{x})$ on P_f.

In this study, a novel methodology that is based on the Bayesian pre-posterior analysis of P_f is developed to deal with SSD. The methodology starts by discretizing the basic variables for which sample sizes need to be determined. The probability mass functions (PMFs) of the discretized variables are then assigned Dirichlet prior distributions. The total probability theorem is employed to express P_f in terms of PMFs of the discretized variables and conditional failure probabilities corresponding to given values of discretized variables. This facilitates the pre-posterior analysis of P_f based on those of the discretized variables. Based on the pre-posterior analysis of P_f and theory of value of information (VoI) (Raiffa and Schlaifer, 1961), a criterion for determining the optimal sample sizes to maximize ENGS is established. Since the Dirichlet distribution can be assigned to the PMF of the random variable with any distribution type, the methodology is applicable to

different probability distributions of the basic variables for which sample sizes need to be determined.

The remainder of this chapter is organized as follows. Section 5.2 provides the formulation of pre-posterior analysis of the PMF of a random variable and P_f. Section 5.3 establishes the SSD criterion based on the quadratic loss function. Two examples of SSD concerning the corrosion assessment of energy pipelines are included in Section 5.4 to demonstrate the SSD results. Moreover, the sensitivity of the SSD results to the discretization of the continuous random variables and equivalent sample size of the prior Dirichlet distribution is also studied in the numerical examples. The chapter is concluded in Section 5.5.

5.2 Pre-posterior analysis

5.2.1 Pre-posterior analysis of PMF

Let Y denote a discrete random variable with m states y_j (j = 1, 2, …, m). The PMF of Y is represented by an m-dimensional vector $\mathbf{W}_Y = \{W_{Y,1}, W_{Y,2}, \ldots, W_{Y,m}\}$ with $\sum_{j=1}^{m} W_{Y,j} = 1$. To model the epistemic uncertainty in the distribution of Y, $\mathbf{W}_Y$ is considered uncertain and hence a random vector. The Dirichlet distribution is often assigned as the prior distribution of uncertain PMFs in the literature concerning the parameter learning of Bayesian networks (Spiegelhalter et al., 1993); that is, $\mathbf{W}_Y \sim \text{Dir}(\boldsymbol{\alpha}_Y)$, where "~" denotes the assignment of a probability distribution, and $\boldsymbol{\alpha}_Y = \{\alpha_{Y,1}, \alpha_{Y,2}, \ldots, \alpha_{Y,m}\}$ is the m-dimensional parameter vector of the Dirichlet distribution. The prior joint PDF of $\mathbf{W}_Y$, $f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y)$, is given by (Jonson and Kotz, 1972),

$$f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y) = \frac{\Gamma(\alpha_{Y0})}{\prod_{j=1}^{m}\Gamma(\alpha_{Y,j})}\prod_{j=1}^{m}(w_{Y,j})^{\alpha_{Y,j}-1} \quad (0 < w_{Y,j} < 1 \text{ and } \alpha_{Y,j} > 0;\ j = 1, 2, \ldots, m) \quad (5.2)$$

where $\mathbf{w}_Y = \{w_{Y,1}, w_{Y,2}, \ldots, w_{Y,m}\}$ is the value of $\mathbf{W}_Y$; $\Gamma(\bullet)$ is the gamma function, and $\alpha_{Y0} = \sum_{j=1}^{m} \alpha_{Y,j}$ is commonly known as the equivalent sample size of the Dirichlet distribution.

The prior mean and variance of $W_{Y,j}$ (j = 1, 2, …, m), $\mu_{W_{Y,j}}^{\pi}$ and $\xi_{W_{Y,j}}^{\pi}$, respectively, are given by,

$$\mu_{W_{Y,j}}^{\pi} = \frac{\alpha_{Y,j}}{\alpha_{Y0}} \tag{5.3}$$

$$\xi_{W_{Y,j}}^{\pi} = \frac{\alpha_{Y,j}(\alpha_{Y0}-\alpha_{Y,j})}{(\alpha_{Y0})^2(\alpha_{Y0}+1)} \tag{5.4}$$

Throughout the chapter, the symbols $\mu_{\bullet}$ and $\xi_{\bullet}$ are used to denote the mean and variance of a random variable $\bullet$, respectively, whereas superscripts π and p are used to denote prior and posterior statistics, respectively. Note that $W_{Y,j}$ and $W_{Y,k}$ ($j, k = 1, 2, \ldots, m; j \neq k$) are correlated with the corresponding covariance, $\omega_{W_{Y,jk}}^{\pi}$, given by,

$$\omega_{W_{Y,jk}}^{\pi} = \frac{-\alpha_{Y,j}\alpha_{Y,k}}{(\alpha_{Y0})^2(\alpha_{Y0}+1)} \ (j \neq k) \tag{5.5}$$

It follows from Eq. (5.5) that any two components in the Dirichlet distribution are negatively correlated, which directly results from the fact that $\sum_{j=1}^{m} W_{Y,j} = 1$. This simple correlation structure is a limitation of the Dirichlet distribution (Caballero el al., 2012).

Now suppose that a set of samples $\mathbf{n}_Y = \{n_{Y,1}, n_{Y,2}, \ldots, n_{Y,m}\}$ are obtained from the outcome space of Y, where $n_{Y,j}$ ($n_{Y,j} \geq 0$; $j = 1, 2, \ldots, m$) represents the number of samples lying in the j-th state. These samples can be used to update the prior distribution of $\mathbf{W}_Y$. The likelihood of $\mathbf{n}_Y$, $L(\mathbf{w}_Y|\mathbf{n}_Y)$, is of the multinomial form as follows,

$$L(\mathbf{w_Y}|\mathbf{n_Y}) = \frac{n_{Y0}!}{\prod_{j=1}^{m} n_{Y,j}!}\prod_{j=1}^{m}(w_{Y,j})^{n_{Y,j}} \tag{5.6}$$

where $n_{Y0} = \sum_{j=1}^{m} n_{Y,j}$, i.e. the total number of samples. Given the conjugacy between the multinomial and Dirichlet distributions, the posterior distribution of $\mathbf{W}_Y$ is also the Dirichlet distribution with the corresponding PDF, $f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y, \mathbf{n}_Y)$, given by (Jonson and Kotz, 1972),

$$f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y, \mathbf{n}_Y) = \frac{\Gamma(\alpha_{Y0}+n_{Y0})}{\prod_{j=1}^{m}\Gamma(\alpha_{Y,j}+n_{Y,j})}\prod_{j=1}^{m}(w_{Y,j})^{\alpha_{Y,j}+n_{Y,j}-1} \tag{5.7}$$

It follows that the parameter vector of the posterior Dirichlet distribution of $\mathbf{W}_Y$ is ($\boldsymbol{\alpha}_Y$ + $\mathbf{n}_Y$). The posterior mean, variance and covariance of $\mathbf{W}_Y$ are then given by,

$$\mu^{p}_{W_{Y,j}} = \frac{\alpha_{Y,j}+n_{Y,j}}{\alpha_{Y0}+n_{Y0}} \tag{5.8}$$

$$\xi^{p}_{W_{Y,j}} = \frac{(\alpha_{Y,j}+n_{Y,j})(\alpha_{Y0}+n_{Y0}-\alpha_{Y,j}-n_{Y,j})}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)} \tag{5.9}$$

$$\omega^{p}_{W_{Y,jk}} = \frac{-(\alpha_{Y,j}+n_{Y,j})(\alpha_{Y,k}+n_{Y,k})}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)} (j \neq k) \tag{5.10}$$

A comparison of statistics of the prior Dirichlet distribution (Eqs. (5.3) through (5.5)) and those of the posterior Dirichlet distribution (Eqs (5.8) through (5.10)) suggests an intuitive interpretation of parameters of the prior Dirichlet distribution: $\alpha_{Y,j}$ is the equivalent (or pseudo) sample count that lie in the j-th state, and α_{Y0} is the total number of equivalent sample count. The values of $\alpha_{Y,j}$ and α_{Y0} relative to $n_{Y,j}$ and n_{Y0} reflect the weight or importance of the prior belief. Note that the conjugacy between the Dirichlet and multinomial distributions has been exploited extensively in the parameter learning associated with the Bayesian network (Feelders and van der Gaag, 2006; Heckerman et al., 1998; Masegosa et al., 2016; Spiegelhalter et al., 1993; Zhou et al., 2016).

If a decision is made to draw a total of n_{Y0} samples but the actual sampling process has not been carried out, the potential sample count in the j-th state (j = 1, 2, …, m) is now a random variable, denoted by $N_{Y,j}$. The posterior statistics of $\mathbf{W}_Y$ then depend on the realization of the random vector $\mathbf{N}_Y = \{N_{Y,1}, N_{Y,2}, \ldots, N_{Y,m}\}$. This is the pre-posterior analysis (Raiffa and Schlaifer, 1961). The marginal (or compound) distribution of $\mathbf{N}_Y$ is the so-called Dirichlet-multinomial distribution, with the corresponding PDF, $f(\mathbf{n}_Y|\boldsymbol{\alpha}_Y)$, given by (Johnson and Kotz, 1972),

$$f(\mathbf{n}_Y|\boldsymbol{\alpha}_Y) = \frac{\Gamma(n_{Y0}+1)\Gamma(\alpha_{Y0})}{\Gamma(n_{Y0}+\alpha_{Y0})} \prod_{j=1}^{m} \frac{\Gamma(\alpha_{Y,j}+n_{Y,j})}{\Gamma(n_{Y,j}+1)\Gamma(\alpha_{Y,j})} \tag{5.11}$$

The mean value and variance of $N_{Y,j}$ are,

$$\mu_{N_{Y,j}} = n_{Y0} \frac{\alpha_{Y,j}}{\alpha_{Y0}} \tag{5.12}$$

$$\xi_{N_{Y,j}} = n_{Y0} \frac{\alpha_{Y,j}}{\alpha_{Y0}} \left(1 - \frac{\alpha_{Y,j}}{\alpha_{Y0}}\right) \left(\frac{\alpha_{Y0}+n_{Y0}}{1+\alpha_{Y0}}\right) \tag{5.13}$$

Furthermore, the covariance of $N_{Y,j}$ and $N_{Y,k}$ ($j, k = 1, 2, \ldots, m; j \neq k$) is given by,

$$\omega_{N_{Y,jk}} = -n_{Y0}\frac{\alpha_{Y,j}\alpha_{Y,k}}{(\alpha_{Y0})^2}\left(\frac{\alpha_{Y0}+n_{Y0}}{1+\alpha_{Y0}}\right) \tag{5.14}$$

Replacing $n_{Y,j}$ and $n_{Y,k}$ in Eqs. (5.8) - (5.10) with $N_{Y,j}$ and $N_{Y,k}$, respectively, one can evaluate the expectations of the posterior mean, variance and covariance of $\mathbf{W}_Y$ with respect to $\mathbf{N}_Y$, respectively, as follows,

$$\mathrm{E}_N\left[\mu^p_{W_{Y,j}}\right] = \frac{\alpha_{Y,j}}{\alpha_{Y0}} \tag{5.15}$$

$$\mathrm{E}_N\left[\xi^p_{W_{Y,j}}\right] = \frac{\alpha_{Y0}}{\alpha_{Y0}+n_{Y0}}\xi^{\pi}_{W_{Y,j}} \tag{5.16}$$

$$\mathrm{E}_N\left[\omega^p_{W_{Y,jk}}\right] = \frac{n_{Y0}\alpha_{Y,j}\alpha_{Y,k}(\alpha_{Y0}+n_{Y0})-\alpha_{Y,j}\alpha_{Y,k}(\alpha_{Y0})^2(\alpha_{Y0}+1)-(n_{Y0})^2\alpha_{Y,j}\alpha_{Y,k}(\alpha_{Y0}+1)-2n_{Y0}\alpha_{Y,j}\alpha_{Y,k}\alpha_{Y0}(\alpha_{Y0}+1)}{(\alpha_{Y0})^2(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+1)(\alpha_{Y0}+n_{Y0}+1)} \tag{5.17}$$

where $\mathrm{E}_N[\bullet]$ denotes the expectation with respect to $\mathbf{N}_Y$. Note that the expectation of the posterior mean (Eq. (5.15)) is the same as the prior mean (Eq. (5.3)). The derivations of Eqs. (5.15) through (5.17) are shown in Appendix D.

5.2.2 Pre-posterior analysis of P_f

Let $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_t\}$ ($t \geq 1$) denote a subset of random variables of $\mathbf{X}$, for which sampling is needed and the corresponding sample sizes need to be determined. In this study, Y_i ($i = 1, 2, \ldots, t$) is treated as a discrete random variable with m_i states; therefore, continuous random variables are discretized. Assuming Y_i ($i = 1, 2, \ldots, t$) to be mutually independent, one can rewrite Eq. (5.1) using the total probability theorem as follows,

$$P_f = \sum_{j=1}^{m} \Pr\left(\text{Failure}|\mathbf{Y} = \mathbf{y}_j\right)W_j \tag{5.18}$$

where $\mathbf{y_j} = \left\{y_{1,j_1}, y_{2,j_2}, \ldots, y_{t,j_t}\right\}$ denotes the j-th state of $\mathbf{Y}$; j_i ($i = 1, 2, \ldots, t$) varies from 1 to m_i; $m = \prod_{i=1}^{t} m_i$ denotes the number of states of $\mathbf{Y}$, and $W_j = \prod_{i=1}^{t} W_{i,j_i}$ denotes the PMF of the j-th state of $\mathbf{Y}$. Given that the PMF of $\mathbf{Y}$, W_j ($j = 1, 2, \ldots, m$), is considered as a random vector, Eq. (5.18) implies that P_f is also a random variable, for which the prior mean value and variance are given by Eqs. (5.19) and (5.20), respectively,

$$\mu_{P_f}^{\pi} = \sum_{j=1}^{m} p_{f,j}\mu_{W_j}^{\pi} \tag{5.19}$$

$$\xi_{P_f}^{\pi} = \sum_{j=1}^{m} p_{f,j}^2 \xi_{W_j}^{\pi} + \sum_{1\le j\le m} \sum_{1\le k\le m, k\ne j} p_{f,j}p_{f,k}\omega_{W_{jk}}^{\pi} \tag{5.20}$$

Once the PMFs of **Y** are updated by a set of samples, the posterior statistics of P_f^p can be obtained as follows:

$$\mu_{P_f}^{p} = \sum_{j=1}^{m} p_{f,j}\mu_{W_j}^{p} \tag{5.21}$$

$$\xi_{P_f}^{p} = \sum_{j=1}^{m} p_{f,j}^2 \xi_{W_j}^{p} + \sum_{1\le j\le m} \sum_{1\le k\le m, k\ne j} p_{f,j}p_{f,k}\omega_{W_{jk}}^{p} \tag{5.22}$$

Equations (5.21) through (5.22) imply that $\mu_{P_f}^{p}$ and $\xi_{P_f}^{p}$ are functions of the samples of **Y**. Given a prescribed sample size $\mathbf{n}_{\mathbf{Y}0} = \{n_{1,0}, n_{2,0}, \ldots, n_{t,0}\}$ for **Y**, the expectations of the posterior mean and variance of P_f, $\mathrm{E}_N\left[\mu_{P_f}^{p}\right]$ and $\mathrm{E}_N\left[\xi_{P_f}^{p}\right]$, with respect to the sampling outcome in the entire state space of **Y**, i.e. $Y_{1,1},\ldots, Y_{1,m_1},\ldots, Y_{t,1}, \ldots, Y_{t,m_t}$, are as follows,

$$\mathrm{E}_N\left[\mu_{P_f}^{p}\right] = \sum_{j=1}^{m} p_{f,j}\mathrm{E}_N\left[\mu_{W_j}^{p}\right] \tag{5.23}$$

$$\mathrm{E}_N\left[\xi_{P_f}^{p}\right] = \sum_{j=1}^{m} p_{f,j}^2\mathrm{E}_N\left[\xi_{W_j}^{p}\right] + \sum_{1\le j\le m}\sum_{1\le k\le m, k\ne j} p_{f,j}p_{f,k}\mathrm{E}_N\left[\omega_{W_{jk}}^{p}\right] \tag{5.24}$$

The derivations of equations for calculating $\mu_{W_j}^{\pi}$, $\xi_{W_j}^{\pi}$, $\omega_{W_{jk}}^{\pi}$, $\mu_{W_j}^{p}$, $\xi_{W_j}^{p}$, $\omega_{W_{jk}}^{p}$, $\mathrm{E}_N\left[\mu_{W_j}^{p}\right]$, $\mathrm{E}_N\left[\xi_{W_j}^{p}\right]$ and $\mathrm{E}_N\left[\omega_{W_{jk}}^{p}\right]$ are shown in Appendix E.

5.3 Sample size determination

As presented in Section 5.2.2, the failure probability, P_f, is a random variable due to the epistemic uncertainties on the distributions of basic random variables. Let p_e denote a point estimate of P_f. In the Bayesian estimation theory, the quadratic loss function is often used to reflect the discrepancy between the point estimate of a parameter and the true parameter (Pham and Turkkan, 1992; Morris, 1968). The quadratic loss function is advantageous in that the evaluated expected loss is proportional to the variance of P_f.

Therefore, the quadratic loss function is employed in this study to model the loss caused by the discrepancy between P_f and p_e as follows,

$$L(P_f, p_e) = C(p_e - P_f)^2 \tag{5.25}$$

where C is the parameter of the quadratic loss function and a positive constant. Since generally accepted rules to quantify C are scarce in the literature, we determine the magnitude of C based on the following simple heuristic. Equation (5.25) suggests that the loss increases as the discrepancy between P_f and p_e increases. The worst loss corresponds to the upper bound of $(p_e - P_f)^2$, i.e. unity, and equals the cost of failure of the structure. Therefore, it is reasonable to assume C to equal the cost of failure, C_F.

The expected loss with respect to the prior distribution of P_f is as follows,

$$\mathrm{E}_{P_f}[L] = \int C(p_e - p_f)^2 f_{P_f}^{\pi}(p_f) dp_f \tag{5.26}$$

It is proved in Appendix F that $p_e = \mu_{P_f}^{\pi}$ is the optimal estimate of P_f in the sense of minimizing $\mathrm{E}_{P_f}[L]$. It follows that the expected prior loss is,

$$\mathrm{E}_{P_f}[L] = \int C\left(\mu_{P_f}^{\pi} - p_f\right)^2 f_{P_f}^{\pi}(p_f) dp_f = C\xi_{P_f}^{\pi} \tag{5.27}$$

Eq. (5.27) is also known as the expected value of perfect information (EVPI) (Morris, 1968; Pham and Turkkhan, 1993). Once $\mathbf{W_Y}$ and P_f are updated by a set of samples $\mathbf{n_Y}$, the posterior expected loss is evaluated as,

$$\mathrm{E}_{P_f}[L|\mathbf{n_Y}] = \int C\left(\mu_{P_f}^{p} - p_f\right)^2 f_{P_f}^{p}(p_f) dp_f = C\xi_{P_f}^{p} \tag{5.28}$$

Equations (5.27) and (5.28) indicate that the expected loss can be expressed as a function of the variance of P_f regardless of its specific distribution type. Given a prescribed sample sizes $\mathbf{n_{Y0}}$ of $\mathbf{Y}$, the expectation of $\mathrm{E}_{P_f}[L|\mathbf{n_Y}]$ with respect to the sampling outcome in the entire space of $\mathbf{Y}$ is,

$$\mathrm{E}_N\left[\mathrm{E}_{P_f}[L|\mathbf{n_Y}]\right] = \mathrm{E}_N(C\xi_{P_f}^{p}) \tag{5.29}$$

It follows that the expected value of sampling information (EVSI) and ENGS are calculated by Eqs. (5.30) and (5.31), respectively.

$$\mathrm{EVSI}(\mathbf{n_{Y0}}) = C\xi_{P_f}^{\pi} - \mathrm{E}_N\left[C\xi_{P_f}^{p}\right] \tag{5.30}$$

$$\mathrm{ENGS}(\mathbf{n_{Y0}}) = \mathrm{EVSI}(\mathbf{n_{Y0}}) - \mathbf{n_{Y0}}\mathbf{C}_s \tag{5.31}$$

where $\mathbf{C}_s = \left[C_{s,1}, C_{s,2}, \ldots, C_{s,t}\right]^{\mathrm{T}}$ denotes the unit cost of sampling for $\mathbf{Y}$. The sample size, $\mathbf{n_{Y0-opt}}$, that maximizes the value of ENGS is the optimal sample size.

Note that Eqs. (5.25) through (5.31) formulate EVPI, EVSI and ENGS by considering the impact of epistemic uncertainty on the failure probability evaluation of a single component. If the epistemic uncertainty influences the failure probability evaluation of a group of components, of which the failure probability of each individual component is evaluated, the total EVPI (EVSI) is equal to the sum of EVPI (EVSI) associated with each individual components.

5.4 Applications

5.4.1 Example 1: SSD for collecting the samples of model error for the pipeline burst capacity model

This example considers the reliability evaluation for a group of corrosion defects on a buried pipeline. The pipeline segment has a nominal outside diameter D_n = 508 mm, a nominal wall thickness w_{tn} = 5.40 mm and a nominal operating pressure o_{pn} = 5.5 MPa. The pipe is made of API 5L Grade X52 steel with the specified minimum yield strength (SMYS) of 359 MPa. It is assumed that the pipeline segment contains 100 corrosion defects that have been detected and sized by a recently conducted inline inspection (ILI). For simplicity, the ILI-reported sizes of different defects are assumed to be identical. The probability of burst of the pipeline at each detected defect is calculated. The burst failure at a given corrosion defect is defined by the following limit state function,

$$g = r_b - o_p \tag{5.32}$$

$$r_b = \kappa \frac{2w_t(\sigma_y + 68.95)}{D} \left[\frac{1 - 0.85\frac{d}{w_t}}{1 - 0.85\frac{d}{Mw_t}} \right] \tag{5.33}$$

where r_b is the burst pressure capacity of the pipe at the defect calculated by the B31G Modified model (Eq. (5.33)) (Kiefner and Vieth, 1989); o_p is the (actual as opposed to nominal) internal pressure of the pipeline; d is the defect depth (i.e. in the through-pipe wall thickness direction); D is the actual outside diameter; w_t is the actual pipe wall thickness; σ_y is the actual yield strength; κ denotes the model error associated with the B31G Modified model, and M is Folias bulging factor which is a function of D, w_t and defect length l (i.e. in the pipe axial direction). The probabilistic properties of the considered random variables are summarized in Table 5.1.

Table 5.1 Probabilistic characteristics of random variables of the pipeline

Parameter	Distribution	Mean	COV (%)	Standard deviation	Source
d	Normal	$0.4w_{tn}$	-	$0.078w_{tn}$	Typical measurement error of ILI tools
l	Normal	150 mm	-	7.8 mm	
D/D_n	Deterministic	1.0	-	-	CSA (2015)
w_t/w_{tn}	Normal	1.0	1.5	-	
σ_y/SMYS	Lognormal	1.1	3.5	-	
o_p/o_{pn}	Gumbel	1.05	3.0	-	
κ	Lognormal	1.297	25.8	-	Zhou and Huang (2012)

The distribution of κ given in Table 5.1 is estimated from burst tests of pipe specimens containing isolated single corrosion defects (Zhou and Huang, 2012). However, suppose that the majority of the defects considered in this example are clustered corrosion defects; the probabilistic characterization of κ given in Table 5.1 does not capture entirely the uncertainty of the burst model for such defects. Given the failure probability is highly sensitive to the probabilistic property of κ (Zhou and Zhang, 2015), it is desirable to perform a number of full-scale burst tests on pipe specimens containing clustered corrosion defects to update the distribution of κ. Since the cost of the burst test is high, the proposed SSD methodology is applied to determine the optimal number of full-scale burst tests. In practice, the cost of the full-scale burst test of a corroded pipe specimen, C_κ, is approximately \$100,000. The failure cost, C_F, is however difficult to quantify, in particular

the indirect cost of failure; we assume that C_F is 500C_κ. Therefore, the relative magnitudes of C_κ and C_F are 1 and 500, respectively.

The prior distribution of κ, which is the one indicated in Table 5.1, is discretized into 40 states, m_κ = 40, and the corresponding PMF is plotted in Fig. 5.1. $\mathbf{W}_\kappa$ is then modeled by a prior Dirichlet distribution $\mathbf{W}_\kappa \sim \text{Dir}(\boldsymbol{\alpha}_\kappa)$, where $\boldsymbol{\alpha}_\kappa = \{\alpha_{\kappa,1}, \alpha_{\kappa,2}, \ldots, \alpha_{\kappa,40}\}$. The equivalent sample size, $\alpha_{\kappa 0} = \sum_{i=1}^{40} \alpha_{\kappa,i}$, of the prior Dirichlet distribution is assumed to be unity, which is commonly assumed in the literature (Zhou et al., 2016). The prior statistics of P_f associated with a single corrosion defect, $\mu_{P_f}^{\pi}$ and $\xi_{P_f}^{\pi}$, are calculated to be 0.0068 and 0.0018, respectively. The $p_{f,i}$ in Eqs. (5.19) and (5.20) is calculated using the simple MC simulation with 1,000,000 trials. Note that in the MC simulation to calculate $p_{f,i}$, the samples of κ is generated from the prior lognormal distribution truncated beyond the boundaries of the state (κ_i, κ_{i+1}] (Straub, 2009; Zwirglmaier and Straub, 2016).

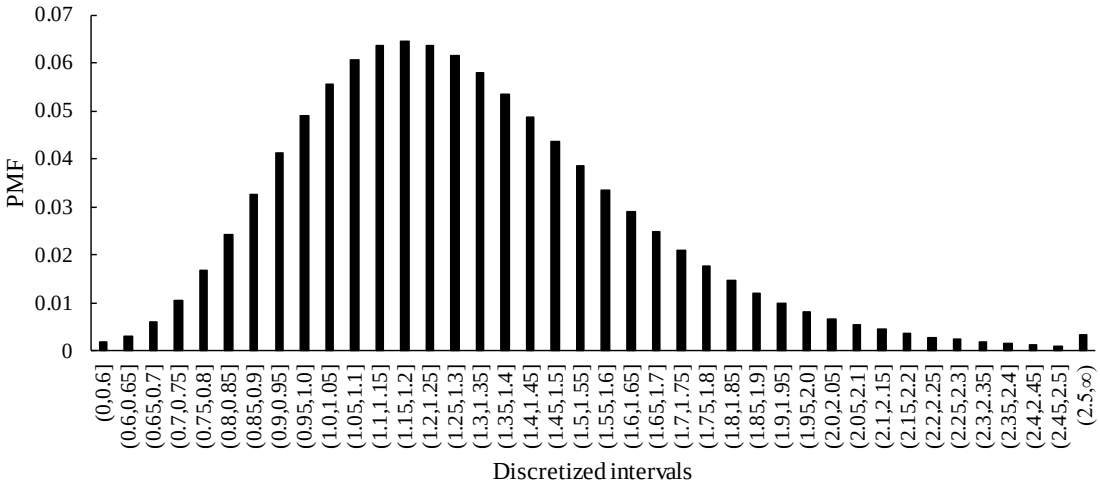


Figure 5.1 Discretization and PMF of κ

To show the impact of the sample size on the uncertainty of failure probability, the variation of $\text{E}_N\left[\xi_{P_f}^{p}\right]/\xi_{P_f}^{\pi}$ with $n_{\kappa 0}$ is plotted in Fig. 5.2, which indicates that the epistemic uncertainty in the failure probability P_f decreases as the sample size increases. The EVPI is calculated to be 97, which defines the upper bound of EVSI. According to Eq. (5.31), the upper bound of EVSI equal to 97 suggests that the sampling value associated with any

sample size large than 97 cannot outweigh the associated sampling cost. The values of EVSI and ENGS corresponding to $n_{\kappa 0}$ are then calculated and plotted in Fig. 5.3. This figure indicates that, as the sample size increases, EVSI increases, whereas the contribution of a unit sample to EVSI decreases. The peak value on the curve corresponding to ENGS indicates that the optimal number of burst tests is 9.

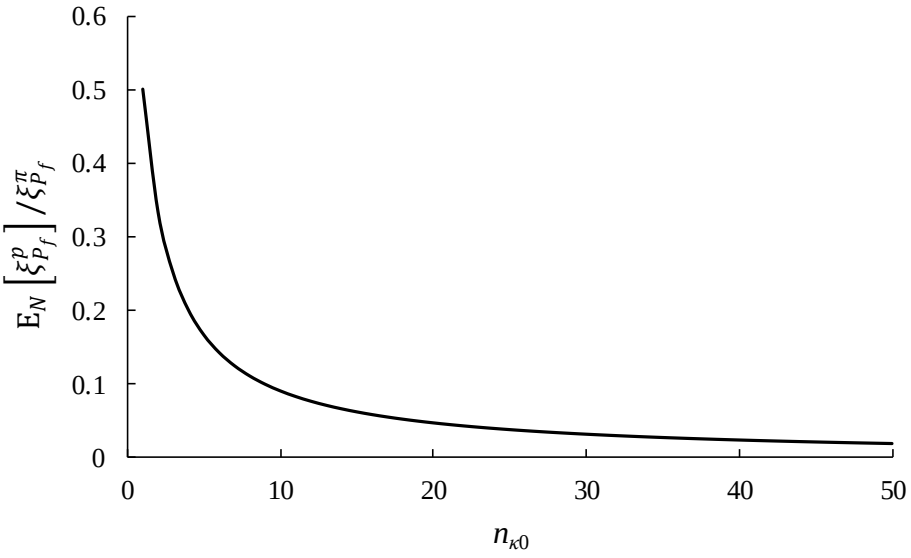


Figure 5.2 Impact of sample size on the uncertainty of failure probability

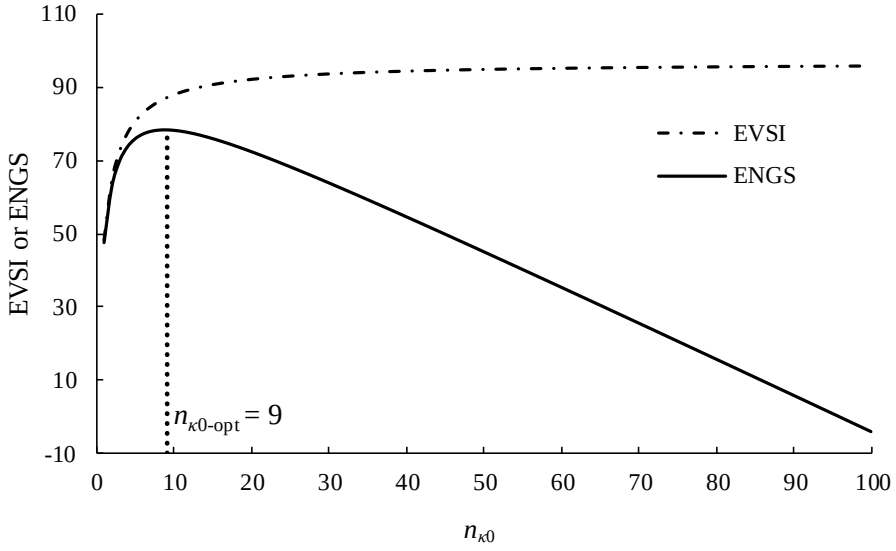
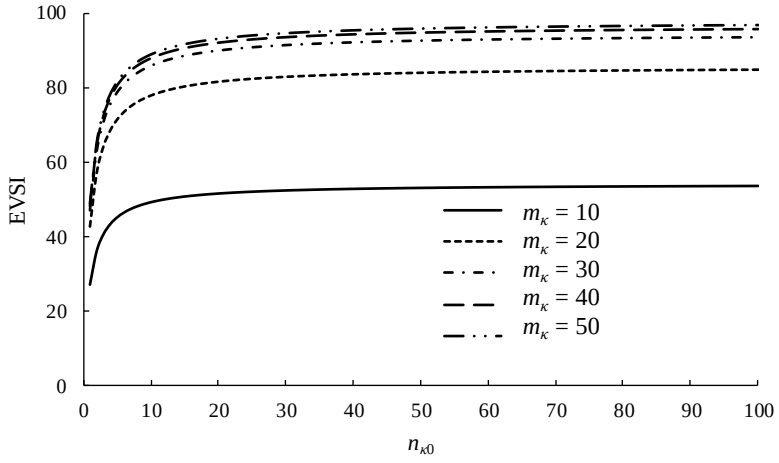
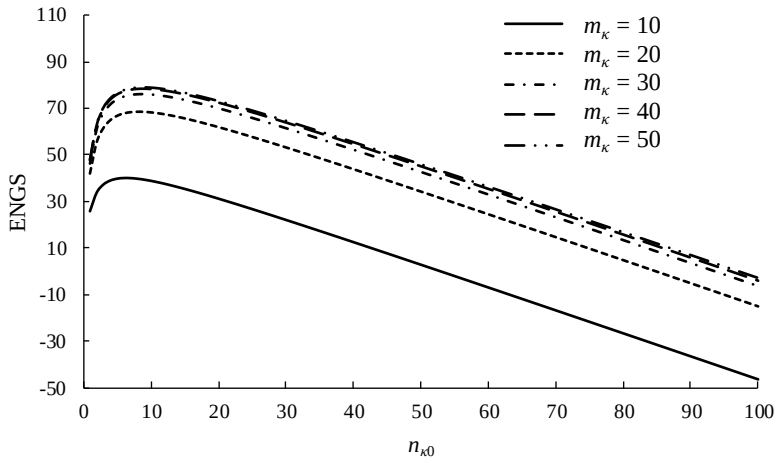


Figure 5.3 The results of EVPI and ENGS

The sensitivity of the SSD results to the number of discretization states of κ is investigated first. All else being equal, the distribution of κ is discretized into 10, 20, 30, 40 and 50 states, respectively, and the corresponding EVSI and ENGS are plotted in Figs. 5.4 (a) and 5.4(b), respectively. If m_κ is equal to or greater than 30, slight changes on EVSI and ENGS are observed as m_κ increases. This suggests that $m_\kappa = 40$ corresponding to the results shown in Fig. 5.1 is an adequate discretization strategy for this example. It should be pointed out that an adequate discretization strategy is problem-specific and generally needs to be determined through a trial-and-error process.



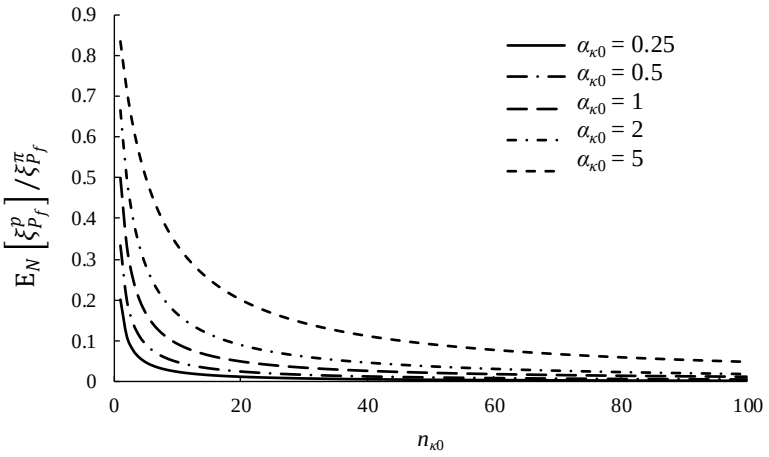
(a) EVSI



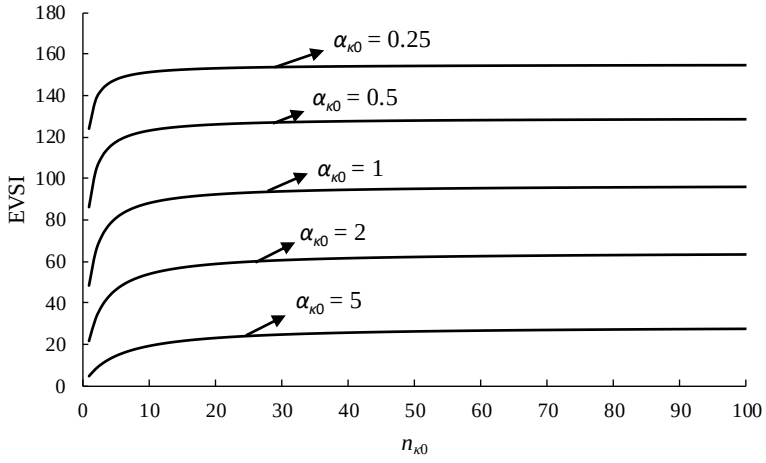
(b) ENGS

Figure 5.4 Sensitivity of SSD results to m_κ

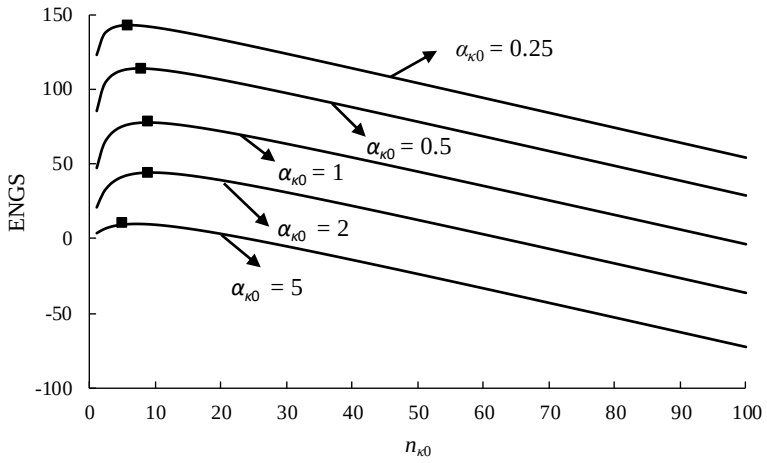
The sensitivity of the SSD results to the equivalent sample size $\alpha_{\kappa 0}$ of the prior Dirichlet distribution is investigated next. All else being equal, $\alpha_{\kappa 0}$ is set to 0.25, 0.5, 1, 2 and 5, respectively. The values of the corresponding EVPI are 154, 129, 97, 65 and 32, respectively, i.e. EVPI decreases as $\alpha_{\kappa 0}$ increases. This is due to that a larger $\alpha_{\kappa 0}$ implies lower uncertainties in the prior Dirichlet distributions as well as P_f. Figures 5.5(b) and 5.5(c) indicate that for a given sample size $n_{\kappa 0}$, EVSI and ENGS decrease too as $\alpha_{\kappa 0}$ increases. However, the same trend does not hold for the optimal sample size: it increases as $\alpha_{\kappa 0}$ increases from 0.25 to 2, but decreases as $\alpha_{\kappa 0}$ increases from 2 to 5. This trend is explained by the trade-off between two influencing factors, the magnitude of EVSI and sensitivity of EVSI to the sample size. Figure 5.5(b) indicates that EVSI increases as $\alpha_{\kappa 0}$ decreases from 5 to 0.25, which tends to lead to a larger optimal sample size according to Eq. (5.31). On the other hand, Figs. 5.5(a) and 5.5(b) indicate that, as $\alpha_{\kappa 0}$ decreases, the sensitivity of $\mathrm{E}_N\left[\xi_{P_f}^p\right]$ and EVSI to the sample sizes increases. In the case associated with small $\alpha_{\kappa 0}$, EVSI is close to its upper bound even for a relatively small sample size; therefore, the benefit of more samples may not outweigh the corresponding sampling cost. It follows that the optimal sample size tends to be smaller as $\alpha_{\kappa 0}$ decreases.



(a) $\mathrm{E}_N\left[\xi_{P_f}^p\right]/\xi_{P_f}^{\pi}$



(b) EVSI



(c) ENGS

Figure 5.5 Sensitivity of SSD results to $\alpha_{\kappa 0}$

5.4.2 Example 2: SSD for collecting samples of corrosion defect sizes for unpiggable pipelines

In practice, there are pipelines for which ILI is infeasible or extremely difficult to conduct due to various reasons such as the tight bends, over- or under-size valves, complicated connections and a lack of launching and receiving stations for ILI tools (Beauregard et al., 2018; Rau and Kirkwood, 2016). Such pipelines are commonly known as unpiggable pipelines. One means to assess the corrosion condition of unpiggable pipelines is to employ the Bayesian methodology to infer the probabilistic distributions of defect sizes and density (Caleyo et al., 2015). To carry out this method, the prior distributions of defect sizes and density can be assumed based on the ILI data of pipelines exposed to similar corrosive environment as the unpiggable pipeline. Then, a number of pipe joints (a pipe joint is typically 12 to 20 m long) of the unpiggable pipelines are excavated and inspected to collect the corrosion data. The collected corrosion data are then used to update the prior distributions. The resulting posterior distributions can be further used to estimate the failure probability of the unpiggable pipeline. Since the cost of excavating pipelines is usually high (Zhang and Zhou, 2014), the developed SSD method is used in the following example to determine the optimal number of excavations by balancing the sampling cost and benefit.

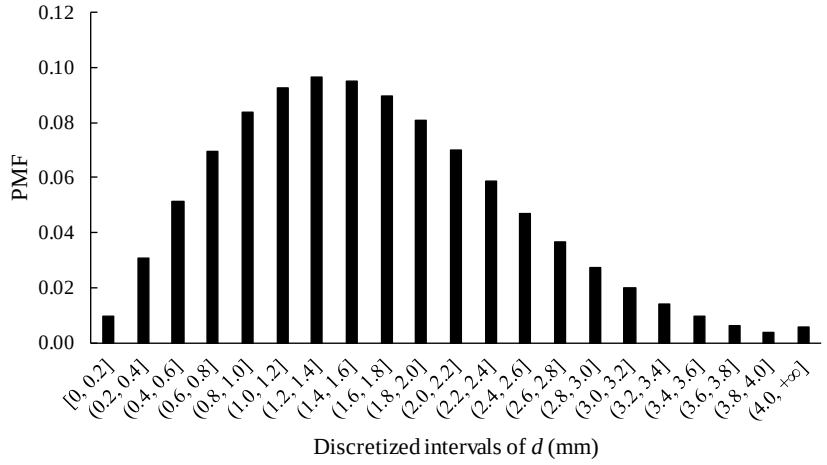
The example considers a 10 km long unpiggable pipeline. The pipeline has D_n = 508 mm, w_{tn} = 5.4 mm and p_n = 5.5 MPa. The pipe steel is API 5L Grade X52 with SMYS = 359 MPa. The pipeline consists of 834 joints, each of which is 12 m long. Usually, one pipeline joint contains multiple corrosion defects. For simplicity, we assume that the failure probability of a pipe joint is dominated by the most critical defect on the joint, defined as the defect at which the pipe joint has the lowest burst pressure capacity. Therefore, only the most critical defect is considered for each pipe joint. It is further assumed that the probabilistic distributions of the depths (lengths) of the critical defects on different joints are identical. A number of pipeline joints will be excavated to collect samples of the defect depth (d) and length (l) and the optimal number of joints to be excavated is determined by the proposed methodology. In practice, the cost of excavating a single pipe joint, C_s, is approximately $200,000. The failure cost, C_F, is assumed to be $250C_s$. It follows that the relative magnitudes of C_s and C_F are 1 and 250, respectively. The probabilistic characteristics of random variables involved in the failure probability evaluation are summarized in Table 5.2. The limited state function defined by Eq. (5.32) and the B31G Modified model defined by Eq. (5.33) are employed in this example to evaluate the failure probability.

Table 5.2 Probabilistic characteristics of random variables

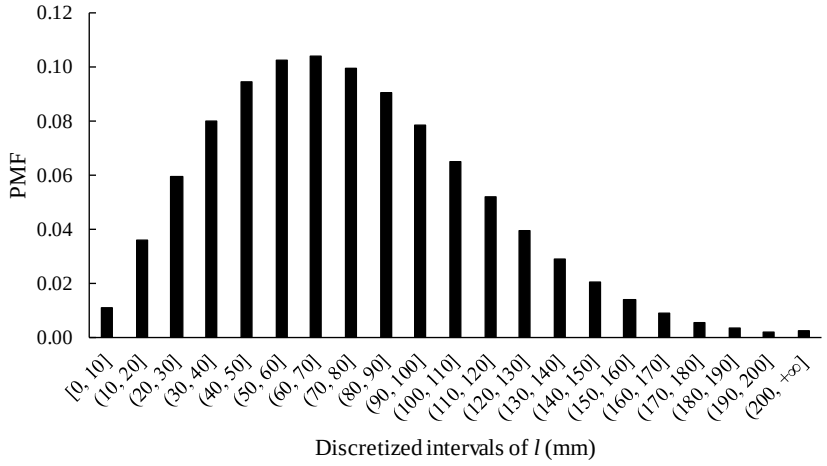
Parameter	Distribution	Mean	COV (%)	Source
d	Weibull	$0.3w_{tn}$	50	Assumed prior distribution
l	Weibull	75 mm	50	
D/D_n	Deterministic	1.0	-	CSA (2015)
w_t/w_{tn}	Normal	1.0	1.5	
p/p_n	Gumbel	1.0	3	
σ_y/SMYS	Lognormal	1.1	3.5	
κ	Lognormal	1.297	25.8	Zhou and Huang (2012)

To apply the proposed methodology to determine the optimal number of joints to excavate, the prior Weibull distributions of d and l defined in Table 5.2 are first discretized. The total number of discrete states, m_d and m_l, are both set to be 21. The PMFs of d and l, $\mathbf{W}_d$ and $\mathbf{W}_l$, are plotted in Figs. 5.6(a) and 5.6(b), respectively. $\mathbf{W}_d$ and $\mathbf{W}_l$ are then modeled by the Dirichlet distributions with α_{d0} and α_{l0} equal to 1. The failure probability of a single pipe joint is evaluated, and $\mu_{P_f}^{\pi}$ and $\xi_{P_f}^{\pi}$ are equal to 0.0078 and 0.000317, respectively.

EVPI is calculated to be 66. Let the sample size, n_0, vary from 1 through 100, and the corresponding EVSI and ENGS are calculated and plotted in Fig. 5.7. ENGS reaches its maximum value at $n_0 = 8$.



(a) *d*



(b) *l*

Figure 5.6 Discretization and PMFs of *d* and *l*

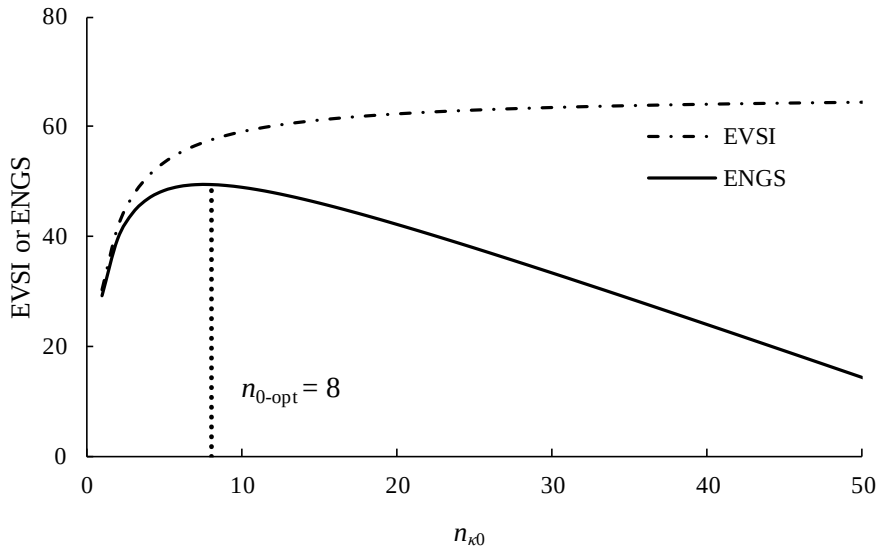
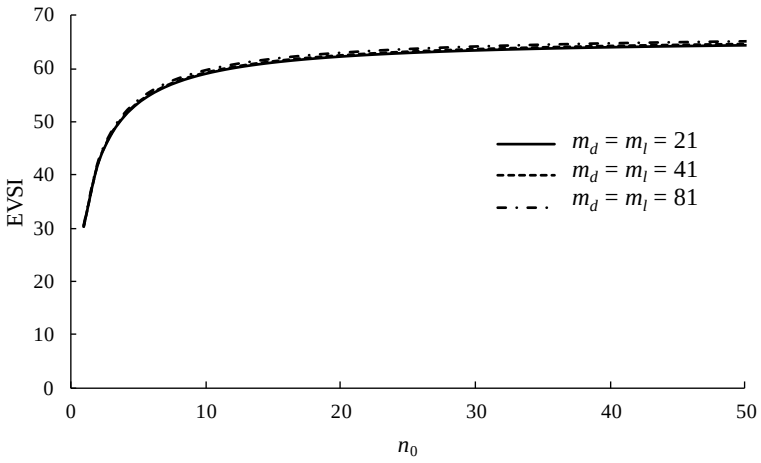
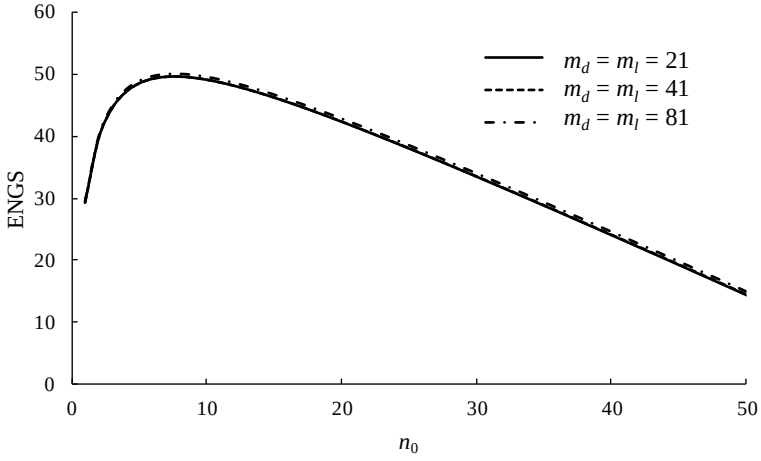


Figure 5.7 The results of EVSI and ENGS

To show that $m_d = m_l = 21$ is adequate for discretization, we consider two more cases of discretization where $m_d = m_l = 41$ and $m_d = m_l = 81$, respectively. The corresponding EVSI and ENGS are plotted in Figs. 5.8(a) and 5.8(b), respectively, which indicate negligible differences among the results associated with the three cases of discretization. Therefore, discretizing the distribution of d and l into 21 states is adequate. This result again demonstrates that the SSD result is insensitive to the discretization of random variables. Next, the sensitivity of the SSD results to the equivalent sample sizes α_{d0} and α_{l0} of the prior Dirichlet distributions is demonstrated. All else being equal, α_{d0} and α_{l0} are set to 0.25, 0.5, 1, 2 and 5, respectively. The values of corresponding EVPI are 116, 93, 88, 42 and 20, respectively. The corresponding EVSI and ENGS are shown in Figs. 5.9(a) and 5.9(b), respectively, which indicates the same trend as observed in Figs. 5.7(a) and 5.7(b). The explanations to Figs. 5.7(a) and 5.7(b) are equally applicable to Figs 5.9(a) and 5.9(b).

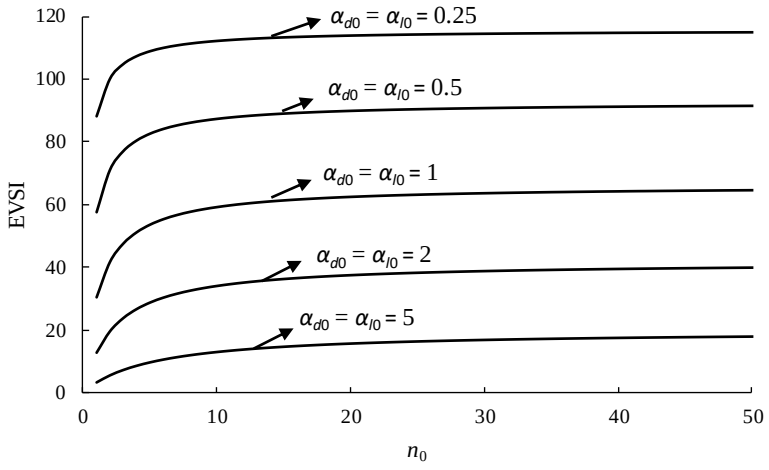


(a) EVSI

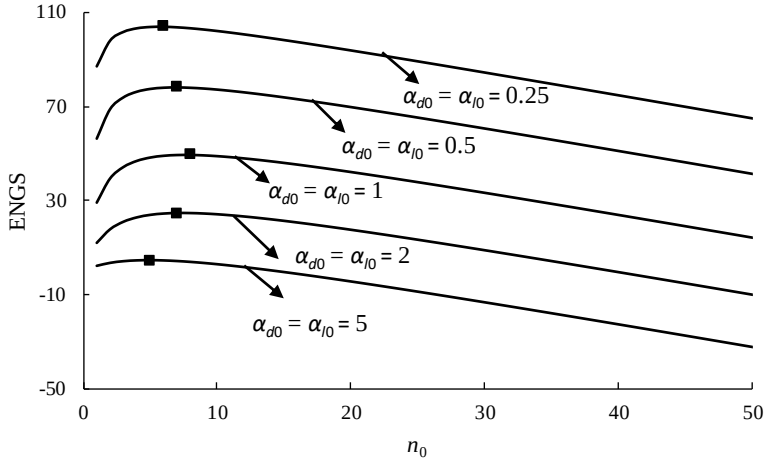


(b) ENGS

Figure 5.8 Sensitivity of SSD results to m_d and m_l



(a) EVSI



(b) ENGS

Figure 5.9 Sensitivity of SSD results to α_{d0} and α_{l0}

5.5 Conclusions

This chapter establishes a methodology of SSD for collecting samples to update the distributions of basic random variables, thus reduce the epistemic uncertainty on the failure probability evaluation. The basic random variable is discretized and a Dirichlet distribution is assigned as its prior PMF. The pre-posterior analysis is performed on the PMFs and failure probability, based on which EVPI, EVSI and ENGS are calculated. The sample

size that maximizes ENGS is the optimal sample size from an economic standpoint. The established methodology has the following two merits: First, the discretization of the continuous random variables and assignment of the Dirichlet distributions to the PMFs make the methodology applicable to a variety of distribution types as opposed to some particular conjugate pairs; second, the analytical solutions of EVPI, EVSI and ENGS are derived, which makes the implementation of the established SSD methodology computationally efficient.

The effectiveness of the proposed methodology is demonstrated by two numerical examples in the context of corrosion assessment of buried pipelines: determining the sample size of the model error of a burst capacity model and determining the number of pipe joints to excavate for the corrosion assessment of unpiggable pipelines. Parametric analysis indicates that the SSD result is insensitive to the discretization of the basic random variables if the random variables are discretized into a fairly large number of states. The SSD result is highly sensitive to the equivalent sample size of the prior Dirichlet distribution. EVPI, EVSI and ENGS decrease as the equivalent sample size increases. The variation of the optimal sample size with the equivalent sample size of the Dirichlet distributions depends on the trade-off between the influence of the equivalent sample size on the magnitude of EVSI and sensitivity of EVSI to sample sizes.

References

Adcock, C.J. (1992) Bayesian approaches to the determination of sample sizes for binomial and multinomial sampling-some comments on the paper by Pham-Gia and Turkkan *The Statistician*, 41, 399-404.

Beauregard, Y., Woo, A., and Huang, T. (2018). Application of In-Line Inspection and Failure Data to Reduce Subjectivity of Risk Model Scores for Uninspected Pipelines. In *Proceedings of the 12th International Pipeline Conference* (pp. V002T07A028-V002T07A028). Calgary, Alberta, Canada.

Caballero, K. L., Barajas, J., and Akella, R. (2012). The generalized Dirichlet distribution in enhanced topic detection. In: Proceedings of the 21st ACM international conference on Information and knowledge management, Maui, HI, USA.

Caleyo, F., Valor, A., Alfonso, L. et al. (2015). Bayesian analysis of external corrosion data of non-piggable underground pipelines. *Corrosion Science*, 90, 33-45.

Canadian Standard Association (CSA). (2015) Oil and gas pipeline systems, CSA Standard Z662-15. Mississauga, Ontario, Canada.

Der Kiureghian, A. (1989). Measures of structural safety under imperfect states of knowledge. *Journal of Structural Engineering*, 115(5), 1119-1140.

Der Kiureghian, A. (2005). First- and second-order reliability methods. In E. Nikolaidis, D. M. Ghiocel, and S. Singhal (Eds.), *Engineering design reliability handbook* (Chap. 14, pp. 1 – 14). Boca Raton: CRC Press.

Der Kiureghian, A. (2008). Analysis of structural reliability under parameter uncertainties. *Probabilistic engineering mechanics*, 23(4), 351-358.

Der Kiureghian, A., and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter?. *Structural Safety*, 31(2), 105-112.

Feelders, A., and Van der Gaag, L. C. Learning Bayesian network parameters under order constraints. *International Journal of Approximate Reasoning*, 42(1-2), 37-53.

Goldsworthy, J.S., Jaksa, M.B., Fenton, G.A., et al. (2007) Effect of sample location on the reliability based design of pad foundations. *Georisk*, 1(3): 155-166.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. *Learning in graphical models*, Springer, Dordrecht; 301-354.

Higo, E., and Pandey, M. D. (2016). Value of information and hypothesis testing approaches for sample size determination in engineering component inspection: a comparison. In: *Proceedings of ASME 2016 Pressure Vessels and Piping Conference* (pp. V005T10A009-V005T10A009), Vancouver, British Columbia, Canada.

Hong, H.P. (1996). Evaluation of the probability of failure with uncertain distribution parameters. *Civil Engineering Systems*, 13(2), 157-168.

Jonson, N. L., and Samuel Kotz. (1972) *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York, USA.

Kiefner, J.F., and Paul H.V. (1989) A modified criterion for evaluating the remaining strength of corroded pipe. No. PR-3-805. Battelle Columbus Div., OH, USA.

Masegosa, A. R., Feelders, A. J., and van der Gaag, L. C. (2016). Learning from incomplete data in Bayesian networks with qualitative influences. *International Journal of Approximate Reasoning*, 69, 18-34.

Melchers, R. E., and Beck, A. T. (2018). *Structural reliability analysis and prediction*. John Wiley and Sons.

Morris, W.T. (1968). *Management science: a Bayesian introduction*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA.

Nishijima, K., and Faber M.H. (2007) Bayesian approach to proof loading of quasi-identical multi-components structural systems. *Civil Engineering and Environmental Systems*, 24 (2), 111-121.

Pham-Gia, A., and Turkkan, T. (1992). Sample size determination in Bayesian analysis *The Statistician*, 41, 105–112.

Raiffa, H., and Schlaifer, R. (1961). *Applied statistical decision theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA, USA.

Rau, J., and Kirkwood, M. (2016). Hydrotesting and In-Line Inspection: Now and in the Future. In *Proceedings of the 11th International Pipeline Conference* (pp. V001T03A055-V001T03A055), Calgary, Alberta, Canada.

Shafieezadeh, A., and Ellingwood, B. R. (2012). Confidence intervals for reliability indices using likelihood ratio statistics. *Structural Safety*, 38, 48-55.

Spiegelhalter, D., Dawid, D., Lauritzen, S., et al. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8, 219-282.

Straub, D. (2009). Stochastic modeling of deterioration processes through dynamic Bayesian networks. *Journal of Engineering Mechanics*, 135(10), 1089-1099.

Zhou, W., and Huang, G. X. (2012). Model error assessments of burst capacity models for corroded pipelines. *International Journal of Pressure Vessels and Piping*, 99, 1-8.

Zhou, Y., Fenton, N., and Zhu, C. (2016). An empirical study of Bayesian network parameter learning with monotonic influence constraints. *Decision Support Systems*, 87, 69-79.

Zhou, W., Gong, C., and Hong, H. P. (2017). New perspective on application of first-order reliability method for estimating system reliability. *Journal of Engineering Mechanics*, 143(9), 04017074.

Zhou, W., and Zhang, S. (2015). Impact of model errors of burst capacity models on the reliability evaluation of corroding pipelines. *Journal of Pipeline Systems Engineering and Practice*, 7(1), 04015011.

Zwirglmaier, K., and Straub, D. (2016). A discretization procedure for rare events in Bayesian networks. *Reliability Engineering and System Safety*, 153, 96-109.

6 Summary, conclusions and recommendations for future study

6.1 General

The work reported in this thesis is focused on employing Bayesian networks and non-parametric Bayesian networks to address four issues in the context of pipeline integrity management with respect to corrosion and third-party damage. Conclusions drawn from the four individual studies are summarized as follows.

6.2 Corrosion growth modeling based on dynamic Bayesian network and parameter learning

Chapter 2 develops a DBN corrosion growth model that incorporates the quantification of measurement errors in ILI data, characterization of corrosion growth, and evaluation of failure probability of the pipeline at the corrosion defect. The model parameters characterizing the errors in ILI data are learned from a dataset consisting of the matched ILI and field-measured corrosion depths using the EM algorithm. The EM algorithm is also employed to learn the model parameters characterizing the annual growth of corrosion depth from a dataset consisting of corrosion depths reported by multiple ILIs.

The effectiveness of the parameter learning for the DBN model is demonstrated by the numerical example involving simulated corrosion data. Application of the DBN model on real corrosion data indicates that the predicted mean corrosion depth in general agree well with the field-measured depth. In comparison with existing corrosion growth models, the developed model is advantageous in the following three respects. First, the integrating and graphical features of the model make the corrosion management more intuitive and transparent to users. Second, the parameter learning technique provides an automated and objective way to extract the parameters of the DBN model from ILI data and field-measured data. Third, the efficient inference algorithm of DBN enables the model updating to be completed highly efficiently.

6.3 Bayesian network model for predicting the probability of third-party damage to underground pipelines

Chapter 3 first develops a BN model to evaluate the probability of pipelines being hit by third-party excavation activities based on a fault tree model widely used in the pipeline industry, and then employs the EM algorithm in the context of parameter learning to learn the parameters of the BN model from two incomplete datasets consisting of individual cases of third-party activities. The TPD datasets simulated by a baseline BN model are first used to examine the effectiveness of the parameter learning, where the KL-divergence between the learned CPT and true CPT is adopted as the metric. The BN model and parameter learning technique are then applied to two real-world TPD datasets collected by a Canadian pipeline operator between 2010 and 2016. The developed model and parameter learning are further validated by the comparison between the empirical value and model-predicted value of two quantities: the probability of a third-party activity being unauthorized and the probability of hit given an unauthorized activity. The results indicate that the probabilities predicted by the BN with the parameters obtained from the parameter learning agree well with the corresponding empirical values.

The developed BN model is advantageous over the existing fault tree model in the following two aspects. First, the BN model can predict the probability of hit under different scenarios of available information, i.e. to predict the probability of hit given a third-party activity with an unknown authorization status, to predict the probability of hit given an authorized activity or unauthorized activity. Second, the BN modeling together with the parameter learning technique provide an effective and efficient means to exploit the historical TPD datasets collected by pipeline operators to learn the failure probabilities of the preventative and protective measures.

6.4 A non-parametric Bayesian network model for predicting the corrosion depth on buried pipelines

Chapter 4 develops an NPBN model for predicting the corrosion depth on underground pipelines. The dependence structure and model parameters, i.e. (conditional) rank correlations are extracted from Velázquez's dataset, which consists of values of the

corrosion depth, pipeline age and nine parameters of surrounding soils from 250 excavation sites in southern Mexico. The empirical correlation matrix evaluated using the samples in Velázquez's dataset indicates that pH value, dissolved chloride, bulk density, water content and pipe-to-soil potential are the most influential soil parameters to the corrosion depth. The 5-fold cross-validation is used to examine the predictive capability of the NPBN model. In the results, the predicted mean corrosion depths in general agree well with the field measurements, and more than 95% field measurements fall in the 5-95 percentile ranges of the predictions. Moreover, the mean value and 5-95 percentile range of corrosion depth associated with clay, clay loam and sandy clay loam are predicted by the NPBN, which indicates that the corrosivity of clay is the highest followed by that of clay loam and sandy clay loam.

In comparison with the regression models, the NPBN can predict the probabilistic distribution of the corrosion depth, which shows the uncertainty associated with the prediction. Moreover, since the correlations between predictor variables are taken into account by the NPBN, the model can handle the prediction of the corrosion depth under the scenarios of missing information, i.e. the values of part of the soil parameters are unavailable. The developed NPBN has significant practical implications in terms of the integrity management of unpiggable pipelines with respect to corrosion.

6.5 Optimal sample size determination based on Bayesian reliability and value of information

Chapter 5 establishes a methodology to determine the optimal sample size for collecting samples to update the distributions of basic random variables, thus reduce the epistemic uncertainty on the failure probability. This methodology first discretizes the basic random variable and assigns a Dirichlet distribution to the PMFs to characterize the epistemic uncertainties. The pre-posterior analysis is performed on the PMFs and failure probability, based on which EVPI, EVSI and ENGS are calculated. The sample size that maximizes ENGS is the optimal sample size from an economic standpoint. The methodology is applied to address two SSD problems in the context of corrosion assessment of buried pipelines: determining the sample size of the model error of a burst capacity model and determining the number of pipe joints to excavate for the corrosion assessment of

unpiggable pipelines. Parametric analyses indicate that the SSD results are insensitive to the discretization of the basic random variables if the random variables are discretized into a fairly large number of states. Since any continuous random variable can be discretized and the Dirichlet distribution can be assigned to the PMF, the application of the methodology is not limited by the original distribution type of the continuous random variable.

6.6 Main assumptions and limitations

The main assumptions based on which the above studies are carried out are emphasized as follows. As a result, the limitations in the conclusions should also be noted.

Chapter 2 assumes that the growth path of defect depth follows a linear function of time. The power-law model is generally considered more appropriate than the linear model to characterize the corrosion growth. This linear assumption is justified in two aspects. First, the growth model is updated continuously with the addition of new ILI data. This allows the predicted growth rate to represent the overall growth path up to the time of the latest ILI. Second, the fact that the interval between subsequent ILIs is usually relatively short, i.e. less than 5 years, implies that the forecasting period over which the linear growth path is extrapolated is relatively short. These two aspects mitigate the error caused by the deviation of the assumed linear growth path from the actual growth path.

In the TPD analysis tools such as fault trees and BN models, the failure probabilities of individual preventative and protective measures are assumed to be objective constants. However, since these measures generally involve human behaviors, the failure probabilities may vary from regions to regions or companies to companies. As a result, the BN model and parameters presented in Chapter 3 is more reflective of the TPD management practice of the company that collects the TPD data. The TPD data from broader sources are desirable to examine the predictive accuracy of the developed BN or update the model parameters before the parameter learning results can be generalized.

It is assumed in Chapter 4 that Velázquez's dataset is drawn from a Gaussian copula, which is validated by a hypothesis test. The employment of the Gaussian copula is primarily for

the reason that it allows the analytical inference. In fact, NPBNs can be developed based on any copula types if only the correlation of zero represents independence between random variables. Therefore, Velázquez's dataset may also be modeled by NPBNs based on other copulas following similar procedures as described in Section 4.2.

6.7 Recommendations for future work

The recommendations for future studies based on the main assumptions, current results and limitations are described as follows.

First, for simplicity, the DBN corrosion growth model assumes that corrosion depth follows a linear function of time with an uncertain growth rate. It is a worthy topic to incorporate more sophisticated models such as the gamma process and power-law function into the DBN growth models, and compare the predictive accuracy of these DBN growth models.

Second, the current DBN growth model evaluates the component failure probability, i.e. the failure probability of the pipeline at a single corrosion defect. Developing a DBN that can model the correlation between the corrosion growths of adjacent defects, thus evaluate the system failure probability (i.e. failure probability of a pipeline segment containing multiple corrosion defects) will benefit the segment-based corrosion management of pipelines.

Third, TPD data from broader sources are desirable to validate the predictive accuracy of the developed BN model or be incorporated into the parameter learning to improve the generality of the learned parameters. Furthermore, the BN model for evaluating the probability of hit due to third-party excavations can be extended to an influence diagram by decision and utility nodes, where the utility nodes characterize the cost and benefit of each individual preventative and protective measures. Such an influence diagram can assist with the allocation of limited management resources by balancing the cost and benefit of individual preventative and protective measures.

Lastly, it is worthwhile to employ NPBNs based on other types of copula to model the Velázquez's dataset and compare the predictive accuracies of these models. Moreover, the

NPBN can also be employed to model the dependence of corrosion length and density (i.e. the number of defects per pipe joints) on the soil parameters if the corresponding datasets are provided by the pipeline operators. Such models combined with the NPBN developed in Chapter 4 can be used to predict the failure probability per pipe joint of unpiggable pipelines using the soil parameters as predictors.

Appendices

Appendix A: Pipeline attributes for the seven TPD regions in the case study

TPD regions	Pipeline attributes								
	A_1:Dig notification requirement	A_2: Public awareness level of one-call	A_3: ROW spacing	A_4: One-call type	A_5: Response time to dig notification	A_6: Patrol frequency	A_7: Locating method	A_8: Response method to notification	A_9: Burial depth
R-1	Required but not enforced	Above average	Intermittent and/or very limited indication	Unified to minimum standard	Three days	Three times per year	Magnetic techniques	Locate/mark/site supervision	1.2 m
R-2	Required but not enforced	Above average	Continuous but limited indication	Unified to minimum standard	Three days	Semi-annually	Magnetic techniques	Locate/mark/site supervision	1.2 m
R-3	Required but not enforced	Average	Continuous but limited indication	Unified to minimum standard	Three days	Three times per year	Magnetic techniques	Locate/mark/site supervision	1.0 m
R-4	Required and enforced	Above average	Continuous but limited indication	Unified to minimum standard	Three days	Weekly	Magnetic techniques	Locate/mark/site supervision	0.6 m
R-5	Required but not enforced	Average	Continuous but limited indication	Unified to minimum standard	Three days	Three times per year	Magnetic techniques	Locate/mark/site supervision	1.5 m
R-6	Required and enforced	Average	Continuous but limited indication	Unified to minimum standard	Three days	Three times per year	Magnetic techniques	Locate/mark/site supervision	1.1 m
R-7	Required but not enforced	Above average	Continuous but limited indication	Unified to minimum standard	Three days	Weekly	Magnetic techniques	Locate/mark/site supervision	1.5 m

Appendix B: Example of evaluating conditional and unconditional rank correlations using Eqs. (4.3) and (4.4)

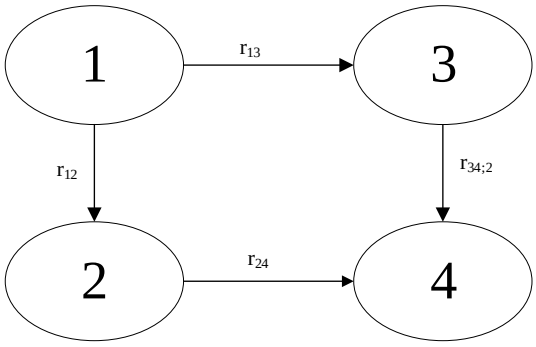


Figure B.1 NPBN with four nodes and four arcs

Consider the NPBN shown in Fig. B.1 and scenario of evaluating the conditional rank correlations given unconditional rank correlations. Let r_{ij} (i, j = 1, 2, …, 4) denote the unconditional rank correlation (in normal space) between nodes i and j. The conditional rank correlation $r_{34;2}$ is determined as follows. According to Eq. (4.4),

$$\rho_{34;2} = \frac{\rho_{34} - \rho_{23} \cdot \rho_{24}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{24}^2)}} \tag{B.1}$$

where ρ_{23}, ρ_{24} and ρ_{34} are transformed from r_{23}, r_{24} and r_{34}, respectively, using Eq. (4.3). Equation (4.3) is then used to evaluate $r_{34;2}$ from $\rho_{34;2}$.

Consider now the NPBN in Fig. B.1 and the scenario of evaluating the rank correlation matrix for the nodes given the conditional rank correlations. Note that r_{12}, r_{13}, r_{24} are given by the NPBN. The evaluation of r_{14}, r_{23}, r_{34} is described as follows. Based on Eq. (4.4),

$$\rho_{23;1} = \frac{\rho_{23} - \rho_{12} \cdot \rho_{13}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}} \tag{B.2}$$

where ρ_{12} and ρ_{13} are evaluated from r_{12} and r_{13} using Eq. (4.3), respectively. Since the missing arcs imply conditional independence (Hanea et al., 2006), $\rho_{23;1} = 0$. Substituting $\rho_{23;1} = 0$ into Eq. (B.2), one can obtain $\rho_{23} = \rho_{12}\, \rho_{13}$.

Based on Eq. (4.4),

$$\rho_{34;2} = \frac{\rho_{34} - \rho_{23} \cdot \rho_{24}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{24}^2)}} \tag{B.3}$$

where ρ_{24} and $\rho_{34;2}$ are evaluated from r_{24} and $r_{34;2}$ ($r_{34;2}$ is given by the NPBN), respectively, using Eq. (4.3), and ρ_{23} has been evaluated before. It follows that ρ_{34} can be obtained from Eq. (B.3).

Based on Eq. (4.4),

$$\rho_{14;23} = \frac{\rho_{14;2} - \rho_{13;2} \cdot \rho_{43;2}}{\sqrt{(1-\rho_{13;2}^2)(1-\rho_{43;2}^2)}} \tag{B.4}$$

where

$$\rho_{13;2} = \frac{\rho_{13} - \rho_{12} \cdot \rho_{23}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)}} \tag{B.5}$$

$$\rho_{43;2} = \frac{\rho_{43} - \rho_{42} \cdot \rho_{23}}{\sqrt{(1-\rho_{42}^2)(1-\rho_{23}^2)}} \tag{B.6}$$

According to the conditional independence implied by the missing arc in the NPBN, $\rho_{14;23}$ =0. Substituting $\rho_{14;23}$ = 0, Eqs. (B.5) and (B.6) into Eq. (B.4), one can evaluate $\rho_{14;2}$. Again, based on Eq. (B.4),

$$\rho_{14;2} = \frac{\rho_{14} - \rho_{12} \cdot \rho_{24}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{24}^2)}} \tag{B.7}$$

ρ_{14} can then be obtained from Eq. (B.7). It follows that r_{14}, r_{23} and r_{24} can be evaluated from ρ_{14}, ρ_{23} and ρ_{24}, respectively, using Eq. (4.3).

Appendix C: The PDF and CDF of Burr distribution

The PDF of Burr distribution is given by,

$$f(x) = \frac{\alpha k\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{k+1}} \qquad (x, k, \alpha, \beta > 0) \tag{C.8}$$

where k and α are shape parameters, and β is the scale parameter.

The CDF of Burr distribution is given by,

$$F(x) = 1 - \left(1 + \left(\frac{x}{\beta}\right)^{\alpha}\right)^{-k} \tag{C.9}$$

Appendix D: The derivation of the pre-posterior statistics of the basic random variable *Y*

For Eq. (5.8), take expectation with respect to $\mathbf{N}_Y$ on both sides,

$$\mathrm{E}_N\left[\mu^p_{W_{Y,j}}\right] = \frac{\alpha_{Y,j}+\mathrm{E}_N[N_{Y,j}]}{\alpha_{Y0}+n_{Y0}} \tag{D.1}$$

Substitute Eq. (5.12) into Eq. (D.1),

$$\mathrm{E}_N\left[\mu^p_{W_{Y,j}}\right] = \frac{\alpha_{Y,j}+\mathrm{E}_N[N_{Y,j}]}{\alpha_{Y0}+n_{Y0}} = \frac{\alpha_{Y,j}+\mu_{N_{Y,j}}}{\alpha_{Y0}+n_{Y0}} = \frac{\alpha_{Y,j}+n_{Y0}\frac{\alpha_{Y,j}}{\alpha_{Y0}}}{\alpha_{Y0}+n_{Y0}} = \frac{\alpha_{Y,j}}{\alpha_{Y0}} = \mu^{\pi}_{W_{Y,j}} \tag{D.2}$$

which proves Eq. (5.15).

For Eq. (5.9), take expectation with respect to $\mathbf{N}_Y$ on both sides,

$$\mathrm{E}_N\left[\xi^p_{W_{Y,j}}\right] = \frac{\alpha_{Y,j}(\alpha_{Y0}+n_{Y0}-\alpha_{Y,j})-\alpha_{Y,j}\mathrm{E}_N[N_{Y,j}]+(\alpha_{Y0}+n_{Y0}-\alpha_{Y,j})\mathrm{E}_N[N_{Y,j}]-\mathrm{E}_N[{N_{Y,j}}^2]}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)} \tag{D.3}$$

Substitute Eqs. (5.12) and (5.13) into Eq. (D.3)

$$\mathrm{E}_N\left[\xi^p_{W_{Y,j}}\right] = \frac{\alpha_{Y,j}(\alpha_{Y0}+n_{Y0}-\alpha_{Y,j})-\alpha_{Y,j}\mu_{N_{Y,j}}+(\alpha_{Y0}+n_{Y0}-\alpha_{Y,j})\mu_{N_{Y,j}}-\xi_{N_{Y,j}}-\left(\mu_{N_{Y,j}}\right)^2}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)}$$

$$= \frac{\alpha_{Y,j}(\alpha_{Y0}+n_{Y0}-\alpha_{Y,j})+n_{Y0}\frac{\alpha_{Y,j}}{\alpha_{Y0}}(\alpha_{Y0}+n_{Y0}-2\alpha_{Y,j})-n_{Y0}\frac{\alpha_{Y,j}}{\alpha_{Y0}}\left(1-\frac{\alpha_{Y,j}}{\alpha_{Y0}}\right)\left(\frac{\alpha_{Y0}+n_{Y0}}{1+\alpha_{Y0}}\right)-\left(n_{Y0}\frac{\alpha_{Y,j}}{\alpha_{Y0}}\right)^2}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)}$$

$$= \frac{\alpha_{Y,j}\alpha_{Y0}(\alpha_{Y0}+n_{Y0})(\alpha_{Y0}-\alpha_{Y,j})(\alpha_{Y0}+n_{Y0}+1)}{(\alpha_{Y0})^2(\alpha_{Y0}+1)(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)}$$

$$= \frac{\alpha_{Y0}}{\alpha_{Y0}+n_{Y0}}\frac{\alpha_{Y,j}(\alpha_{Y0}-\alpha_{Y,j})}{(\alpha_{Y0})^2(\alpha_{Y0}+1)}$$

$$= \frac{\alpha_{Y0}}{\alpha_{Y0}+n_{Y0}}\xi^{\pi}_{W_{Y,j}} \tag{D.4}$$

which proves Eq. (5.16).

For Eq. (5.10), take expectation with respect to $\mathbf{N}_Y$ on both sides,

$$\mathrm{E}_N\left[\omega^{p}_{W_{Y,jk}}\right] = \frac{-\alpha_{Y,j}\alpha_{Y,k} - \alpha_{Y,j}\mathrm{E}_N[N_{Y,k}] - \alpha_{Y,k}\mathrm{E}_N[N_{Y,j}] - \mathrm{E}_N[N_{Y,j}N_{Y,k}]}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)} \tag{D.5}$$

Substitute Eqs. (5.12) through (5.14) into Eq. (D.5),

$$\mathrm{E}_N\left[\omega^{p}_{W_{Y,jk}}\right] = \frac{-\alpha_{Y,j}\alpha_{Y,k} - \alpha_{Y,j}\mu_{N_{Y,k}} - \alpha_{Y,k}\mu_{N_{Y,j}} - \mu_{N_{Y,j}}\mu_{N_{Y,k}} - \omega_{N_{Y,jk}}}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)}$$

$$= \frac{n_{Y0}\alpha_{Y,j}\alpha_{Y,k}(\alpha_{Y0}+n_{Y0}) - \alpha_{Y,j}\alpha_{Y,k}(\alpha_{Y0})^2(\alpha_{Y0}+1) - (n_{Y0})^2\alpha_{Y,j}\alpha_{Y,k}(\alpha_{Y0}+1) - 2n_{Y0}\alpha_{Y,j}\alpha_{Y,k}\alpha_{Y0}(\alpha_{Y0}+1)}{(\alpha_{Y0})^2(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+1)(\alpha_{Y0}+n_{Y0}+1)}$$

(D.6)

which proves Eq. (5.17).

Appendix E: The derivation of the prior, posterior and pre-posterior statistics of W_j

The probability of the *j*-th state of the vector **Y** representing the basic random variables is denoted as,

$$W_j = \prod_{i=1}^{t} W_{i,j_i} \tag{E.1}$$

where W_{i,j_i} denotes the j_i-th PMF of the *i*-th basic random variable.

The prior statistics of W_j are calculated by Eqs. (E.2) through (E.4) as follows,

$$\mu_{W_j}^{\pi} = \mathrm{E}_Y^{\pi}\left[\prod_{i=1}^{t} W_{i,j_i}\right] = \prod_{i=1}^{t} \mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right] = \prod_{i=1}^{t} \mu_{W_{i,j_i}}^{\pi} \tag{E.2}$$

$$\begin{aligned}\xi_{W_j}^{\pi} &= \mathrm{V}_Y^{\pi}\left[\prod_{i=1}^{t} W_{i,j_i}\right] = \prod_{i=1}^{t} \mathrm{E}_Y^{\pi}\left[{W_{i,j_i}}^2\right] - \prod_{i=1}^{t}\left(\mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right]\right)^2 \\ &= \prod_{i=1}^{t}\left(\mathrm{V}_Y^{\pi}\left[W_{i,j_i}\right] + \left(\mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right]\right)^2\right) - \prod_{i=1}^{t}\left(\mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right]\right)^2 \\ &= \prod_{i=1}^{t}\left(\xi_{W_{i,j_i}}^{\pi} + \left(\mu_{W_{i,j_i}}^{\pi}\right)^2\right) - \prod_{i=1}^{t}\left(\mu_{W_{i,j_i}}^{\pi}\right)^2\end{aligned} \tag{E.3}$$

$$\begin{aligned}\omega_{W_{jk}}^{\pi} &= \mathrm{Cov}_Y^{\pi}\left[\prod_{i=1}^{t} W_{i,j_i}, \prod_{i=1}^{t} W_{i,k_i}\right] \\ &= \mathrm{E}_Y^{\pi}\left[\prod_{i=1}^{t} W_{i,j_i} \prod_{i=1}^{t} W_{i,k_i}\right] - \mathrm{E}_Y^{\pi}\left[\prod_{i=1}^{t} W_{i,j_i}\right]\mathrm{E}_Y^{\pi}\left[\prod_{i=1}^{t} W_{i,k_i}\right] \\ &= \prod_{i=1}^{t} \mathrm{E}_Y^{\pi}\left[W_{i,j_i} W_{i,k_i}\right] - \prod_{i=1}^{t} \mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right]\mathrm{E}_Y^{\pi}\left[W_{i,k_i}\right] \\ &= \prod_{i=1}^{t} \mathrm{E}_Y^{\pi}\left[W_{i,j_i} W_{i,k_i}\right] - \prod_{i=1}^{t} \mu_{W_{i,j_i}}^{\pi} \mu_{W_{i,k_i}}^{\pi}\end{aligned} \tag{E.4}$$

where,

$$\mathrm{E}_Y^{\pi}\left[W_{i,j_i} W_{i,k_i}\right] = \mathrm{Cov}_Y^{\pi}\left[W_{i,j_i}, W_{i,k_i}\right] + \mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right]\mathrm{E}_Y^{\pi}\left[W_{i,k_i}\right] = \omega_{W_{i,jk}}^{\pi} + \mu_{W_{i,j_i}}^{\pi} \mu_{W_{i,k_i}}^{\pi} \text{ if } j_i \neq k_i$$

$$\mathrm{E}_Y^{\pi}\left[W_{i,j_i} W_{i,k_i}\right] = \mathrm{V}_Y^{\pi}\left[W_{i,j_i}, W_{i,k_i}\right] + \mathrm{E}_Y^{\pi}\left[W_{i,j_i}\right]\mathrm{E}_Y^{\pi}\left[W_{i,k_i}\right] = \xi_{W_{i,j_i}}^{\pi} + \left(\mu_{W_{i,j_i}}^{\pi}\right)^2 \text{ if } j_i = k_i$$

The posterior statistics W_j are calculated by Eq. (E.5) through (E.7) as follows,

$$\mu_{W_j}^p = \mathrm{E}_Y^p\left[\prod_{i=1}^t W_{i,j_i}\right] = \prod_{i=1}^t \mathrm{E}_Y^p\left[W_{i,j_i}\right] = \prod_{i=1}^t \mu_{W_{i,j_i}}^p \tag{E.5}$$

$$\begin{aligned}\xi_{W_j}^p &= \mathrm{V}_Y^p\left[\prod_{i=1}^t W_{i,j_i}\right] = \prod_{i=1}^t \mathrm{E}_Y^p\left[{W_{i,j_i}}^2\right] - \prod_{i=1}^t\left(\mathrm{E}_Y^p\left[W_{i,j_i}\right]\right)^2 \\ &= \prod_{i=1}^t\left(\mathrm{V}_Y^p\left[W_{i,j_i}\right] + \left(\mathrm{E}_Y^p\left[W_{i,j_i}\right]\right)^2\right) - \prod_{i=1}^t\left(\mathrm{E}_Y^p\left[W_{i,j_i}\right]\right)^2 \\ &= \prod_{i=1}^t\left(\xi_{W_{i,j_i}}^p + \left(\mu_{W_{i,j_i}}^p\right)^2\right) - \prod_{i=1}^t\left(\mu_{W_{i,j_i}}^p\right)^2\end{aligned} \tag{E.6}$$

$$\begin{aligned}\omega_{W_{jk}}^p &= \mathrm{Cov}_Y^p\left[\prod_{i=1}^t W_{i,j_i}, \prod_{i=1}^t W_{i,k_i}\right] \\ &= \mathrm{E}_Y^p\left[\prod_{i=1}^t W_{i,j_i}\prod_{i=1}^t W_{i,k_i}\right] - \mathrm{E}_Y^p\left[\prod_{i=1}^t W_{i,j_i}\right]\mathrm{E}_Y^p\left[\prod_{i=1}^t W_{i,k_i}\right] \\ &= \prod_{i=1}^t \mathrm{E}_Y^p\left[W_{i,j_i}W_{i,k_i}\right] - \prod_{i=1}^t \mathrm{E}_Y^p\left[W_{i,j_i}\right]\mathrm{E}_Y^p\left[W_{i,k_i}\right] \\ &= \prod_{i=1}^t \mathrm{E}_Y^p\left[W_{i,j_i}W_{i,k_i}\right] - \prod_{i=1}^t \mu_{W_{i,j_i}}^p\mu_{W_{i,k_i}}^p\end{aligned} \tag{E.7}$$

where,

$$\mathrm{E}_Y^p\left[W_{i,j_i}W_{i,k_i}\right] = \mathrm{Cov}_Y^p\left[W_{i,j_i}, W_{i,k_i}\right] + \mathrm{E}_Y^p\left[W_{i,j_i}\right]\mathrm{E}_Y^p\left[W_{i,k_i}\right] = \omega_{W_{i,j_ik_i}}^p + \mu_{W_{i,j_i}}^p\mu_{W_{i,k_i}}^p \text{ if } j_i \neq k_i$$

$$\mathrm{E}_Y^p\left[W_{i,j_i}W_{i,k_i}\right] = \mathrm{V}_Y^p\left[W_{i,j_i}, W_{i,k_i}\right] + \mathrm{E}_Y^p\left[W_{i,j_i}\right]\mathrm{E}_Y^p\left[W_{i,k_i}\right] = \xi_{W_{i,j_i}}^p + \left(\mu_{W_{i,j_i}}^p\right)^2 \text{ if } j_i = k_i$$

The pre-posterior statistics W_j are calculated by Eq. (E.8) through (E.10) as follows,

$$\mathrm{E}_N\left[\mu_{W_j}^p\right] = \prod_{i=1}^t \mathrm{E}_N\left[\mu_{W_{i,j_i}}^p\right] \tag{E.8}$$

$$\mathrm{E}_N\left[\xi_{W_j}^p\right] = \prod_{i=1}^t\left(\mathrm{E}_N\left[\xi_{W_{i,j_i}}^p\right] + \mathrm{E}_N\left[\left(\mu_{W_{i,j_i}}^p\right)^2\right]\right) - \prod_{i=1}^t \mathrm{E}_N\left[\left(\mu_{W_{i,j_i}}^p\right)^2\right] \tag{E.9}$$

where

$$\mathrm{E}_N\left[\left(\mu_{W_{i,j_i}}^p\right)^2\right] = \frac{\alpha_{i,j_i}\alpha_{i,k_i} + 2\alpha_{i,j_i}\mu_{N_{i,j_i}} + \left(\mu_{N_{i,j_i}}\right)^2 + \xi_{N_{i,j_i}}}{(\alpha_{i0} + n_{i0})^2}$$

$$\mathrm{E}_N\left[\omega_{W_{jk}}^p\right] = \prod_{i=1}^t \mathrm{E}_N\left[E_Y^p\left[W_{i,j_i}W_{i,k_i}\right]\right] - \prod_{i=1}^t \mathrm{E}_N\left[\mu_{W_{i,j_i}}^p \mu_{W_{i,k_i}}^p\right] \tag{E.10}$$

where,

$$\mathrm{E}_N\left[\mathrm{E}_Y^p\left[W_{i,j_i}W_{i,k_i}\right]\right] = \begin{cases} \mathrm{E}_N\left[\xi_{W_{i,j_i}}^p\right] + \mathrm{E}_N\left[\mu_{W_{i,j_i}}^p \mu_{W_{i,k_i}}^p\right] & j_i = k_i \\ \mathrm{E}_N\left[\omega_{W_{i,j_i k_i}}^p\right] + \mathrm{E}_N\left[\mu_{W_{i,j_i}}^p \mu_{W_{i,k_i}}^p\right] & j_i \neq k_i \end{cases}$$

where

$$\mathrm{E}_N\left[\mu_{W_{i,j_i}}^p \mu_{W_{i,k_i}}^p\right] = \begin{cases} \frac{\alpha_{i,j_i}\alpha_{i,k_i} + 2\alpha_{i,j_i}\mu_{N_{i,j_i}} + \left(\mu_{N_{i,j_i}}\right)^2 + \xi_{N_{i,j_i}}}{(\alpha_{i0}+n_{i0})^2} & j_i = k_i \\ \frac{\alpha_{i,j_i}\alpha_{i,k_i} + \alpha_{i,k_i}\mu_{N_{i,j_i}} + \alpha_{i,j_i}\mu_{N_{i,k_i}} + \mu_{N_{i,j_i}}\mu_{N_{i,k_i}} + \omega_{N_{i,j_i k_i}}}{(\alpha_{i0}+n_{i0})^2} & j_i \neq k_i \end{cases}$$

The prior, posterior and pre-posterior statistics of the PMFs of individual basic random variables, W_{i,j_i} and W_{i,k_i}, and the statistics of the counts of samples, N_{i,j_i} and N_{i,k_i}, involved in Eqs. (E.2) through (E.10) can be calculated by Eqs. (5.3) through (5.5), Eqs. (5.8) through (5.10) and Eqs. (5.12) through (5.17).

Appendix F: Optimal point estimate of failure probability

The expected loss is evaluated as follows,

$$\mathrm{E}_{P_f}[L] = \int C\left(p_\mathrm{e} - p_f\right)^2 f_{P_f}(p_f)\,dp_f \tag{F.1}$$

The derivative of Eq. (F.1) with respect to p_e is,

$$\frac{\mathrm{dE}_{P_f}[L]}{\mathrm{d}p_\mathrm{e}} = 2C\int\left(p_\mathrm{e} - p_f\right) f_{P_f}(p_f)\,dp_f = 2C\left(p_\mathrm{e} - \mu_{P_f}\right) \tag{F.2}$$

Let Eq. (F.2) equal to zero. It follows that the minimum value of $\mathrm{E}_{P_f}[L]$ occurs at $p_\mathrm{e} = \mu_{P_f}$.

Curriculum Vitae

Name:	Wei Xiang
Post-secondary Education and Degrees:	Huazhong University of Science and Technology Wuhan, Hubei, China 2008-2012 B.Eng. Huazhong University of Science and Technology Wuhan, Hubei, China 2012-2015 M.Eng Western University London, Ontario, Canada 2015-2019 Ph.D.
Honours and Awards:	Ontario Trillium Scholarship 2015-2019 Western Graduate Research Scholarship 2015-2019
Related Work Experience	Teaching Assistant and Research Assistant Western University 2015-2019

Publications:

Articles in refereed journals:

[1] **Xiang, W.** and Zhou, W. (2019). Integrated pipeline corrosion growth modeling and reliability analysis using dynamic Bayesian network and parameter learning technique. *Structure and Infrastructure Engineering*. (accepted on August 1, 2019).

[2] Zhou, W., **Xiang, W.** and Hong, H.P. (2017). Sensitivity of system reliability of corroding pipelines to modelling of stochastic growth of corrosion defects. *Reliability Engineering and System Safety*, 167, 428-438.

[3] Zhou, W., **Xiang, W.** and Cronin, D. (2016). Probability of rupture model for corroded pipelines. *International Journal of Pressure Vessels and Piping*, 147, 1-11.

[4] Wang, D., Chen, Z., **Xiang, W.**, and Zhu, H. (2017). Experimental investigation of damage identification in beam structures based on the strain statistical moment. *Advances in Structural Engineering*, 20(5), 747-758.

[5] Wang, D., **Xiang, W.**, Zeng, P., and Zhu, H. (2015). Damage identification in shear-type structures using a proper orthogonal decomposition approach. *Journal of Sound and Vibration*, 355, 135-149.

[6] **Xiang, W.**, Wang, D., and Zhu, H. (2014). Damage identification in a plate structure based on strain statistical moment. *Advances in Structural Engineering*, 17(11), 1639-1655.

[7] Wang, D., **Xiang, W.,** and Zhu, H. (2014). Damage identification in beam type structures based on statistical moment using a two step method. *Journal of Sound and Vibration*, 333(3), 745-760.

Conferences papers:

[1] **Xiang, W**., and Zhou, W. (2019). Optimal Sample Size Determination based on Bayesian Reliability and Value of Information. In *Proceedings of the 13th International Conference on Applications of Statistics and Probability in Civil Engineering* (ICASP-13), Seoul, South Korea.

[2] **Xiang, W.**, and Zhou, W. (2018). Corrosion Growth Modeling by Learning a Dynamic Bayesian Network From Multiple In-Line Inspection Data. In *Proceedings of the 12th International Pipeline Conference*, Calgary, Alberta, Canada.