DATA MINING OF MARINE ACCIDENT/INCIDENT DATABASE FOR USE IN RISK-BASED SHIP DESIGN

Dracos Vassalos, d.vassalos@na-me.ac.uk
Wenkui Cai, wenkui.cai@strath.ac.uk
The Ship Stability Research Centre, Department of Naval Architecture and Marine Engineering
Universities of Glasgow and Strathclyde, United Kingdom

Dimitris Konovessis, d.konovessis@newcastle.ac.uk
School of Marine Science and Technology, Newcastle University Marine International (NUMI),
Singapore Singapore Newcastle University

ABSTRACT:

Engineers are always confronted with the challenge of deriving factual and reliable information from accident/incident data and updating selected risk models for use in design in conjunction with results obtained from first-principle approaches. Remarkable achievements in computer science provide a promising forward, through the use of data mining techniques. The work presented in this paper aims to employ Bayesian Networks (BN) in developing risk models and its pertinent "mining" algorithms so that data is explored exhaustively. As a result, use of subjective judgment is reduced (or even eliminated) and the potential of objective resource/means could be fully released. Ultimately, the output is expected to provide a concrete foundation to assist decision-making.

Keywords: *Risk-Based Ship Design, Data Mining, Bayesian Network*

1. INTRODUCTION

Rapid evolution of computer technologies and corresponding propagated computational algorithms enable engineers to step out the slough of rough estimation through a qualitative approach and to tackle it using more scientific quantitative strategies – in the case of safety, this is translated to the adoption of consistent measures and frameworks for addressing safety through quantitative risk analysis. In the marine industry, the concept of risk-based ship design has been advocated for more than one decade, in promoting systematic integration of risk assessment with the conventional ship design process and allowing flexible trade-offs between cost, earning, performance and safety. Extensive collaboration within the marine industry has been achieved through a number of significant large-scale European research projects, namely, SAFER EURORO I & II, SAFEDOR. Risk-based design under the theme "Design for Safety" has demonstrated its brawny vitality through consecutive introduction of guidelines for damage stability and fire safety (MSC82/24/Add.1 Annex 2, 2007, MSC/Circ.1002, 2001, MSC.1/Circ.1212, 2006). Furthermore, shipyards also show great interest to the early signal of liberation, which is particularly true for yards delivering high-end products, cruise ships and RoPax. One of the pioneering projects MV Genesis, the world biggest cruise ship, has adopted risk-based methodologies in her design and is due to be delivered at the end of year 2009. Early results through "risk screening" indicated satisfactory performance in terms of safety (Vassalos, 2009).

To employ risk assessment, engineers are always confronted with the challenge of deriving factual and reliable information from accident/incident data and updating selected risk models in conjunction with results obtained from first-principle approaches. This is mainly practiced through subjective judgement (expert opinion and interpretation). As we are stepping towards a more rigorous stage of development of risk-based design, where usage of subjective information could be kept to a minimum, uncertainty should be quantified and minimised through more consistent and objective means. Rapid advancement of science and computer technologies could equip us with gigantic storage space and fast process capability, which makes it possible to develop and employ sophisticated methodologies that one could not perform in the past. Documentation of past accident/incident becomes a common practice within all sectors of the marine industry with the view that critical lessons can be learned. However, current approaches towards accident/incident database can be at best described as simple statistical analysis, while it lacks a systematic and reliable means to integrate data with risk-based design in an efficient and consistent manner.

Accompanied with the development of computer technologies ever growing knowledge in computer science points towards a promising approach to tackle this problem, in employing data mining techniques to constantly learn from data and feed selected risk models. Selecting Bayesian Networks (BN) as the platform for the development of risk models is justified due to their superiority over traditional tools such as fault and event trees, with respect to the sophisticated causality modelling offered, probabilistic data processing and updating, flexibility and transparency. Hence not only data collected from actual accident/incident but also data generated from simulations as well as established parametric models can be united and effectively utilised due to the wide adaptability of BN.

The paper starts with a brief introduction of data mining and BN in Section 2 which is also accompanied by a limited survey of relevant work in the engineering domain. Section 3 summarises the approach adopted. Detailed approach and a preliminary case study are depicted in Sections 4 and 5, respectively. Finally conclusions are drawn in Section 6.

2. BACKGROUND KNOWLEDGE

2.1 Data Mining & Bayesian Networks

Data mining, also popularly referred to as "knowledge discovery from data" (KDD), can hardly be given a unified definition as its territory is expanding at a fast pace. It makes use of works at multidisciplinary fields, such as database technology, machine learning, pattern recognition, statistics, data visualization, etc., so as to discover meaningful new correlations, patterns, and trends. The typical process is demonstrated in the flow chart of Figure 1. Various tasks can be carried out by data mining techniques including association rules, quantifying the relationship between two or more attributes, classification rules, predicting categorical labels, and lastly clustering, grouping a set of physical objects into similar classes (Han & Kamber, 2006). Equipped with these functionalities data mining techniques have been employed across a wide spectrum of fields and applications: retailing (Han et al., 2005; Zheng et al., 2001) concerning customer shopping habit, financial system concerning fraud detection (Bishop, 2006; Chan et al., 1999; Guo et al., 2008), several engineering fields such as the chemical (Anand et al., 2006) and aviation industry (Nazeri et al., 2001) concerning safety data analysis. In contrast, limited literature and applications are observed in the marine industry in the area of data mining.

One of the reasons is the fact that a platform capable of coupling both "mining" techniques and probabilistic inference algorithms is still unclear for the time being. Nevertheless,

an approach of integrating data mining method with uncertainty reasoning technique (Chen, 2001), covering Bayesian Networks, Artificial Neural Networks, Fuzzy Logic and Genetic Algorithms, provides a promising way forward. More specifically, BN is identified as the platform to be used for representing underlying characteristics of the data, namely, causal structure and uncertainties, and also performing probabilistic inference in the case of model updating. Furthermore, BN is also generally considered to be one of the most promising tools for use in marine risk analysis, due to its superiority in modelling sophisticated relationships among physical variables, in comparison with fault and event tree modelling.

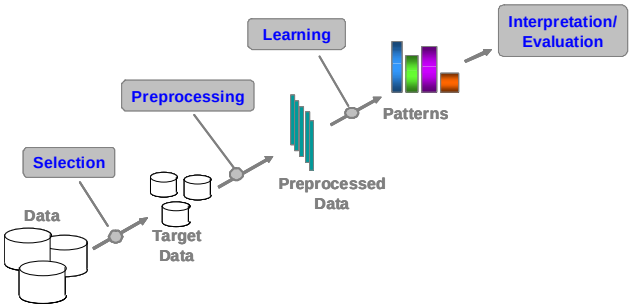

Figure 1. Flow chart of data mining, (Han & Kamber, 2006).

A BN is a graphical structure that consists of a set of random variables, represented as nodes, and a set of directed edges between variables, known as arcs. In the case of discrete variables, a conditional probability table (CPT) is attached to each variable having parents. The only constraint that BN has to comply with is that the nodes together with arcs have to form "directed acyclic graphs": one cannot return to a node by following the arcs. Another attractive feature of BN is the robust capability of carrying out probabilistic inference tasks on the basis of the Bayesian Theorem, by Equation (1)

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \qquad (1)$$

With the backup of sound mathematical theory, one can constantly employ BN to represent and reason about an uncertain domain in a flexible manner. This can be illustrated as shown in Figure 2, where four types of reasoning are depicted. The diagnostic type, as the name suggests, is based on reasoning from symptoms to causes in a way that evidence is given to the end node and information propagates backwards, whereas the predictive type is to set evidence for root node and to infer forward. Analogical techniques are also applied to the two other types of reasoning.

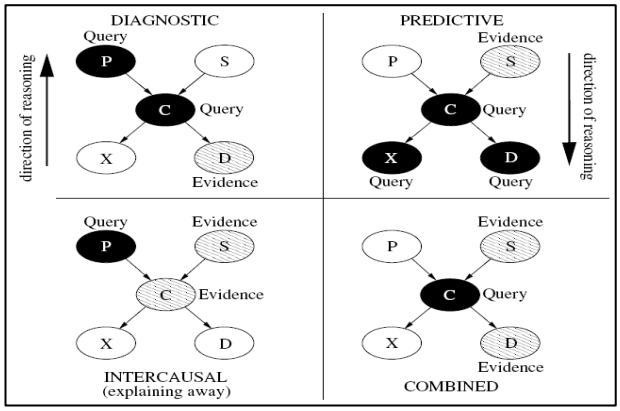

Figure 2. Types of reasoning for BN, (Korb et al., 2004).

2.2 Limited Survey of Similar Approach

Research conducted at the U.S. Federal Aviation Administration made use of an early data mining research product from Microsoft Corporation, known as WinMine (Milburn et al., 2006). In this piece of work, aviation accident data were classified using the Human Factors Analysis and Classification System (HFACS) to investigate interrelationships between various causal factors for a single type of accident (controlled flight into terrain). From this preliminary examination it was revealed that WinMine is a useful tool to illustrate interaction among causal factors in a way that traditional approach cannot. A particular feature is that the network is updated dynamically with the changing level of strength of dependencies. However, users of the tool initially have no information of how strong the strength of dependency is; they can simply adjust the slider bar (indicated by red arrow on the left hand side in Figure 3) to a particular location. Thus

this system would be more appropriate for ranking the relevant importance of causal factors. More importantly, the network generated is not readily available for further processing due to that is not quantified.

An integrated approach towards modelling of complex interaction of causal factors for aviation accident was presented in (Luxhøj et al., 2003). Again BN was placed at the centre of the proposed framework. The approach initiates with the identification of case-based scenarios. Influence diagrams are then constructed to model the interactions among causal factors which are followed by finalising the BN model including parameters quantification. Eventually the projected risk is displayed on a relative risk intensity graph accompanied with a series of risk intervention modules for lowering the risk. As a result, effectiveness of various risk control options can be assessed with respect to specific variable so that the objective of assisting decision-making can be achieved. Though an obvious shortcoming of the proposed approach relates to the significant involvement of personnel opinion which would greatly compromise the reliability of the output, this approach does set a visible target in terms of applying BN as a risk model to assist decision making.

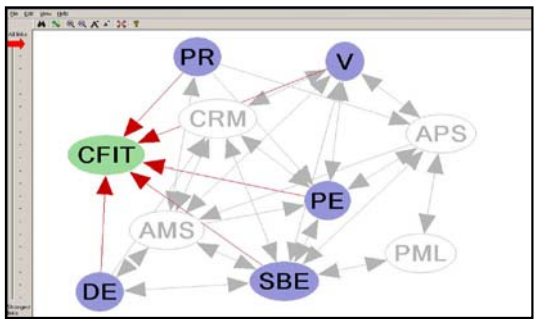

Figure 3. Low strength of dependency from WinMine, (Milburn et al., 2006).

Besides the research of applying BN and its relevant mining approaches mentioned above, there were also a very limited number of applications concerning employment of BN as a tool to assist decision-making in the marine industry. An extensive examination of BN was carried out in Friis-Hansen's PhD thesis (2000), which tried to incorporate it into risk analysis at different stages and fields of marine operation, while DNV (2003) has also used BN for risk screening of collision and grounding events for large passenger ships. More recently, research carried out at SAFEDOR (Ravn, 2006a, Ravn, 2006b) attempted to model causal factors for ships under power and to capture the failure of propulsion and steering gear system. It has demonstrated that BN has huge potential to play a much important role in the risk assessment process; nevertheless, a key bottleneck needing to be tackled urgently is how historical data and first-principle data can be utilized for generating BN risk model so that reliable and optimal results can be obtained without difficulty.

3. APPROACH PROPOSED

A BN risk model to be developed and employed for a marine application is generally characterised by a number of preselected variables with links connecting between variables to reflect belief of the causal relationships. The current approach towards constructing a BN model is often discredited because of unauthentic sources, which significantly weaken its applicability. As a result this research focuses on how a BN model can be derived without using subjective means.

An initial task was to set up an accident/incident database within the Ship Stability Research Centre (SSRC) with first-hand accident/incident data provided by collaborating partner Royal Caribbean Cruise Lines (RCCL). The database is designed to store and retrieve data confidentially while the main focus is to flexibly extract selected data for designated query and pre-process it to feed the BN learning algorithm.

On the foundation of the accident/incident database system, the BN structure can be derived out of the data using specific learning techniques. There are two major types of learn-

ing: constraint-based learning and score-based learning.

The BN network obtained in this way should then be quantified with probabilities or conditional probabilities through data using parameters learning algorithms. However, data employed for parameters learning should not only cover factual data collected from the industry but also data generated through simulation from first-principle tools. This is due to the fact that factual data may not be suitable for a brand new and innovative design where no historical experience exists.

4. INTEGRATION OF DATA WITH BN

4.1 Accident/Incident Database System

The database was developed using Microsoft Access in the form of a relational database. Key data concerning vessel information, voyage information, human factors, fire, collision, grounding, machinery failure, root causes, and consequences have been programmed in separate relational tables. Raw accident/incident data provided by RCCL was carefully interpreted and analysed so that each event can be encoded and stored in the database. A number of 576 cases of fire accident/incident for cruise ships are stored and are readily available for further processing, while data for three other types of accident, namely, collision, grounding, and machinery failure, are currently under processing.

4.2 Structure Learning

Structure learning of BN can be broadly classified as constrained-based learning and score-based learning. Constrained-based learning aims to determine independencies and conditional independencies (CI) between variables using statistical measures. Traditional approach is to make null hypothesis testing of dependencies between two variables so as to identify the significance of association which will be checked against the predefined confidence level. This is feasible to identify association between two variables, while more advanced mathematical models are needed to identify CI among three variables or more. In contrast, the score-based learning approach assesses a number of potential BN patterns in terms of probability of a certain "optimal" BN pattern given the data provided. An important module needs to be developed beforehand in order to explore potential BN patterns.

4.2.1 Constrained-Based Learning

Due to the particular characteristics of accident/incident data and the format of data recorded in the database, categorical data analysis would be the best technique for CI discovery. Two mathematical models were adopted for dependency analysis: logistic regression model, also known as Logit, and log-linear model.

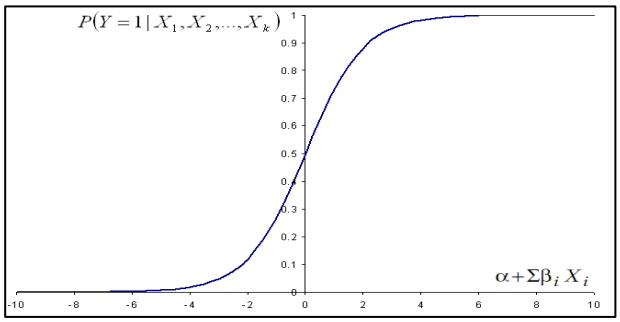

Figure 4. Logistic regression functions.

Logistic regression model is an important model for data with categorical response variable. One of the primary reasons that this model won great favour is because the function will always produce a value between zero and one, which is ideal for modelling probabilistic functions. Furthermore, the S-shaped curve that is similar with cumulative distribution function of the standard normal distribution is another attractive feature, depicted in Figure 4. For binary response variable, Y = (0, 1) with k independent variable values X_1, X_2, up to X_k, the logistic model is defined as,

$$P(Y=1|X_1,X_2,...,X_k)=\frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}} \quad (2)$$

The terms α, β_i represent the unknown coefficient parameters to be determined using data. In order to assess the goodness of fit of the estimated model, likelihood-ratio chi-square test or Pearson chi-square test are used to compare the observed counts and fitted values. The maximum likelihood technique is generally adopted for model fitting with estimated coefficients to be used for examination of significance of association.

On the other hand, the loglinear model is a model commonly used for contingency table analysis by modelling cell counts and eventually deriving the association and interaction patterns among variables. A saturated loglinear model for three variables (X, Y, Z) can be defined as,

$$\begin{aligned}\log\ \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\ &+ \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}\end{aligned} \quad (3)$$

The doubly subscripted terms, such as $\{\lambda_{ij}^{XY}\}$, are association terms that reflect deviations from independence. The likelihood-ratio statistic (deviance) or Pearson statistic is the tool for assessing goodness of fit. Through backward elimination of saturated model one can identify the most simplified model, whilst still satisfying predefined significance level. Consequently CI can be easily derived.

With a collection of independencies and CI, proper learning algorithms could then be processed. One of the most widely accepted approach, known as PC algorithm (Spirtes, 2000), was selected for BN structure derivation. The PC algorithm is briefly introduced here:

- Start with a complete undirected graph in which each variable is linked with all other variables with undirected arcs.
- Iterate throughout the graph to remove the link (say, X — Y) from the graph if there is $I(X,Y|S)$, where S denotes any node of the set of adjacent nodes of X and Y. $I(X,Y|S)$ simply indicates X and Y are conditionally independent given S.
- Iterate throughout the network with each uncoupled meeting (X — Y — Z) and orient as (X → Y ← Z) if X and Z are found to be independent given a set of variables which do not contain Y. For the rest of links, the arrow should be directed in a way that no more head-to-head links is created.

4.2.2 Score-Based Learning

The score-based learning algorithm uses scoring functions to compare various BN structures and select the best fit of the data. While the score function is generally known as Bayesian scoring criterion,

$$score = P(d|G)$$

$$=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij})}{\Gamma(N_{ij}+M_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(a_{ijk}+s_{ijk})}{\Gamma(a_{ijk})} \quad (4)$$

Where the variables N, M, a, s are originated from the Dirichlet density function, given in (5).

$$\rho(f_1,f_2,...,f_{r-1})=\frac{\Gamma(N)}{\prod_{k=1}^{r}\Gamma(a_k)}f_1^{a_1-1}f_2^{a_2-1}...f_r^{a_r-1} \quad (5)$$

Where: $N=\sum_{k=1}^{r}a_k$, $0\le f_k\le 1$, $\sum_{k=1}^{r}f_k=1$

To obtain the optimal BN model, a heuristic searching algorithm is needed to generate all promising network patterns for evaluation. Care should be taken when performing the searching process due to that a locally maximized solution can be easily confused with the globally optimal one. An algorithm called "Greedy Search" was proven to be an effective approach if the data set is sufficiently large (Korb et al., 2004). The detail of this algorithm is not further discussed in this paper since it is currently under development.

Both constrained-based and score-based strategies have their superiorities and limitations. The score-based approach may be inac-

curate when CI are not obvious, while the score-based approach may not guarantee a global optimal structure. As a result, it is suggested to generate a preliminary solution first through constrained-based approach before performing optimisation using score-based technique (Neapolitan, 2004).

4.2.3 Parameters Learning

Strictly speaking, the Bayesian scoring criterion mentioned above in score-based learning approach is a conditional probability, which can be identified only after the network is quantified. Since parameter learning aims to derive probabilities and conditional probabilities out of data, one has to be clear that probabilities are relative frequencies. Using Dirichlet distribution to represent our belief concerning a relative frequency is a common practice. Initial condition for a Dirichlet distribution with parameters $a_1, a_2, ...a_r$ is,

$$\rho(f_1, f_2, ..., f_{r-1}) = Dir(f_1, f_2, ... f_{r-1}; a_1, a_2, ...a_r) \quad (6)$$

The distribution function can be updated by additively taking into account new evidence,

$$\rho(f_1, f_2, ..., f_{r-1} \mid d) = Dir(f_1, f_2, ... f_{r-1}; a_1 + s_1, a_2 + s_2, ...a_r + s_r) \quad (7)$$

By doing so, the updated probability for a specific variable given the evidence of data is:

$$\rho(X^{(M+1)} \mid d) = \frac{a_k + s_k}{N + M} \quad (8)$$

Where: $N = \sum_{k=1}^{r} a_k$, $M = \sum_{k=1}^{r} s_k$

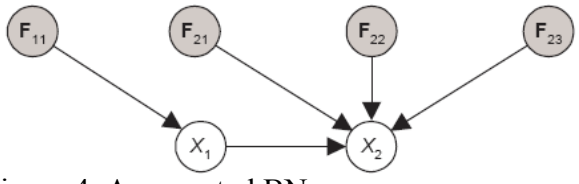

Figure 4. Augmented BN.

Concerning the characterisation of the distribution function, augmented BN was defined so that each variable in the network is given node(s) to incorporate our belief about relative frequency. For instance, X_1 was given one node (with three states) to define its beliefs while X_2 is given three augmented nodes with each node corresponding to a state of its parent node X_1 as shown in Figure 4. Consequently every augmented node is quantified and updated.

5. CASE STUDY

Five variables were extracted from the database to examine the feasibility of adopting the BN learning approach. The five variables selected are generally considered to be important factors affecting the risk level of fire accident for cruise ships. The number of variables was predefined at five as the computation complexity increases exponentially where automation may be needed. A total of 576 cases were retrieved for the five variables:

- T denotes event time (the time is commonly agreed to be an important factor affecting the percentage of fatalities onboard ship for day time and night time if fire escalated from the space of origin with smoke propagated. Event time is classified into daytime and night time)
- L denotes vessel location (the location of vessel when accident occurs could affect the consequences also due to people onboard would have much chance to survive if vessel is near the harbour. It has two categories: at sea, in port)
- C denotes whether crew is present (as crews are generally considered to be trained people onboard who are familiar with what proper actions should be taken at the first instance, so the chance of putting out a fire is much larger if any crew is at the scene. It is treated as binary variable denoted as true, false)
- H denotes human factor (the human factor still accounts for 80% of the causes of marine incidents. It presented as true, false)
- S denotes severity (event severity is classified as minor, moderate and serious).

An approach of integrating both constraint-based approach and score-based approach is adopted for the network structure. In constrained-based learning statistical analysis is carried out to identify dependencies and CI. Following that, the network obtained is quantified, first by using parameter learning method as score-based approach needs to make use of the quantified augmented nodes. Lastly, in the score-based process an initial testing was given to various randomly chosen actions instead of carrying out extensive exploration of the candidate BN .

Table 1. Independence analysis between two variables.

Independence		DF	Value	Probability
T,L	Pearson Chi-Square	1	6.4016	0.011
	Likelihood ratio Chi-Square	1	6.666	0.010
T,H	Pearson Chi-Square	1	0.9482	0.330
	Likelihood ratio Chi-Square	1	0.9445	0.331
T,S	Pearson Chi-Square	2	4.6203	0.099
	Likelihood ratio Chi-Square	2	4.1503	0.126
T,C	Pearson Chi-Square	1	0.1828	0.669
	Likelihood ratio Chi-Square	1	0.1836	0.668
L,H	Pearson Chi-Square	1	0.1531	0.696
	Likelihood ratio Chi-Square	1	0.1529	0.696
L,S	Pearson Chi-Square	2	0.8505	0.654
	Likelihood ratio Chi-Square	2	0.8922	0.640
L,C	Pearson Chi-Square	1	0.0235	0.878
	Likelihood ratio Chi-Square	1	0.0235	0.878
H,S	Pearson Chi-Square	2	5.5868	0.061
	Likelihood ratio Chi-Square	2	5.5231	0.063
H,C	Pearson Chi-Square	1	0.004	0.950
	Likelihood ratio Chi-Square	1	0.004	0.950
S,C	Pearson Chi-Square	2	4.5819	0.101
	Likelihood ratio Chi-Square	2	4.3091	0.116

According to the PC algorithm, the graph starts with each variable linked with all other variables. Results for two variable dependencies analysis indicate six pairs of independencies as highlighted in red in Table 1. The analysis was performed on the basis of a logistic regression model using SAS, an integrated software system for data analysis. As a result, the corresponding links were removed from the graph, which is shown in Figure 6. It is noted that the significance of dependencies is not so strong between event time and event severity (90% of confidence) comparing with the level that is generally adopted (95% of confidence), however, such confidence level is still reasonable and the two relevant variables (event time and vessel location) could be isolated from the rest three variables if this link is removed.

The refined graph needs to be further examined against CI among three and more variables. For three variable CI analysis, the loglinear model was employed to assess the possible combinations of any three variables on the basis of refined graph. Four possible combinations (L, T, S) (T, S, C) (T, S, H), (C, S, H) were checked using the backward elimination approach. Consequently three CI were identified, $I(L,S\,|\,T)$, $I(T,H\,|\,S)$, $I(T,C\,|\,S)$. Any direct link between two variables that are conditionally independent on a third variable should be removed, however no action was needed due to the special case that the existing graph satisfies all constraints. Judging by the limited links within the obtained graph and small amount of variables it is understood that further operation is not necessary, nevertheless higher number of variables combination were analyzed to further confirm and validate the previous result. Five variables loglinear models were assessed step by step from saturated model to the simplest one (Figure 5), suggestingthe same graph as the one originally obtained.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Time	1	10.26	0.0014
Location	1	75.43	<.0001
Time*Location	1	6.07	0.0137
Severity	2	281.61	<.0001
Time*Severity	2	4.49	0.1059
Human	1	0.63	0.4278
Human*Severity	2	5.02	0.0814
Crew	1	0.37	0.5423
Crew*Severity	2	3.97	0.1376
Likelihood Ratio	34	23.27	0.9173

Figure 5. SAS output for Loglinear model with five variables.

Link orientations for the resultant graph were then determined using the PC algorithm. Previous CI analysis of three variable combinations suggest little evidence from the combination (S, H, C), thus head-to-head orientation was added first as shown in Figure 6. Following which, arrow was added from S pointing T as the algorithm allows none new head-to-head situation to be generated;a similar principle ap-

plies to the link from T to L. Consequently, an initial BN structure was learned.

Parameter learning was then performed with the assistance of the augmented BN. It is important each variable to have equivalent sample size with the rest so as to suit the exchangeability of dataset. The initial network before accounting for data and the updated network are displayed in Figures 6 and 7, respectively. The probability of getting the identified network as the optimal model given the set of factual data can be easily estimated through Bayesian scoring criterion. Since the searching algorithm of the score-based approach is still under development, random modifications were made to demonstrate the feasibility of using score functions for judging various alternatives. Removal of orientation from S to T and from H to S has been tested; the initial model gave a probability of 0.296, the model without link from S to T gave 0.189, and lastly the model without link from H to S had 0.515. It appears that the last network model is the one to be chosen within the limited space of candidate networks, however one should always bear in mind that the space in this case does not cover promising candidate networks exhaustively.

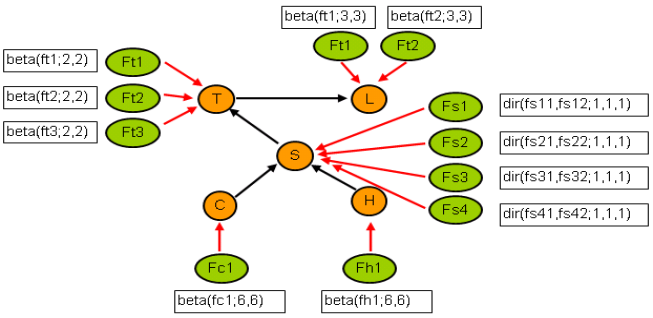

Figure 6. Initial augmented BN.

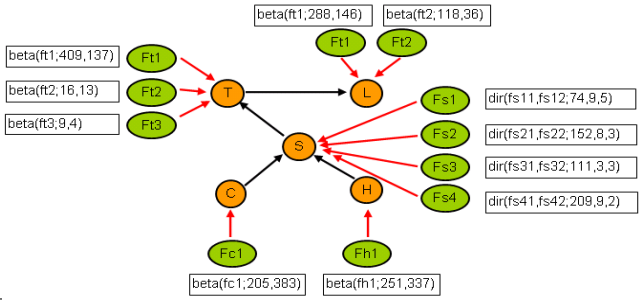

Figure 7. Updated augmented BN.

6. CONCLUSION

This paper presents a systematic approach of constructing risk models using Bayesian Networks by learning from data using data mining techniques. The data can be historical accident/incident or data generated through simulations from first-principle tools. An accident/incident database system has been established for information storage and retrieval. Two BN structure learning approaches together with a parameter learning technique were examined on the basis of five preselected variables from the data, where the preliminary results suggest both constraint-based and score-based approaches can provide a promising way of constructing a BN structure with much improved reliability when compared to the current subjective approach of using expert judgement.

Further work on determining dominant variables is needed for each specific accident category so that fast and reliable computation can be carried out to facilitate future areas of research such as assistance of decision making, etc. Moreover, presented learning techniques need to be further detailed to tackle missing data and to quantify uncertainties. An algorithm for fast identification of promising BN patterns is also an important module for score-based learning.

7. ACKNOWLEDGEMENTS

The financial support and data input by Royal Caribbean Cruise Lines is gratefully acknowledged. The authors would also like to thank researchers at SSRC specifically Dr. Jasionowski Andrzej, Mr. Mermiris George, Dr. Puisa Romanas for their invaluable help to this research.

8. REFERENCES

ANAND, S., KEREN, N., TRETTER, M. J., WANG, Y., O'CONNOR, T. M. & MANNAN, M. S., 2006, “Harnessing data mining to explore incident databases”, Journal of Hazardous Materials, 130, 33-41.

BISHOP, C. M., 2006, Pattern recognition and machine learning, New York, Springer.

CHAN, P. K., FAN, W., PRODROMIDIS, A. L. & STOLFO, S. J., 1999, Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and Their Applications, 14, 67-74.

CHEN, Z., 2001, Data mining and uncertain reasoning: an integrated approach, New York ; Chichester, Wiley.

DNV, 2003, FSA Main Technical Report. 2003-0277

FRIIS-HANSEN, A., 2000, Bayesian Networks as a Decision Support Tool in Marine Applications, Dept. of Naval Architecture and Offshore Eng. Technical University of Denmark.

GUO, T. & LI, G.-Y., 2008, Neural data mining for credit card fraud detection. Kunming, China, Inst. of Elec. and Elec. Eng. Computer Society.

HAN, E.-H., KARYPIS, G. & KUMAR, V., 2000, Scalable parallel data mining for association rules. IEEE Transactions on Knowledge & Data Engineering, 12, 337-352.

HAN, J. & KAMBER, M., 2006, Data mining: concepts and techniques, Amsterdam ; London, Elsevier.

IMO, 2001, Guidelines on Alternative Design and Arrangements For Fire Safety. IMO MSC/Circ.1002.

IMO, 2006, Guidelines on Alternative Design and Arrangements For SOLAS Chapters II-1 and III. IMO MSC.1/Circ.1212.

IMO, 2007, Resolution MSC.216(82) - Adoption of Admendments to the International Convention for the Safety of Life At Sea, 1974, as Amended. IMO MSC 82/24/Add.1 Annex 2.

KORB, K. B. & NICHOLSON, A. E., 2004, Bayesian artificial intelligence, Boca Raton; London, Chapman & Hall/CRC.

LUXHØJ, J. T., JALIL., M. & JONES., S. M., 2003, A Risk-Based Decision Support Tool for Evaluating Aviation Technology Integration in the National Airspace System. Proceedings of the AIAA's 3rd Annual Aviation Technology, Integration, and Operations (ATIO) Technical Forum. Denver, Colorado.

N. J. MILBURN, L. DOBBINS, J. POUNDS, and S. GOLDMAN, 2006, "Mining for information in accident data," Federal Aviation Administration, Washington, DC.

NAZERI, Z., BLOEDORN, E. & OSTWALD, P., 2001, Experiences in mining aviation Safety data. Santa Barbara, CA, United States, Association for Computing Machinery.

NEAPOLITAN, R. E., 2004, Learning Bayesian networks, Harlow, Prentice Hall.

NORRINGTON, L., QUIGLEY, J., RUSSELL, A. & VAN DER MEER, R., 2008, Modelling the reliability of search and rescue operations with Bayesian Belief Networks. Reliability Engineering and System Safety, 93, 940-9.

RAVN, E., 2006a, SAFEDOR D 2.4.3 Modelling of Causation Factor for Ship Under Power.

RAVN, E., 2006b, SAFEDOR D 2.4.7 Risk-Based Model for Failure of Propulsion and Steering Gear System.

SPIRTES, P., GLYMOUR, C. N. & SCHEINES, R., 2000, Causation, prediction, and search, Cambridge, MA, MIT Press.

VASSALOS, D., 2009, Project Genesis Risk-Based Design Implementation. SAFEDOR Closing Meeting. GL Hamburg, Germany.

ZHENG, Z., KOHAVI, R. & MASON, L., 2001, Real world performance of association rule algorithms. San Francisco, CA, United States, Association for Computing Machinery.