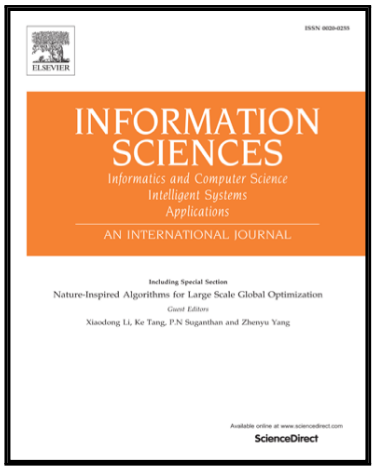
Journal Pre-proof



Multi-State Deterioration Prediction for Infrastructure Asset: Learning from Uncertain Data, Knowledge and Similar Groups

Haoyuan Zhang, D. William R. Marsh

PII: S0020-0255(19)31061-8
DOI: https://doi.org/10.1016/j.ins.2019.11.017
Reference: INS 15009

To appear in: *Information Sciences*

Received date: 27 February 2019
Revised date: 12 October 2019
Accepted date: 12 November 2019

Please cite this article as: Haoyuan Zhang, D. William R. Marsh, Multi-State Deterioration Prediction for Infrastructure Asset: Learning from Uncertain Data, Knowledge and Similar Groups, *Information Sciences* (2019), doi: https://doi.org/10.1016/j.ins.2019.11.017

Highlights

- How to encode data with censorship and expert knowledge in a Bayesian network model for deterioration learning.
- How to elicit and translate engineering knowledge about deterioration behaviours into a Weibull distribution.
- A random forest to select a subset of features, with the selected features, assets are separated into several groups by their similarity.
- A hierarchical BN model that can learn deterioration from other groups.
- Comparing with other existing methods, we show our method gives better performance in predicting asset deterioration, especially for asset groups with little data.

Multi-State Deterioration Prediction for Infrastructure Asset: Learning from Uncertain Data, Knowledge and Similar Groups

Haoyuan Zhang*, D. William R. Marsh

Risk and Information Management Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

Abstract

Infrastructure assets such as bridges need to be inspected regularly for signs of deterioration. Although a fixed inspection interval could be used, an estimate of the rate of deterioration allows us to schedule the next inspection more cost-effectively. Our earlier work outlined a Bayesian framework that uses both data and knowledge to predict the transition between assets, which has been extended and realised in this paper for asset deterioration prediction. In the Bayesian model, censorship is modelled to incorporate the uncertainty from inspection records and prior of the parameter is used to express expert knowledge. In particular, we also suggest how the prior probabilities of the parameters of a Weibull distribution can be set in practice using expert estimates such as the maximum and average times of a transition from one state to another.

Furthermore, assets with similar characteristics may deteriorate similarly. We propose to separate related assets into groups and learn deterioration between these groups. This assumption allows us to tackle the challenge of limited data further and is experimented with the deck inspection records from the National Bridge Inventory database in Wyoming. This database includes over 100 features of each bridge such as structure type and average daily traffic: we use a modified random forest to select a subset of important features to separate assets into groups. The model is extended into hierarchical Bayesian models to learn between groups with the help of hyper-parameters and an aggregated variable from the feature levels. Performance of our method is compared with other existing approaches from various aspects.

Keywords: Deterioration prediction, Multi-state system, Weibull distribution, Feature selection, Hierarchical Bayesian networks, Prior elicitation

1. Introduction

The prediction of asset deterioration can support the decisions on inspection and answer questions like when to inspect the asset or which asset we should inspect. The process of deterioration, such as a bridge deteriorates from a good condition to a structurally deficient condition, is described by the decrease in asset condition over time. Historical inspection records provide information about the asset's previous conditions over time. We can use

*Corresponding author.
Email address: haoyuan.zhang@qmul.ac.uk (Haoyuan Zhang)

these records to infer relevant data to model the deterioration process of an asset. By doing so, we can predict the condition of an asset by estimating how soon it is likely to deteriorate into an unacceptable level in the future.

The asset condition, describing the level of deterioration of an asset, is recorded during each inspection. The publicly available National Bridge Inventory (NBI) database is one example that contains a wide variety of information generated from each inspection and it is studied in this paper. NBI archives unified information of over half a million bridges and tunnels in the United States under the national inspection standard since 1992. The health of a bridge is monitored through periodic bridge inspection. The inspection evaluates the conditions of the deck, superstructure, substructure and culvert on a 9 to 0 scale with a one-point interval. In the grading system, State 9 (S9) represents an excellent condition, and State 0 (S0) represents a failed condition. In general, a condition rating of 4 or lower quantifies a bridge as structurally deficient. As a result, this bridge may require speed or load restriction to ensure safety.

The deterioration time of each state can be inferred from the condition data and their corresponding inspection time. Since most bridges are inspected periodically, the inspected state may not reveal the actual deterioration time of the structure. For example, the most recent inspection shows a structure is in S8 but S9 in the last inspection. These records do not imply the deterioration time from S9 to S8 is the time gap between these two inspections. In fact, the deterioration may happen anytime between these two inspections. Censorship is often introduced to encode this uncertainty [1]. In the usual case, one might have more confident in saying the structure deteriorated before an inspection (left censored), between two consecutive inspections (interval censored), or after the most recent inspection (right censored), rather than a specific time point. In reality, we usually only have a small number of inferred deterioration time data due to the long inspection interval (average bridges in the NBI were inspected every two years). Besides, infrastructure like bridges has a slow decay process, which leads to limited number of state transitions. Within the NBI database (26-year inspection history by far), some bridges do not even deteriorate once. Also, most of the bridges are in good or fair conditions (at S5 or over) because maintainers often intervene when assets are in poor condition to prevent further deterioration. As of 2017, only 7.7% of the bridges are at S4 or less that could produce deterioration data for lower states.

The deterioration rate may vary from asset's characteristics (features). For example, a bridge with more traffic loading may deteriorate more rapidly than a less loaded bridge. When estimating the rate of decay, by considering the impact of different features on an individual asset, it is possible to give a more accurate prediction [2]. By assuming assets with the same feature levels to share the same deterioration rate, we can separate the overall asset population into different groups, where, within each group, individuals are considered with the same deterioration behaviours. However, each asset often associated with many features. For example, ranging from year of built, to maintenance agency, there are over a hundred features associated with each bridge in the NBI database. Considering too many features can be a disadvantage as it takes too many resources and results in slow computations. More importantly, with the increase in the number of features, after reaching an optimal amount of features, many approaches exhibit a decrease of accuracy [3], for example, by overfitting.

Therefore, to provide accurate individualised deterioration predictions, one of the challenges is to reduce the feature dimensionality and to make use of the feature information to learn the rate of decay.

This paper extends and implements the previous work on Bayesian Network (BN) models for deterioration prediction outlined in Zhang and Marsh [4] to tackle the challenges addressed above. The remainder of the paper is organised as follows. In section 2, we review the approaches for deterioration prediction and discuss the advantages and challenges of current BN models. The contribution of this paper starts from Section 3, it tackles the discussed challenges following a real case study. We show how to select a subset of features to group assets, and extend the BN models into hierarchical BN models that can learn deterioration between groups to provide an individualised prediction for each asset group. We also emphasise how to elicit knowledge from engineers that can be included in the models. We compare the performance of our approach with other existing methods in section 4, and conclusions are drawn on section 5.

2. Approach to Deterioration Prediction

Markov process and its variants are commonly used to predict asset deterioration in asset management system such as Bridge Management Systems (BMSs)[5]. A Markov process is a stochastic process where the future state only depends on the present state (Markov property). In deterioration prediction, it is represented by the transition probability from one state to another state (see Frangopol et al. [6] for a comprehensive review). Though these studies have shown remarkable performance for prediction within the global population, they also agreed current approaches suffer difficulty to provide an accurate prediction for individuals. Also, in these studies, deterioration time was assumed to be perfect (complete data) and rich, where in practice, as discussed in the previous section, data are often uncertain, and for some specific type of assets, the data size is relatively small. Some studies had tackled the data uncertainty limitation, for example, Kobayashi et al. [7] addressed the uncertainty by modelling the deterioration using a hidden Markov model. However, they only modelled the measurement errors of asset states given by engineers, while ignoring the random effect on the monitoring from devices by assuming these data are precise. Ferguson et al. [8] instead modelled the data as censored survival time within a Markov model, which can potentially employ in multi-state deterioration prediction. To evaluate the transition probabilities, they employed a non-parametric estimator called Datta-Satten estimator for hazard rate function estimation. Later in section 4.3, we compare its performance with our approach. But in summary, most of these methods are data-driven and only provide aggregated prediction for all assets rather than tailored predictions for individuals.

Apart from traditional Markov-based approaches, parametric statistical distributions are also often studied for deterioration prediction [6]. Various statistical distributions have been used to fit asset deterioration. Among them, Weibull distribution is commonly used for its flexibility in analysing the time to failure, for example, bridge components deterioration in Le [9]. One of the advantages of using a parametric distribution is its interpretability. For example, in a Weibull distribution, its probability density function (pdf) over time t is:

$$f(t) = \frac{\beta}{\eta}(\frac{t}{\eta})^{\beta-1} e^{-(\frac{t}{\eta})^{\beta}}, f(t) \geq 0, \beta \geq 0, \eta \geq 0 \tag{1}$$

it is characterised by parameter shape β and parameter scale η. These parameters quantify the behaviours of the distribution. For example, with a shape β value that is lower than 1, we can assume it has a failure rate that decreases with time, and a shape greater than 1 describing the wear-out failure, giving an increasing failure rate. This property provides a natural way to extract knowledge from engineers. To include this kind of knowledge about the deterioration characteristics, enabled by Bayes' theorem, the priors in BN provide a suitable framework. As a result, it could ease the burden of needing many data.

2.1. Bayesian Networks for Deterioration Prediction

To learn the parametric distribution for deterioration prediction within a Bayesian framework, prior knowledge becomes part of the model as the prior of the parameter; data on known deterioration time is entered, updating the distribution over the parameters. The posterior pdf of a Weibull distribution is:

$$f(\beta, \eta | \, Data) = \frac{L(Data|\beta, \eta)\varphi(\beta)\,\varphi(\eta)}{\iint_0^{\infty} L(Data|\beta, \eta)\varphi(\beta)\,\varphi(\eta)\, d\beta d\eta} \tag{2}$$

$L(\beta, \eta)$ is the likelihood function based on Weibull distribution (Equation 1) and Data $1, \ldots, n$:

$$L(Data|\beta, \eta) = \prod_{i=1}^{n} f(t_i|\, \beta, \eta) = \prod_{i=1}^{n} \frac{\beta}{\eta}\left(\frac{t_i}{\eta}\right)^{\beta-1} e^{-\left(\frac{t_i}{\eta}\right)^{\beta}} \tag{3}$$

Given Equation 1 and 2, the posterior distribution of deterioration time is:

$$f(T|\, Data) = \iint_0^{\infty} f(T|\, \beta, \eta)\, f(\beta, \eta|\, Data)\, d\beta d\eta \tag{4}$$

This Bayesian parameter estimation framework has been applied in several studies to predict asset deterioration, for example, in Enright and Frangopol [10] for bridge condition prediction and in Hong and Prozzi [11] for road pavement performance prediction. Extended uses of Bayesian parameter estimation were also developed to consider more practical assumptions, such as data with censorship (e.g. Lu [12] and Coolen [13]) and deterioration across multiple states (e.g. Han et al. [14]).

Figure 1 presents a BN model constructed on these principles developed from Zhang and Marsh [4]. The time each asset transits from one state to another state follows a Weibull distribution, which can be inferred when data on the past transition of the assets of the same class are entered as evidence. Take Asset 4 as an example, the entered data 60 representing it takes 60 months for Asset 4 to transit from S2 to S1. While for Assets 1 to 3 and 5, their data are censored. Instead of entering observations, we modelled them with additional Boolean variables as constraints on the deterioration time. To represent a left-censored data, for example, Asset 1 has a deterioration time less than 24 months, the variable is expressed

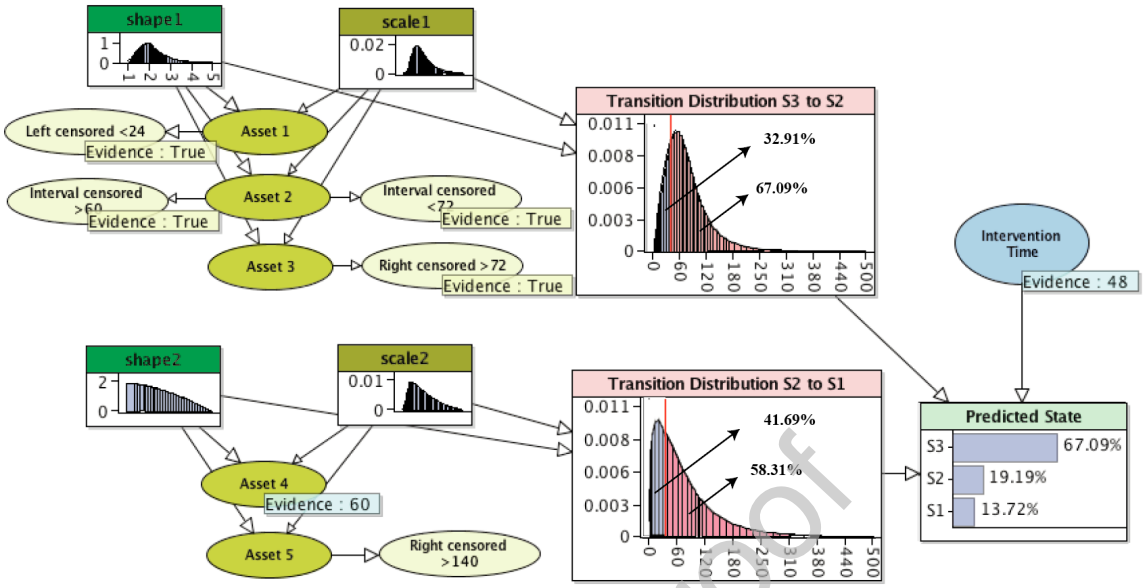


Figure 1: Condition prediction of a multi-state asset after 48 months.

with a logical expression $if(T_{Asset1} < 24, \text{“}True\text{”}, \text{“}False\text{”})$, and the **True** state is observed. Similar constraints are used for right-censored data. For interval-censored data, two Boolean variables are built applying the same principles.

Because Assets 1 to 3 are considered as the same type of assets, we assume that they deteriorate following the same Weibull distribution, meaning that they share the same shape and scale. The posterior distribution of transition time from S3 to S2 is used to predict future deterioration from S3 to S2 for assets in the same class. It is showed as the variable Transition Distribution S3 to S2 in the figure, derived from the parameters learnt from data and prior knowledge of shape 1 and scale 1. The same principle applies to transition from S2 to S1.

$\varphi(\beta)$ is the prior of the parameter shape β, and $\varphi(\eta)$ is the prior of parameter scale η. Distributions, such as normal distribution or uniform distribution, could be used to express uncertainty over the parameters as their prior distributions. Some experts find it easier not to specify their opinions with absolute precision but providing value intervals [15]. Following Marquez et al. [16], triangular distribution, measure the uncertainty by a collection of lower limit, median and upper limit, is used in this model for its advantage in extracting expert knowledge into value ranges.

To model the deterioration of assets across multiple states, we need to learn and link multiple distributions for prediction. Assuming the asset is currently in S3, Figure 1 presents an example to predict its further deterioration after 48 months. Therefore, three states (S3, S2 and S1) are modelled using a categorical variable, with the transition from S3 to S2 (T_3), and S2 to S1 (T_2). Each transition is modelled by a separate parameter learning model. We assume deterioration progresses with the sequence of the rating system, that is, the transition from S3 to S1 must go through the transition of S2. Hence, denotes Interven-

tion Time as T, in this example, the expression for the predicted state variable becomes: $if(T < T_3, \text{“}S3\text{”}, if(T < T_2, \text{“}S2\text{”}, \text{“}S1\text{”}))$. Since the starting state of this asset is in S3, the query from Intervention Time will first visit the Transition Distribution S3 to S2. In the pdf of this variable, only 32.91% of the area of this distribution is smaller than 48 months, that means, 67.09% of this asset will still stay at S3. For those transit to S2, only 41.69% will further transit to the S1 showed in Transition Distribution S2 to S1, that is, 32.91%*41.69% = 13.72%. The deterioration prediction distribution is, therefore, showed in variable Predicted State.

We have introduced how to learn deterioration from data with uncertainty and knowledge, and how to use them to predict multi-state deterioration. However, to assign knowledge, for example, this asset has a faster deterioration compares to others, into the priors, we need to interpret the behaviours of the parameters from a non-statistician perspective so that the engineers can be confident in expressing the knowledge. Though including expert knowledge within the Bayesian framework has already significantly reduced the need for large-sized data, we would also like to make use of the feature information to provide more accurate prediction for individuals.

2.2. Learning From Features

A framework to give individualised deterioration predictions was developed in Chang [17]. It first reduced the feature dimension of the dataset to select a small subset of features as important features, and then assets with the same feature values were grouped. Each group was modelled in the form of a Markov model to model multi-state deterioration, the transition probabilities of the Markov model were learned using logistic regression. Individualised deterioration prediction was then performed based on the group the asset belongs to.

To reduce the dimension, in Chang [17], covariance analysis was first conducted to remove highly correlated variables and penalised regression was later performed to rank features based on their importance in deciding the deterioration time. Various studies have been proposed to study the feature importance [18], of which, random forest is one of the most popular methods. Unlike penalised regression applied in Chang [17], random forest does not assume a linear relationship between the features and the response variable (i.e. deterioration time). Random forest is an ensemble method in which variable importance evaluation is performed by voting multiple independently developed decision trees from bagged training samples. Two importance measurement functions of the features proposed by Breiman [19] are commonly used in the random forest: Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA). Strobl et al. [20] made a comparison between these two measurements. It reveals that MDI measurement is biased towards features with more categories. The bias is a challenging problem for the NBI dataset because it involves with mixed data types, where some are binary (e.g. whether the structure is flared) and some have more than 20 categories (e.g. feature maintenance responsibility has 29 categories). Though Strobl et al. [21] later claimed that MDA measurement is also biased, it favours features with highly correlated variables. To avoid this, pre-processing to remove highly correlated variables between each feature is necessary.

Though Chang [17] provides us with a framework to perform individualised prediction, the challenges of uncertainty in the deterioration data and learning deterioration for those groups with small data size, remain untouched. Since it is possible to separate assets into different groups by their feature values, it becomes more interesting if we could learn between these groups considering we understand the influence of features on their deterioration rates.

We can extend the BN models with hierarchy to leverage the learning of groups with little data from groups with more data. In a hierarchical BN, the parameter of the prior distribution is called its hyper-parameter (that is, the parameter of the parameter); the result is that we can model the uncertainty about the parameters themselves. The prior of the local parameter represents the local deterioration behaviours of that subgroup, while the prior of the hyper-parameter (called hyper-prior) governs the parameters of all the subgroups. Therefore, by modelling multiple layers of information, hierarchical BN can leverage different groups by their similarity, and further, to provide individualised deterioration prediction [22, 4]. But to construct a hierarchical BN model, we still need to elicit the prior distribution for the hyper-parameter.

Developed from Neil et al. [23], Zhang and Marsh [4] developed a collection of generic BN models for asset management. They also extended the modelling to include censored data and multi-state asset modelling as we introduce in section 2.1. However, in their paper, the prior elicitation is not explained in details, the selection of features and their impact on the parameters are estimated purely by experts, and the model is built on a hypothetical problem instead of a real case study. Asking experts to assign relevant features and quantify their impacts on the deterioration rate would require high-level expertise. This request is difficult in practice, especially for cases like the NBI database, where each bridge is associated with over a hundred features. A more realistic method is required.

In the following, we extend and instantiate the models developed from Zhang and Marsh [4] using the real deck structure records in Wyoming from the NBI dataset to predict multi-state asset deterioration. We show how to select a subset of representative features and develop a modified hierarchical BN model that can learn the influence of different features on deterioration rather than entirely estimated by experts. We also provide a comprehensive and intuitive explanation about how to derive knowledge to represent the priors of the (hyper) parameters of a Weibull distribution. At last, by comparing with other existing methods, we show the advantages of our methodology in individualised deterioration prediction.

3. Hierarchical BNs for Individualised Deterioration Prediction

This section extends the previously discussed BN models into hierarchical BN models so that we can learn deterioration between groups. We illustrate this process with a case study about bridge deck structure deterioration in Wyoming in the NBI database. The NBI database encodes over a hundred features for each bridge. Therefore, before building a deterioration prediction model, we first reduce the dimensionality of the database to a small number of features that is representative enough in deciding the deterioration rate. The selected features are used to classify bridges into groups based on their individual feature levels. For each transition, a deterioration model that can learn parameters from other groups is built. They

are further assembled to predict deck structure condition that is rated by multiple states. At last, we show how we can elicit expert knowledge for the parameters of a Weibull distribution so that we can assign values for the priors.

3.1. Dimension Reduction in NBI dataset

In addition to the over a hundred features recorded in the NBI database, feature age, an indirect feature is also generated in our study due to its popularity in predicting deterioration [17]. The age of a bridge is calculated by the duration between the year of built and the corresponding inspection year. The reconstruction year is not used to infer bridge age because of the poor quality of this variable. For example, out of 3,127 bridges in Wyoming in 2017, there are only 168 bridges that have valid records in this variable, and among these bridges, some decks' conditions remain unchanged or even deteriorated since the previous inspection. This could be caused by entry errors [24] or they only replaced or rebuilt one part of the bridges (e.g. just superstructure or substructure) [17]. An exploratory data analysis was then performed to reduce the quantity of the features: features with no association to deterioration rate, for example, structure number; and for features where over 95% of their population are missing, such as critical feature inspection date, are removed.

A preliminary correlation matrix, of which Pearson correlations are used for continuous variables, and polychoric correlations are used for categorical variables, is developed to measure the statistical relationships between pairwise variables. This results in some strongly correlated variables, for example, feature maintenance responsibility and feature owner of the structure has a 0.94 correlation. One of them is considered redundant. The latter feature is removed from the feature candidate due to feature maintenance responsibility has a higher correlation with the deterioration time. This process results in a candidate pool of over 40 features.

Lastly, to further reduce the dimensionality of the dataset, feature selection is performed using an R package called Boruta [25], which is built on a modified random forest to evaluate the feature importance. First, Boruta adds and shuffles the value of duplicated features from the original features to remove their correlation with the response variable deterioration time. These new features called shadow attributes are combined with the original feature space. Then a random forest classifier is performed, and MDA measurement is used in our case. Denote $\mathcal{B}^t$ as the out-of-bag samples for a tree t and $L(T_t(X_i), y_i)$ as the accuracy at the ith training sample measured by root mean square. Let $X_{i,\,\pi_j} = X_{i,1}, ..., X_{\pi_j(i),j}, X_{i,j+1}, ..., X_{i,p}$, and π_j is a random permutation of n integers. Follow Breiman [19], the importance for feature Xj is computed as the sum of the importance over all trees (*ntree*) in the forest is defined as

$$Imp\left(X_j\right) = \frac{\sum_{t\in\mathcal{B}} \sum_{i\in\mathcal{B}^t} L\left(T_t\left(X_i\right),\ y_i\right) - L\left(T_t\left(X_{i,\,\pi_j}\right),\ y_i\right)}{ntree} \tag{5}$$

The bias caused by MDA metric is avoided thanks to the pre-processing procedure mentioned before by removing highly correlated variables. Additional Z scores among those shadow attributes are computed by dividing the mean of accuracy loss by its standard deviation in order to take accuracy loss fluctuation into account. Features with significantly high

Z scores are tagged as important and vice versa. This process is repeated until all attributes are labelled. Details of this algorithm can be found in Kursa and Rudnicki [25].

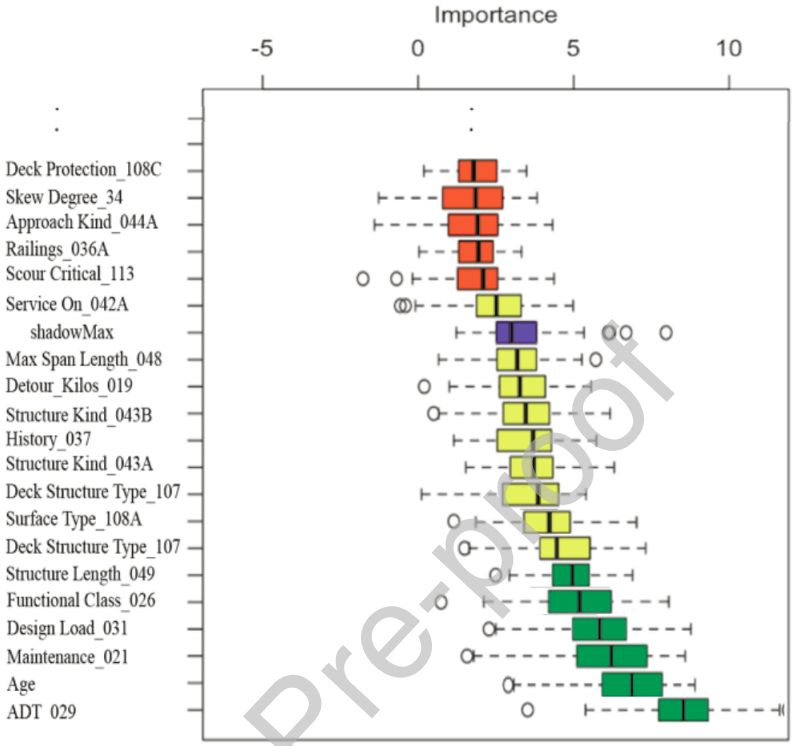


Figure 2: Feature selection for deck structure in Wyoming from the NBI database.

Figure 2 shows the features importance for our case study. Rankings of features are separated by the Z score, which is coloured by blue box indicated as shadow attribute. Green boxes represent accepted features that have higher importance values, that is, are more predictive in evaluating deterioration time. Yellow boxes are tentative features that are medium significant, which are taken into a backup plan when needed. Red boxes are rejected or insignificant features. They are removed from the feature candidate pool. The labels on the left side are the features from the NBI database, and the numbers associated are their item ID, which can refer to the detail explanation of each feature in the NBI manual [26]. A brief introduction of each accepted feature follows the sequence of its ranking is:

- **ADT**: average daily traffic, records the most recent average daily vehicles traffic volume on the structure.
- **Age**: an indirect feature generated from Item 27 built year and Item 90 inspection date, represents the current age of the structure.
- **Maintenance**: the maintenance agency that is responsible for the structure. Notes that different agencies may have different inspection training and regulations.
- **Design load**: the designed live load of the structure.

- **Functional class**: the functional classification of structure, for example, whether it is designated in a rural area or urban.
- **Structure length**: the length of the roadway supported by the structure.

3.2. Assigning Feature Levels

After the important features are identified, the possible values of each feature are categorised into several levels representing their indication of deterioration time. Three levels, from low, medium to high are used here. We can use a linear scalar variable in the BN, for example, a ranked node [27], to model each feature.

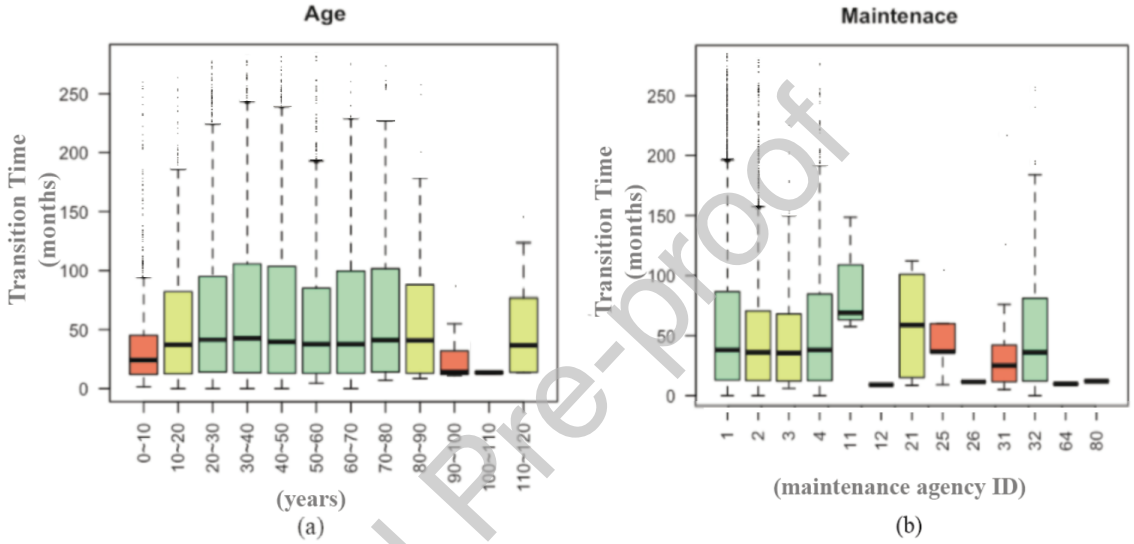


Figure 3: Assigning feature levels: (a)feature age; (b)feature maintenance. In the boxplot, the box measures the spread of data, called interquartile range (IQR). The box marks the lower (Q1) percentile to upper (Q3) percentile of the data, and the black line within the box marks the median. Follow the rule of interquartile range, data falling outside the Q1-1.5IQR to Q3+1.5IQR range (represented as whiskers in the plot) are identified as outliers.

Figure 3 presents two examples of how to assign the level of each feature. In the figure, the red bar represents level low (meaning this structure has a higher probability to deteriorate faster), the yellow bar represents level medium, and the green bar represents level high. Since feature age is a continuous variable, it is first discretised into several bins with a 10-year interval. By plotting against the deterioration time within the training dataset, we can see different age categories have different deterioration time distributions. For example, for age between 0 to 10, though several outliers having long deterioration time, most of them deteriorate within 50 months. Therefore, it is rated as low. Feature maintenance is a categorical variable. We can see that for example, a structure that is maintained by agency 1 - state highway agency, normally has a longer deterioration time, it is rated high.

3.3. Building Hierarchical BN Model to Learn Transition Distribution Between Groups

After the level of each feature is assigned, bridges can be separated into groups by their feature levels. Assets with the same combination of feature values can be grouped into the

same group assuming having the same deterioration characteristics. While for assets within different groups, their feature values can be an indication of how similar their deterioration is.

For each transition, we build a hierarchical BN model to learn the transition distributions between different groups as shown in Figure 4. Notes that for demonstration simplicity, Figure 4 only shows the top two most important features, ADT and Age. Later in the validation, we will discuss how many features we should consider. In this example, assets with a low feature level in ADT and a medium feature level in Age are grouped as Group 1 (Asset 1, 23 and 25); assets with a high feature level in ADT and a medium feature level in Age are grouped as Group 2 (Asset 4 and 9).

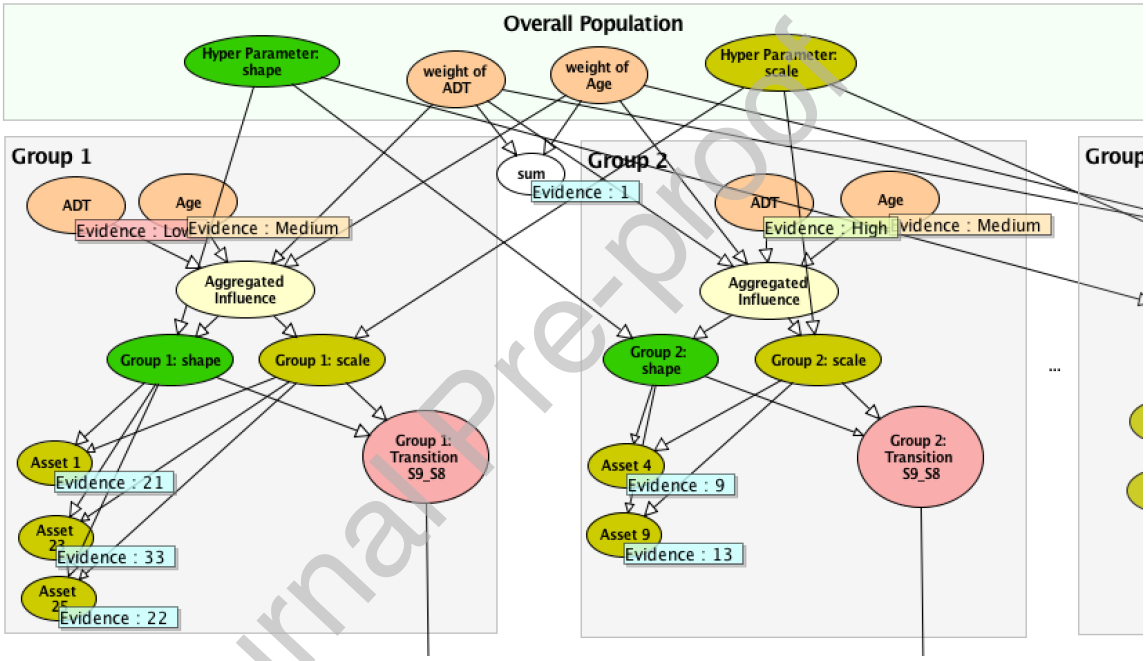


Figure 4: Hierarchical Bayesian Network to learn between groups.

Different features may have different strengths (weights) in influencing the deterioration of assets. For example, the environmental condition may have more influence on the deterioration of a metal bridge than its service type. Experts could have knowledge about the weights of these features and assign them directly. In the case where we lack this type of knowledge, we can assign a hyper-parameter for each feature that represents its weight (variable weight of ADT and weight of Age). For feature k, the weight w_k is modelled with a prior of an uniform distribution with a lower bound L_k of 0 and upper bound U_k of 1. All the weights of the feature are linked together with a variable that sums the weights to 1 (variable sum). The weights converged and learned where there are many different groups of assets (with varying combinations of the feature) with deterioration data. However, if we do not have many groups, due to the convergence, it is better to rate weights by experts or to give more informative priors for the weight variables, rather than learn them. The priors for the

overall population hyper-parameters are defined as below, where $Z = \{\beta, \eta\}$:

$$w_k \sim Uniform\,(L_k, U_k) \tag{6}$$

$$\varphi_Z \sim Triangular\ (a_Z, b_Z, c_Z) \tag{7}$$

In the hierarchical BN model, within each group i, instead of eliciting $\varphi_{i,z}$ from experts, the prior of the local parameter is learned from hyper-parameter φ_Z together with an indicator called aggregated influence $AggI_i$ resulting from each group's feature levels $FL_{i,1}, ..., FL_{i,k}$. The syntax for the parameters in Group i:

$$FL_{i,k} \sim Ranked\,(L_{i,k}, M_{i,k}, H_{i,k}) \tag{8}$$

$$AggI_i|FL_{i,1}, ..., FL_{i,k}, w_1, ..., w_k \sim TNormal(wmean(FL_{i,1}, w_1, ..., FL_{i,k}, w_k), \sigma_i^2, 0, 1) \tag{9}$$

$$\varphi_{i,z}|AggI_i, \varphi_Z \sim TNormal(Partitioned(AggI_i, \varphi_Z),\ \sigma_{i,z}^2,\ L_{i,z},\ U_{i,z}) \tag{10}$$

Each feature level $FL_{i,k}$ is modelled by a ranked node, which is an ordinal categorical variable that can be observed from low, medium to high. The aggregated influence is essentially a linear scalar that aggregates all the relevant features using a Truncated Normal (TNormal) distribution into the range of 0 to 1 [27]. Its mean is the weighted mean of the feature levels, and the variance σ^2 represents the certainty about the aggregation. The influence is further mapped to low (0 to 0.333), medium (0.333 to 0.666) and high (0.666 to 1) three levels, where for example, a low level influence indicates a faster deterioration. Each parameter is then partitioned by the level of the aggregated influence and its corresponding hyper-parameter φ_Z following another TNormal distribution. Next, we will introduce how to elicit priors for a Weibull distribution so that we can elicit the variance $\sigma_{i,z}^2$, representing the uncertainty about the grouping, as well as lower $L_{i,z}$ and upper $U_{i,z}$ limits about the deterioration time.

3.4. Prior Knowledge Elicitation of A Weibull Distribution

It is difficult for non-statisticians to evaluate the values of shape β and scale η directly, but can be made easier by understanding the characteristics of the distribution and trends of parameters. This subsection focuses on the interpretability of the characteristics of parametric statistical distributions using the example of a Weibull distribution.

3.4.1. Shape Parameter

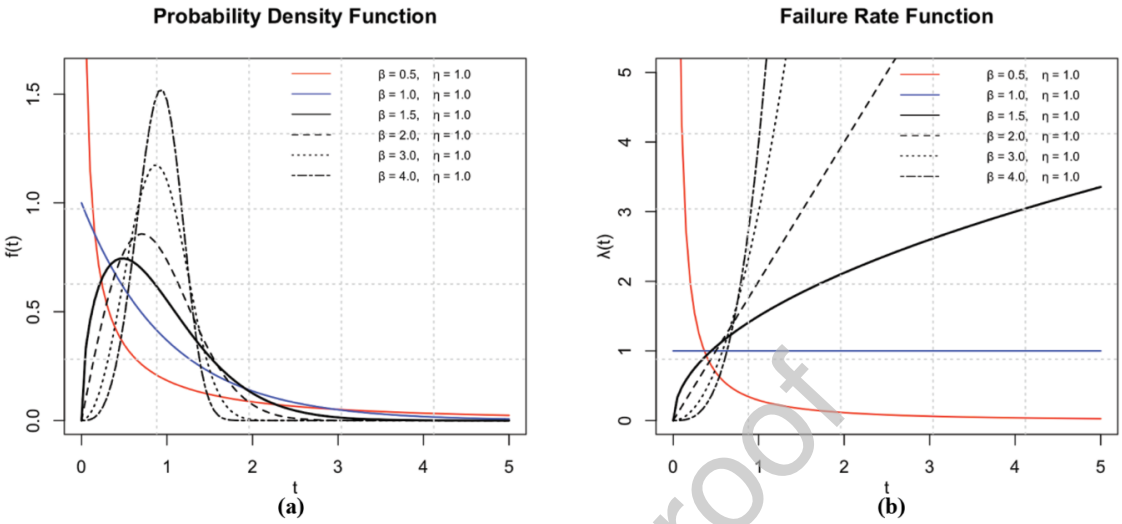


Figure 5: pdf and failure rate function of Weibull distributions with different shapes.

Figure 5 shows examples of the Weibull distributions' pdfs with different shape parameter β. Empirically, the shape parameter of Weibull distribution offers us great flexibility in modelling a variety of distribution with diverse physical behaviours: it becomes an exponential distribution when β equals to 1, a Rayleigh distribution when β equals to 2. Also, studies have shown that the skewness of Weibull distribution has a strong relationship with the value of β [28]. The skewness decreases with β: it has a positive skewness to the left side of the pdf when β is smaller than 3.6, and it approximates symmetrically as a normal distribution when β is near 3.6 with a skewness value approaching 0, and it skews to the right with a negative skewness when it is greater than 3.6 [29].

To understand its characteristics more intuitively, we have its cumulative distribution function (cdf), which is also known as the distribution's unreliability:

$$F(t|\beta,\eta) = \int_0^{\infty} f(t|\beta,\eta)\,dt = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{11}$$

While the survival function (reliability) of a distribution is the complementary of the cdf, representing the probability of the asset will survive after a given time:

$$S(t|\beta,\eta) = Prob[T > t] = 1 - F(t|\beta,\eta) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{12}$$

By representing Event A a situation where the asset will fail between a small enough interval between t and $t + \Delta t$, and Event B a situation where the asset will survive after time t, using the laws of conditional probability, the probability of Event A given Event B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{f(t|\beta,\eta)\,dt}{S(t|\beta,\eta)} = \lambda(t|\beta,\eta)\,dt \tag{13}$$

From this, we have the instantaneous failure rate function at any time point t:

$$\lambda(t|\beta,\eta) = \frac{f(t|\beta,\eta)}{S(t|\beta,\eta)} = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \tag{14}$$

As a natural extension of exponential distribution (when shape $\beta = 1$), Weibull has a polynomial failure rate with an exponent $(\beta-1)$. The characteristics in the failure rate function are useful to help experts define the priors of the parameters. The corresponding failure rate functions of the distribution in Figure 5 (a) are showed in Figure 5 (b). A β value that is between 0 and 1 describes a decreasing failure rate over time. This often happens when an asset is in the burn-in phase that shows early degradation, which may be caused by the problematic building process or infrastructure operation. In the case where we believe the asset's failure rate would not change over time, that is, it has a constant failure rate, we can suggest β equals to 1. When describing infrastructure in a wear-out failure phase, we can recommend the prior of β greater than 1 indicating an increasing failure rate.

3.4.2. Variance

By looking at Figure 5, we can also notice that with the increase in shape parameter β, the pdf gets narrower, which represents a smaller variance. The variance σ^2 of a Weibull distribution is:

$$\sigma^2 = \eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right] \tag{15}$$

validated in Abramowitz and Stegun [30], the difference between $\Gamma(1+\frac{2}{\beta})$ and $(\Gamma(1+\frac{1}{\beta}))^2$ becomes smaller with the increase in β, that is, the variance σ^2 decreases with the increase in β. When $\beta \to \infty$, the variance of the Weibull distribution approaches 0.

3.4.3. Scale Parameter

By holding the same shape β, increase the scale parameter η has an effect of stretching out the pdf, this can be seen in Figure 6 (a): with the increase in η, the range of the pdf gets wider with a lower crest. The quantiles of the distribution can explain this, given the cdf in Equation 11, by setting $F(tp) = p$, we have:

$$t_p = \left\{ln\left(\frac{1}{1-p}\right)\right\}^{\frac{1}{\beta}}\eta \tag{16}$$

As shown by the red and black solid lines in Figure 6 (b), having the same shape β, it takes longer to reach the same quantile with a higher scale η. Additionally, when time to failure t equals to η, from the cdf in Equation 11, we have $1 - e^{-1} \approx 0.632$, that is, as shown in the shaded area in Figure 6 (a) and (b), regardless the value of β, when time t equals to η, 63.2% of the population will fail. Together with the skewness information of Weibull distribution mentioned in the shape parameter, this can give experts confidence to estimate the scale parameter η given knowledge of the average failure time.

Also, given the mean of Weibull distribution μ is,

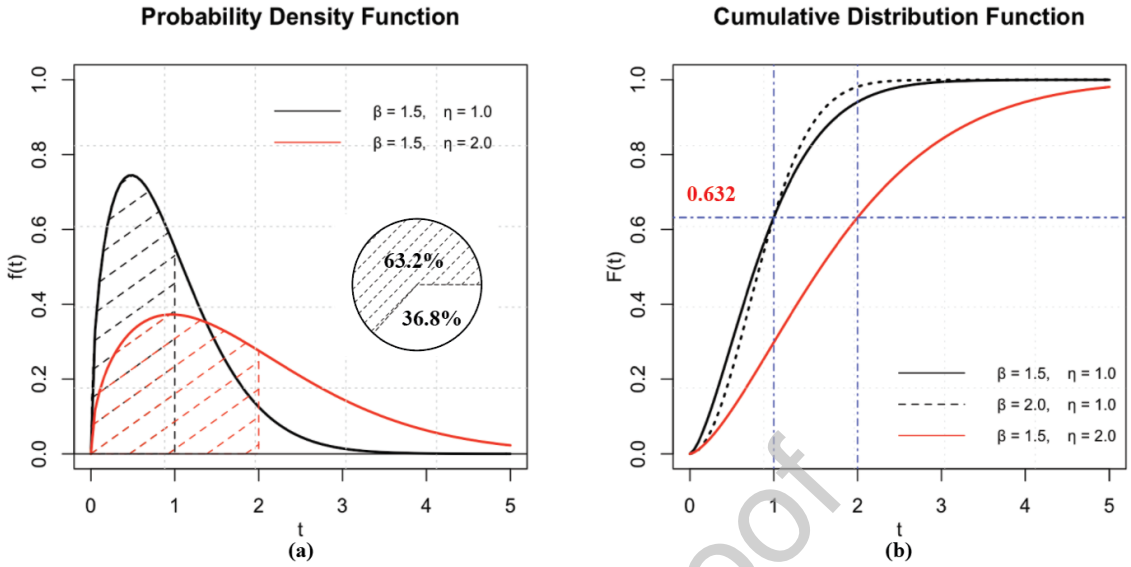


Figure 6: (a) pdf of Weibull distributions with different scales; (b) cdf of Weibull distributions with different shape and scale.

$$\mu = \eta\Gamma\left(1 + \frac{1}{\beta}\right) \tag{17}$$

we have, $\frac{\eta}{\mu} = \frac{1}{\Gamma(1+\frac{1}{\beta})}$. Given $\beta \geq 0$, we have $1 + \frac{1}{\beta} \geq 1$. From Figure 7 (a), we can see that for a gamma function $\Gamma(x)$ with $x \geq 1$, it has a minimum value at $\Gamma(1.462..) \approx 0.886$. Therefore, the maximum of $\frac{\eta}{\mu} \approx \frac{1}{0.886} \approx 1.129$, and from $1 + \frac{1}{\beta} = 1.462$, we can obtain the corresponding shape $\beta \approx 2.166$. This can be interpreted as the maximum of scale is 1.129 times of the mean of the Weibull distribution μ. This validates the examples in Figure 7: given a mean of the Weibull distribution μ, with the increase in shape value β, the scale η reaches its peak when the value of shape at around 2.166, which is about 1.129 times of its mean μ. And after that point, the value of η decreases slightly and gradually stabilised reaching its mean μ when $\beta \to \infty$. By knowing this, experts can form an idea of the relationship between the value of scale and mean time to failure, and provide more confidence for them to extract knowledge for the prior of scale.

3.4.4. Summary

To quantify prior knowledge from experts, it is vital to explain the characteristics of Weibull distribution in a way that engineers can understand. Instead of asking their opinions about the shape parameter β directly, we may interpret the question into the form of failure rate. For example, with the increase in time, does the asset get more likely to deteriorate? If the answer is yes, we can define a prior for β that is greater than 1.

Also, we assume experts are more likely to have knowledge such as the mean deterioration time of this type of asset μ, which can be used to narrow down the range of β and estimate the mean of the scale parameter η. Given the μ, in the overall population distribution of

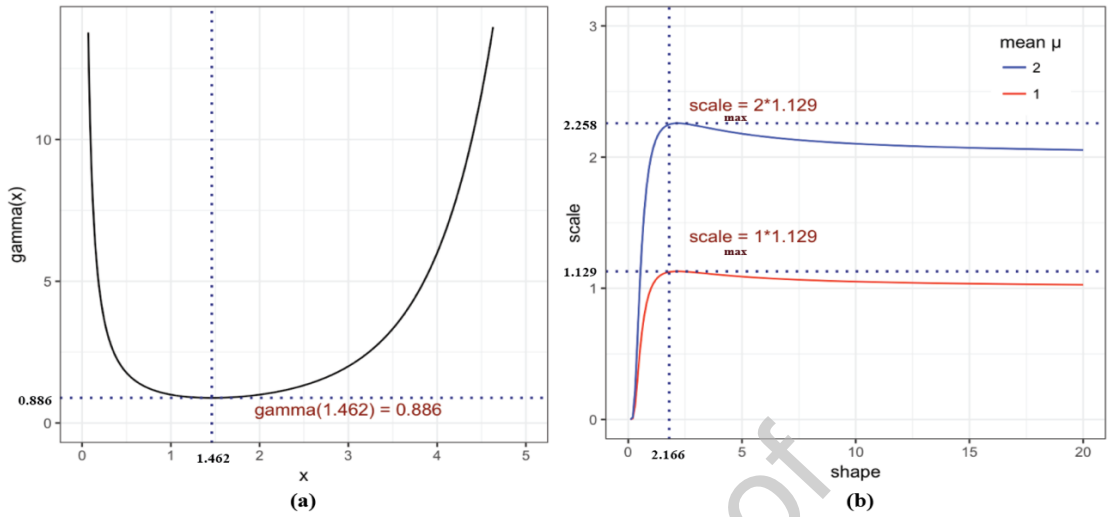


Figure 7: (a) gamma function; (b) mean μ of Weibull distributions.

this type of asset, whether the assets are more likely to deteriorate before μ or after? If the answer is before μ, we can interpret it as a left-skewed distribution, which represents the β is ranged from 1 to 3.6. This is also validated by most studies, where most infrastructure's failure time follows a β between 1 to 3.6 (see examples in Le [9]). To further narrow down β, we can increase its left bound if the experts are very confident in providing the above information. This is contributed by the information that the higher β is, the smaller variance σ^2, as discussed before.

Given the information of the μ and the skewness, and the interpretation that the value of scale parameter η represents 63.2% of the failure population, we can estimate the mean value of η is slightly higher than μ. Also, given the maximum of the scale η is 1.129 times of its μ, we can use it to set the upper bound of η.

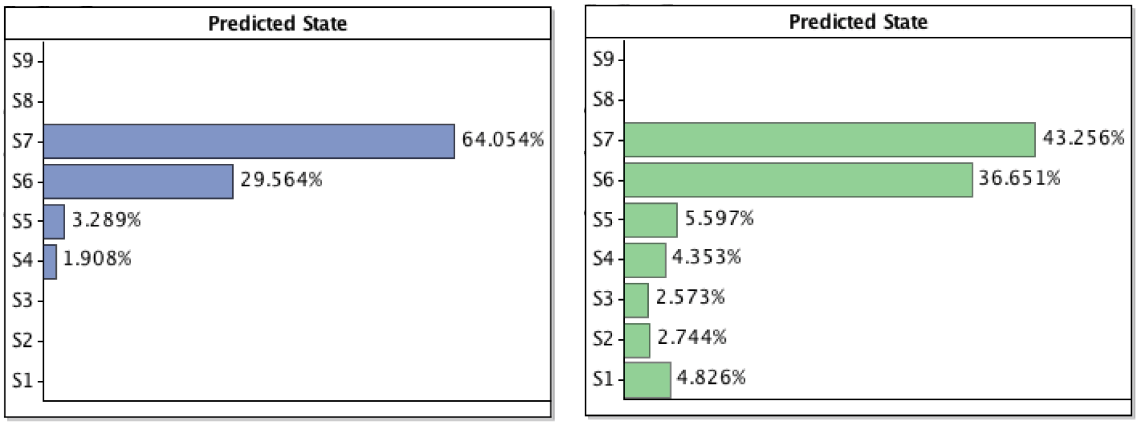
After the deterioration models for all transitions are built and priors for the (hyper) parameters are obtained, we assemble them for the multi-state deterioration prediction follows the model discussed in section 2.1. For each deck, we query the related transition models based on the group it belongs to, and enter its observation of the intervention time to produce an individual prediction.

4. Validation of Deterioration Prediction

The section introduces how to measure the performance of a probabilistic prediction, and investigates the prediction performance of the proposed approach from different perspectives.

4.1. Measurement Metric

Accuracy rate is one of the most common measurement metrics by evaluating the fraction of the classification. For example, given a structure in year 2010 is in S7, we want to predict its condition in year 2017: if our prediction is S6 which matches the actual inspection result



(a) (b)

Figure 8: Two predicted condition distributions.

in year 2017, we have an accuracy rate of 100%. By repeating this process for all the test datasets, we can have the average accuracy rate to evaluate the performance of the model.

However, the inspection result itself is often uncertain. In most cases, inspection is carried out visually, though inspectors are trained with standard inspection manuals, the inspection result is still covered with a range of noise. The noise comes from, such as the inspection tools they used, environmental factors, or human subjectivity. An experiment was carried out by Phares et al. [31], which shows that for the same structure, 68% of the ratings given by different inspectors fell into a one-point interval range differences, and 95% fell into a two-point interval range differences. Therefore, in practice, when predicting structure's deterioration, rather than providing a single estimation of the condition, a probabilistic prediction like what our models produce, that considers variation for a multi-state prediction may be more applicable.

Given a query, the developed BN model predicts and outputs a probabilistic distribution over a discrete range of possible outcomes, rather than the most likely outcome. Figure 8 presents two predictions for the condition distribution of a deck after 45 months given its initial state is at S7. Though the actual inspection result is this structure deteriorated into S6 after 45 months, that is, both predictions are wrong if we only consider the accuracy rate by selecting the most likely state. However, compare to result in Figure 8(a), (b) gives a better prediction with a closer distance to S6.

To take this phenomenon into account, Ranked Probability Score (RPS) is chosen for its ability in measuring multi-state probabilistic prediction performance with orders. RPS is an extension of the Brier score for multi-class classification that also takes the distance between different possible outcomes into account. Assume there are K categorical events, the cumulative observation X_m and prediction Y_m can be defined by a vector of the observation's probability components and a vector of the prediction's probability components respectively::

$$X_m = \sum_{k=1}^{m} x_k,\ Y_m = \sum_{k=1}^{m} y_k\ , m = 1,\ \ldots,\ K \tag{18}$$

the RPS is the sum of the squared difference between the components of these two vectors

$$RPS\ = \frac{1}{J} \sum_{k=1}^{J} (X_m - Y_m)^2, J = K - 1 \tag{19}$$

A perfect prediction, for example, event k, would assign all the probability to x_k ($x_k = 1$), so the difference would be 0 with an RPS = 0, while the worst score is $K - 1$ due to the accumulation [32]. That is, the smaller the RPS we have, the better the prediction we made. In the examples in Figure 8, (a) has an RPS of 0.052 while (b) has a score of 0.033. Hence, (b) is a better prediction. This calculation yields the RPS for a single event. To evaluate the performance of a prediction for a collection of events n, the average RPS can be defined as

$$R\bar{P}S\ = \frac{1}{n} \sum_{i=1}^{n} RPS_i \tag{20}$$

In addition to accuracy rate and RPS, Kappa statistic and Kendall's coefficient are another two commonly used metrics for classification problems. Kappa statistic handles imbalanced classification problems well, it compares the observed agreement between observation and prediction with the expected agreement between observation and random guess according to the frequency of each class. It ranges from 0 to 1, and Landis and Koch [33] suggest that Kappa statistic <0.20 indicates the performance is poor, 0.21-0.40 is fair, 0.41-0.60 is moderate, 0.61-0.80 is good and >0.81 is very good. Kendall's coefficient handles ranked multi-class problems well, it is a non-parametric test that assesses the agreement among the prediction by computing a normalised score for concordant rankings between the observation and prediction. It ranges from 0 indicating no agreement to 1 indicating complete agreement.

4.2. Number of Features

If all relevant factors are used to distinguish structure into different groups, then there is likely to be insufficient historical data to estimate every transition distribution parameters, even with data-rich group. Therefore, groups need to be defined by the features that are most important to the deterioration time but within a limited amount. After the feature selection in section 3.1, there are 6 accepted features considered as important. However, each feature is quantified into 3 levels depending on the values, considering 6 features would result in $3^6 = 243$ groups, which is not only expensive for prior elicitation, but also expensive to do inference within the model.

The NBI deck information from year 1992 to 2010 in Wyoming are used to learn the deterioration distributions of S7 (because this state has the largest amount of data). To mitigate the influence of prior knowledge on the prediction performance, all the priors in this section are provided uninformatively with an uniform distribution from 0 to 20 in the shape parameter and an uniform distribution from 0 to 500 in the scale parameter. We predict whether a

deck in S7 will deteriorate into another state after the next inspection to investigate how the performance changes over the number of feature (from 0 feature to 6 features).

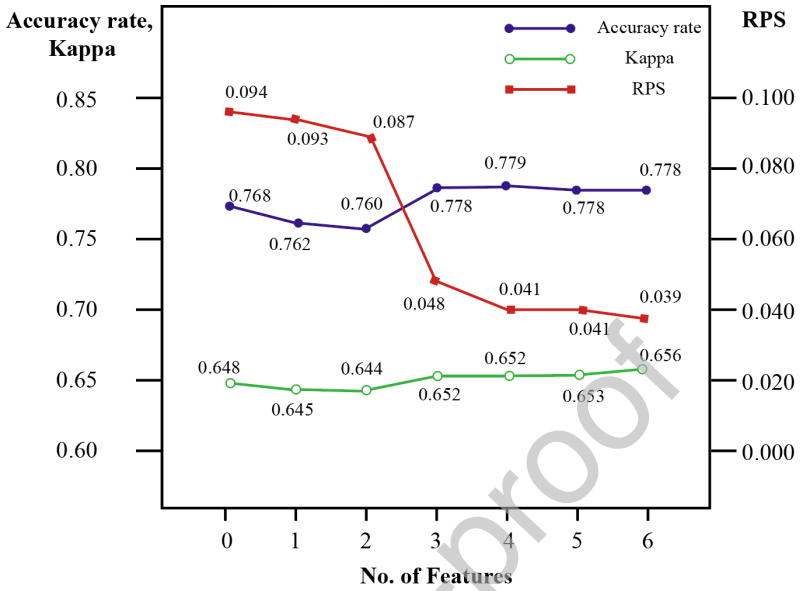


Figure 9: Accuracy rate (higher better), Kappa (higher better) and RPS (lower better) performance with the increase in feature amount.

Since this experiment is a binary classification problem (stay in S7 or not), we only measure the performance with accuracy rate, Kappa statistic and RPS. Figure 9 presents the result of the experiments. The blue line measures the average accuracy rate by the most likely state (the higher, the better), green line measures the Kappa statistic (the higher, the better) and the red line measures the average RPS (the lower, the better). Though there is a slight decrease in the accuracy rate when only one or two features are considered, in general, the accuracy rate increases with the increase in the number of features. While the RPS and Kappa performance always gets better with the increase in the number of features, that is, the prediction gets closer to the actual inspection.

After a drastic increase in the performance when 3 features are considered, all measurement metrics exhibit steady performance afterwards. Reason for this steady performance could be most of the structures may possess the same set of feature combinations. For example, with 6 features, there are 243 possible groups with different feature combinations, only 92 of them exists in the database. Among these 92 groups, 80% of them have less than 2 candidates. Therefore, it becomes unnecessary to consider too many features to form groups. Since considering 3 features already gives us a reasonable performance, to avoid the increase in modelling and inference complexity, it is applied in the following experiments.

4.3. Multi-State Prediction Performance

This subsection compares the multi-state prediction performance between different approaches. The NBI deck information from year 1992 to 2010 in Wyoming are used to learn

the deterioration distributions between different states. In this experiment, given the most recent deck condition in Wyoming by 2010, our model predicts its deck condition after the next inspection (second inspection). In total, there are 2249 testing data points in this case. Since no record of decks in S0 in the testing data set, only performance of S9 to S1 are considered here. The models are compared with several other approaches:

- **HierBN**: the hierarchical BN models developed from section 3 that learn between 27 groups separated by 3 selected features: ADT, age and maintenance, to give individualised predictions;
- **BN**: the BN models developed from section 2.1 that learn the parameters using the overall population deterioration data to give aggregated predictions;
- **MCLR**: a Markov model where its transition probability matrix is estimated using logistic regression designed for the same NBI dataset developed by Chang [17] to give aggregated predictions;
- **MCLR_G**: Markov models where their transition probability matrixes are estimated using logistic regression designed for the same NBI dataset but separated into 20 groups according to Chang [17] to give individualised predictions;
- **Mssurv**: a Markov model that use the Datta-Satten estimator to learn the transition probabilities. It allows the modelling of censored data in a multi-state system developed by Ferguson et al. [8]. It is implemented here to give aggregated predictions.

In order to measure the accuracy rate, in each prediction, the state with the highest probability is considered as the predicted state. Therefore, we can compare the prediction against the actual observation. As a result, confusion matrixes for different approaches are generated.

Table 1: Confusion matrix of deck condition prediction in Wyoming after the next inspection given the most recent condition by 2010 using HierBN

Observation	Prediction								
	S9	S8	S7	S6	S5	S4	S3	S2	S1
S9	**0**	0	0	0	0	0	0	0	0
S8	0	**28**	0	0	0	0	0	0	0
S7	0	12	**613**	2	0	0	0	0	0
S6	0	0	145	**657**	2	0	0	0	8
S5	0	0	21	131	**403**	0	0	0	0
S4	0	0	1	9	35	**105**	0	0	0
S3	0	0	0	2	3	7	**44**	0	0
S2	0	0	0	1	1	2	4	**11**	0
S1	0	0	0	0	0	0	0	0	**2**

Table 1 presents one of the confusion matrix for HierBN. The number marked in bold represents its prediction matches its observation. By evaluating the performance purely from the quantity of correct prediction, the proposed approach HierBN gives the best prediction with 1863 correct predictions, closely followed by MCLR_G with 1842 and MCLR with 1840.

However, using a confusion matrix for a multiple states classification problem may create confusion in deciding the accuracy rate for each transition. For example, in the testing dataset, there are five decks start with S9, four of them transited to S8, and one transited to S7 in the second inspection. To evaluate the accuracy rate for Transition 9 (T9, the transition from S9 to other states), it is difficult to reason from the above confusion matrixes since the records are overlapped with records that start with S8 and S7. Hence, in addition to the confusion matrixes, an accuracy rate for each transition is also provided. For example, using HierBN, we predict all five decks started with S9 deteriorated into S8 in the second inspection. This gives us an 80% accuracy rate for T9 in HierBN. In order to provide a more in-depth evaluation for each approach, we also provide the average RPS, Kappa statistic and Kendall's coefficient as suggested in Section 4.1. The results are summarised in Table 2.

Table 2: Multi-state prediction performance of decks in Wyoming after the next inspection given the most recent condition by 2010.

Methods	Accuracy rate								RPS	Kappa	Kendall
	T9	T8	T7	T6	T5	T4	T3	T2			
HierBN	**0.80**	**0.69**	**0.78**	0.81	0.90	**0.92**	0.92	**1.00**	**0.047**	**0.763**	0.859
BN	0.00	0.23	0.77	**0.82**	0.88	0.42	0.75	0.73	0.062	0.701	0.816
MCLR	0.00	0.40	**0.78**	**0.82**	**0.91**	0.89	0.92	0.09	0.053	0.747	**0.861**
MCLR_G	0.20	0.49	0.76	**0.82**	**0.91**	0.90	**0.94**	**1.00**	0.049	0.748	0.855
Mssurv	0.40	**0.69**	0.53	**0.82**	**0.91**	**0.92**	0.92	0.09	0.071	0.634	0.769

When predicting deck deterioration from states between S7 to S3 (i.e. T7 to T3), almost all methods have similar performance with relatively high accuracy rates. This is because the data amount within these states is considerably rich (over 300 data in each state). Hence, with enough data, these approaches perform similarly. Within these states, our approach guarantees a satisfying and steadily good performance compares to others. This is contributed by having successfully identified the small subgroups in our method and leveraged their deterioration learning specifically based on their features. Though the accuracy rates are not significantly higher than others, we suspect this is due to the population of the subgroup itself is small. Hence, its impact on all measurements is consequently small.

However, for data-poor states like S9, S8 and S2, our model gave a promising performance compared to others. Though in these states, MCLR_G also gave decent accuracy rates comparing to MCLR, we believe this is benefited from its policy to group similar structures for estimating the transition probability matrix individually. But the reason that our model outperforms MCLR_G and others in most cases is that we not only consider separating the population into related groups but also learn from each subgroup.

Figure 10 compares the real observed distribution with the distributions learned from BN and HierBN. These decks belong to a group that their features ADT, age and maintenance are all rated as low level. The distributions showed in the figure are the pdf of Transition 8 (T8). Without grouping and learning, a BN learned distribution based on limited available data is showed with the purple bins, while the learned distribution of HierBN is shown in yellow bins. The simple BN gives a relatively uninformative distribution with a wide range,

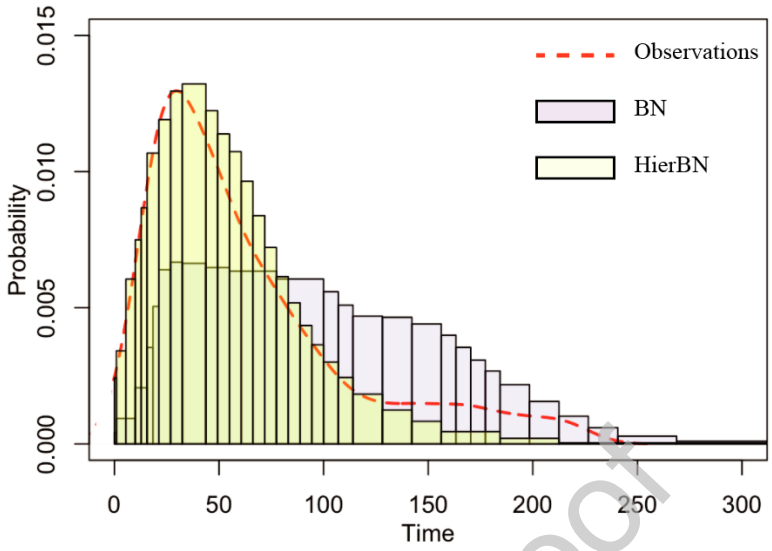


Figure 10: Probability density functions of Transition 8 (T8) of group rated as low-levels in all features.

but HierBN successfully predicts with a focus on early deterioration time that matches the pdf of the real observations (showed in red dotted line).

We measure the Kullback-Leibler (K-L) divergence to quantify the distance between the learned distribution and the observed distribution. The divergence ranges from 0 to infinity, and the smaller K-L divergence is, the closer the distribution is to the observed distribution. Comparing with the observation, distribution fitted by BN has a K-L divergence of over 2.3. It indicates the fitted distribution has a considerable distance with the observation. In contrast, fitted distribution by HierBN presents a satisfying performance with a K-L divergence of 0.041.

The last three columns in Table 2 give the average RPS, Kappa statistic and Kendall's coefficient of the overall prediction. The scores are not drastically different since the majority of the structure population are between S7 and S3 where almost all methods have good performance. The candidate pool of the other states' data is small, hence; the average performance of HierBN in RPS only excels slightly. All Kappa statistics are over 0.60 indicating good performance in classifying imbalanced data in all investigated approaches. So as Kendall's coefficients, indicating good performance in ordinal classification. Notes that Mssurv has a decent performance in accuracy rates but poor in RPS, Kappa and Kendall's coefficient compared to others, the reason we suspect is its predictions are usually too extreme and did not take too much uncertainty into consideration, which leads to a penalty in these metrics when the prediction is wrongly classified.

The complete BN model for this experiment contains more than 1000 variables where most of them are observed by the training data. The inference of this model was done using separate submodel for each transition, which varies in size and computational time. For example, it took less than 1 minute to compute some decks' transition from S7 to S6 as the

data was sufficient and do not require learning from others, while it took almost 1 hour for some decks' transition from S5 to S4 as it requires learning from other groups.

4.4. Future Prediction

This subsection investigates the prediction performance for future inspection. Same training data as in the previous subsections are used to learn the distributions, while data from year 2010 to 2017 are used to validate the performance of different approaches: experiments are performed to predict the deck structure's conditions after one year (year 2011) to seven years (year 2017). Figure 11 presents the results measured by the average RPS of all states (the accuracy rates, Kappa statistics and Kendall's coefficients between methods are all quite close but with the same trending as RPS, hence not displayed).

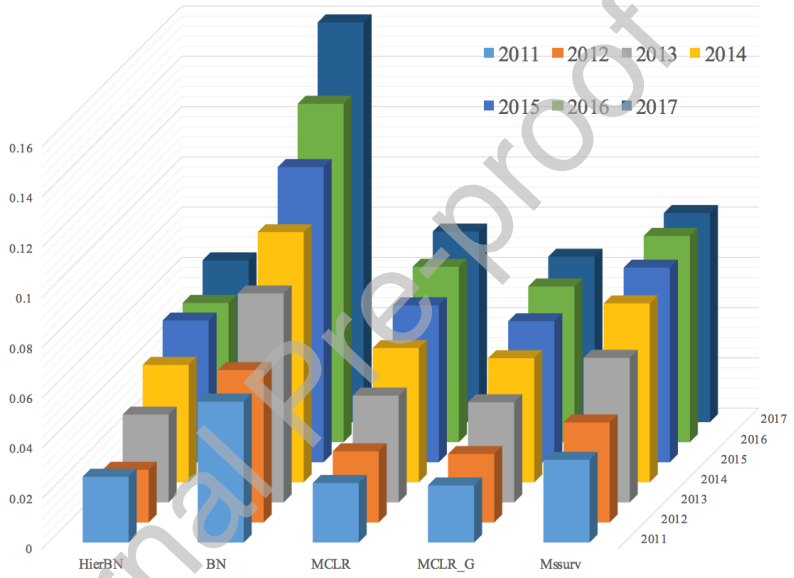


Figure 11: RPS of different approaches with the increase in prediction time.

Illustrated by this figure, all methods possess an increasing RPS over time; that is, the performance gets worse with the increase in prediction time. Our approach HierBN prevail other methods by having the lowest scores across these seven years. By comparing to the regular BN, we can also see the drastic improvement bring by separating the population into groups and learning between them. MCLR_G has a very close performance with the proposed approach. However, compares to MCLR, the benefit of grouping in MCLR_G is not significant in short future prediction but slowly increase over the years.

But we can achieve a reasonable prediction for a 2-year inspection with an RPS of 0.021, accuracy rate of 87.6%, Kappa score of 0.821 and Kendall's coefficient of 0.912, and a 4-year inspection with an RPS of 0.047, accuracy rate of 77.5%, Kappa score of 0.763 and Kendall's coefficient of 0.859 in HierBN. This counts for one and two biennially inspections in a regular inspection scheme respectively, which is valuable information for future inspection planning and resources scheduling.

5. Conclusion

To tackle the challenge of having uncertain and limited deterioration data for asset deterioration learning, hierarchical BN models for deterioration prediction, which leverages the learning from data, knowledge and similar assets, are developed and experimented in this paper. We summarise the use of data, knowledge and learning from others:

- **Censored data:** the exact time of an asset transits from one state to another state is not always available. We adopt the modelling of censored data in section 2.1 so that records from the periodic inspection can be used in place of exact transition times.
- **Expert knowledge:** the use of Bayesian parameter learning framework allows us to include knowledge from experts as the priors of a statistical distribution's parameters. The assignment of the priors is made possible by understanding the characteristics of each parameter. Section 3.4 explained how to elicit these priors in a Weibull distribution, we show how to translate engineering knowledge that is possible to be derived from experts into mathematical terms.
- **Learning from others:** some assets may only have little deterioration data. Therefore, we pool deterioration data with related asset types that are different but have related ageing processes as shown in section 3. By leveraging the assumption that assets with similar characteristics may deteriorate similarly, we separate assets into groups to provide individualised predictions, where, within each group, individuals are considered having the same deterioration behaviours. We show how to reduce the dimension of feature space and to select a few numbers of predictive features. To further tackle the challenge of limited data, especially after the grouping, where some groups may have more data while others have less, we extended the deterioration model into a hierarchical BN model to learn between groups. This hierarchical framework allows us to model multiple groups of assets in the same model and uses the hyper-parameters to leverage the strongly learned local parameters from groups with more data to infer the local parameters from groups with little data.

We also measure the performance of the developed deterioration prediction models using the deck data in Wyoming in NBI. We first evaluate how many features to consider in the models. The example shows that the performance improves at a rate up to three features; after this, more features can only improve the performance slightly. We also compare the performance of our models with other available approaches for correctly predicting the condition of a multi-state system. Our results show that our proposed models excel at most predictions, especially for cases where there is little data. We also show that as the prediction time increases, the accuracy of the prediction drops. The proposed models can provide reasonably good accuracy over 1 or 2-biennially inspections in a regular inspection scheme, which is useful for inspection planning.

Traditional asset management system, such as BMS, still heavily relies on Markov-based models for deterioration prediction [5]. These models suffer to provide accurate individual prediction as well as are computational challenging. Studies such as Chang [17] have

shown us how to group similar assets to provide individualised deterioration prediction using Markov models, but it suffers the limitation to learn the deterioration rate when the asset group has little data. Our hierarchical BNs addressed this limitation by incorporating expert knowledge and learning the deterioration between groups, and have achieved better performance in most metrics.

In an asset management system, we often have many assets and each asset comprises of multiple components rated by multiple states. It was pointed out that Markov models can easily become computationally intractable as the size of models grow exponentially with the increase of the number of states, components and assets [34]. BN model can tackle this challenge from its inference and its modelling. There are two common types of inference for BNs, Monte Carlo based inference and discretisation based inference. For Monte Carlo based inference, Le [9] showed that it is significantly faster than Markov models with the increase of model size, and for discretisation inference, we can appoint binary factorisation to reduce the computational complexity as introduced in Neil et al. [35]. In addition, we can extend the concept of Object Oriented in the BN modelling, which allows us to inference on specific compiled parts of the model that are queried, instead of inferencing the joint distribution of a whole model [36].

References

[1] D. M. Louit, R. Pascual, A. K. Jardine, A Practical Procedure for the Selection of Time-To-Failure Models Based on the Assessment of Trends in Maintenance Data, Reliability Engineering & System Safety 94 (2009) 1618–1628.

[2] G. Washer, M. Nasrollahi, C. Applebury, R. Connor, A. Ciolko, R. Kogler, P. Fish, D. Forsyth, Proposed Guideline for Reliability-Based Bridge Inspection Practices, Transportation Research Board, 2014.

[3] R. Kohavi, G. H. John, Wrappers for Feature Subset Selection, Artificial Intelligence 97 (1997) 273–324.

[4] H. Zhang, D. W. R. Marsh, Generic Bayesian Network Models for Making Maintenance Decisions from Available Data and Expert Knowledge, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 232 (2018) 505–523.

[5] G. Bu, J. Lee, H. Guan, M. Blumenstein, Y.-C. Loo, Improving Reliability of Markovian-Based Bridge Deterioration Model Using Artificial Neural Network, 35th International Symposium on Bridge and Structural Engineering (IASBE) (2011).

[6] D. M. Frangopol, M. Kallen, J. M. Van Noortwijk, Probabilistic Models for LifeCycle Performance of Deteriorating Structures: Review and Future Directions, Progress in Structural Engineering and Materials 6 (2004) 197–212.

[7] K. Kobayashi, K. Kaito, N. Lethanh, A Statistical Deterioration Forecasting Method Using Hidden Markov Model for Infrastructure Management, Transportation Research Part B: Methodological 46 (2012) 544–561.

[8] N. Ferguson, S. Datta, G. Brock, Mssurv, an R Package for Nonparametric Estimation of Multistate Models, Journal of Statistical Software 50 (2012) 1–24.

[9] B. Le, Modelling Railway Bridge Asset Management, Ph.D. thesis, University of Nottingham, 2014.

[10] M. P. Enright, D. M. Frangopol, Condition Prediction of Deteriorating Concrete Bridges, Journal of Structural Engineering 125 (1999) 1118–1125.

[11] F. Hong, J. A. Prozzi, Estimation of Pavement Performance Deterioration Using Bayesian Approach, Journal of Infrastructure Systems 12 (2006) 77–86.

[12] J.-C. Lu, Bayes Parameter Estimation for the Bivariate Weibull Model of Marshall-Olkin for Censored Data (Reliability Theory), IEEE Transactions on Reliability 41 (1992) 608–615.

[13] F. Coolen, On Bayesian Reliability Analysis with Informative Priors and Censoring, Reliability Engineering & System Safety 53 (1996) 91–98.

[14] D. Han, K. Kaito, K. Kobayashi, Application of Bayesian Estimation Method with Markov Hazard Model to Improve Deterioration Forecasts for Infrastructure Asset Management, KSCE Journal of Civil Engineering 18 (2014) 2107–2119.

[15] L. Scholten, A. Scheidegger, P. Reichert, M. Maurer, Combining Expert Knowledge and Local Data for Improved Service Life Modeling of Water Supply Networks, Environmental Modelling & Software 42 (2013) 1–16.

[16] D. Marquez, M. Neil, N. Fenton, A New Bayesian Network Approach to Reliability Modelling, Mathematical Methods in Reliability (MMR07) (2007).

[17] M. Chang, Investigating and Improving Bridge Management System Methodologies Under Uncertainty, Ph.D. thesis, Utah State University, 2016.

[18] P. Wei, Z. Lu, J. Song, Variable Importance Analysis: A Comprehensive Review, Reliability Engineering & System Safety 142 (2015) 399–432.

[19] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32.

[20] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution, BMC Bioinformatics 8 (2007) 25.

[21] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional Variable Importance for Random Forests, BMC Bioinformatics 9 (2008) 307.

[22] A. R. Andrade, P. F. Teixeira, Statistical Modelling of Railway Track Geometry Degradation Using Hierarchical Bayesian Models, Reliability Engineering & System Safety 142 (2015) 169–183.

[23] M. Neil, M. Tailor, D. Marquez, N. Fenton, P. Hearty, Modelling Dependable Systems Using Hybrid Bayesian Networks, Reliability Engineering & System Safety 93 (2008) 933–939.

[24] Z. U. Din, P. Tang, Automatic Logical Inconsistency Detection in the National Bridge Inventory, Procedia Engineering 145 (2016) 729–737.

[25] M. B. Kursa, W. R. Rudnicki, Feature Selection with the Boruta Package, Journal of Statistical Software 36 (2010) 1–13.

[26] W. Weseman, Recording and Coding Guide for the Structure Inventory and Appraisal of the Nation's Bridges, 1995. Report to United States Department of Transportation, Federal Highway Administration, USA.

[27] N. E. Fenton, M. Neil, J. G. Caballero, Using Ranked Nodes to Model Qualitative Judgments in Bayesian Networks, IEEE Transactions on Knowledge and Data Engineering 19 (2007) 1420–1432.

[28] R. A. Groeneveld, Skewness for the Weibull Family, Statistica Neerlandica 40 (1986) 135–140.

[29] J. McCool, Using the Weibull Distribution: Reliability, Modeling, and Inference, John Wiley& Sons, 2012.

[30] M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables, Dover Publications, 1964.

[31] B. M. Phares, D. D. Rolander, B. A. Graybeal, G. A. Washer, Reliability of Visual Bridge Inspection, Public Roads 64 (2001).

[32] D. S. Wilks, Statistical Methods in the Atmospheric Sciences, Academic Press, 2011.

[33] J. R. Landis, G. G. Koch, The Measurement of Observer Agreement for Categorical Data, Biometrics (1977) 159–174.

[34] W. T. Scherer, D. M. Glagola, Markovian Models for Bridge Maintenance Management, Journal of Transportation Engineering 120 (1994) 37–51.

[35] M. Neil, X. Chen, N. Fenton, Optimizing the Calculation of Conditional Probability Tables in Hybrid Bayesian Networks Using Binary Factorization, IEEE Transactions on Knowledge and Data Engineering 24 (2011) 1306–1312.

[36] A. Pfeffer, D. Koller, B. Milch, K. T. Takusagawa, SPOOK: A System for Probabilistic Object-Oriented Knowledge Representation, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (1999) 541–550.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: