

Rough Set Approach for Identification of Accident on Water Route Segment

Hao ZHANG, Ying-jie XIAO and Liang CHEN

Shanghai Maritime University, Shanghai 201305, China
haozhang@shmtu.edu.cn

Abstract

This paper presents a novel non-parametric methodology – rough set theory – for accident occurrence exploration. The rough set theory allows researchers to analyze accidents in multiple dimensions and to model accident occurrence as factor chains. Factor chains are composed of Seaman characteristic, ship's characteristics, navigational behavior and environment factors that imply typical accident occurrence. Rose2 software tool is used. The purpose of this application is to find out the critical attributes to reduce the number of the fatality in maritime accidents. This paper explains the application on the accident reports of Accident Data database, containing data records for all categories of maritime accidents between the years of 2003 and 2009. Variable precision rough set is used to reduce the attributes of data set. The categorization tools and decision trees are used to find the relations and rules about the accidents resulted in fatality. Some rules about the fatality are obtained and also the attributes that affect the fatality in the incident have determined.

Keywords: *Rough set, maritime accident, route segment, machine learning*

1. Introduction

With the rapid growth in maritime transportation in China, number of ships, size of ships, traffic conflicts and delays, and accidents has also dramatically increased in recent years. Marine traffic risk has been a core subject in maritime studies coupled with transport safety, shipping efficiency, distribution accessibility, reliability and loss prevention. Regulatory changes in the maritime industry are distilled from past experience, mainly related to ship accidents. Safety objectives and functional requirements would be more useful, requiring safety performances to be met both for technical and operational aspects. This can be achieved through the consideration of accident analysis. In recent years, China Maritime Safety Administration (MSA) has identified the importance for the implementation of advanced statistical methods in maritime accident data mining. Identifying the premonitory factors for maritime traffic accidents is an important task for the China MSA.

Scientific literature on risk and accident analysis shows a recent diffusion of the use of mathematical methods. Several techniques have been applied in maritime accident analysis, including:

BBN modeling of HOFs can be used in risk analysis to identify further opportunities of risk mitigation acting at the organizational and regulatory level of the MTS, the model could be used as a tool for supporting retrospective analyses based on incident reporting or accident investigation, such as the identification of latent failures at the organisational level (P. Trucco, 2007).

The fuzzy and bayes probability method are an efficient approach for maritime risk assessment. The technique allows solving problems related to dealing the imprecise and uncertain data (Hu S. P, 2007; P. Trucco, 2007; Liu, *et al.*, 2005, Eleye-Datubo, *et al.*, 2008; Yang, *et al.*, 2008; Jean-François Balmat 2011). Maritime traffic poses various

risks in terms of human, environmental and economic loss, besides (Uluscu, OS, 2009), Passenger ship and High Speed Craft (HSC) ship are high-risk vessels. Scenarios for use in advanced evacuation analyses of passenger ships scenarios based on risk assessment can be related to actual accident scenarios, covering the major hazards passenger ships (Erik Vanem, Rolf Skjong, 2006; P. Szwed, J. Rene van Dorp, J.R.W. Merrick). Maritime queuing and traffic simulation model are significant methods to risk assessment (Merrick JRW, Drop JR, 2003; Dimitrios Mavrakis, Nikolaos Kontinakis, 2008; Ozbas, B,2009; Floris Goerlandt,2011). Marine traffic management currently puts most of its emphasis on controlling traffic. Statistical models, such as linear regression, logistic regression, Poisson, negative binomial or zero-inflated count models, have been widely applied in analyzing causes of accidents. The statistically significant evidence reveals that the port of registration, the vessel type and the accident type are critical to the number of injuries and fatalities. The findings have identified factors that can contribute to reducing the severity level of port accidents (2007).

Formal Safety Assessment (FSA) methodology has been successful in nuclear and offshore industries and IMO is following this. FSA has been developed by the International Maritime Organization (IMO) as a structured and systematic methodology, aimed at enhancing maritime safety, including protection of life, health, the marine environment and property by using risk analysis and cost-benefit assessment (Jin Wang, 2006; Pedro Antao, 2008; Oyvind Berle, 2011). This is in essence a risk management technology and a scientific, casualty-date dependent approach.

Casualty database for risk based approach is essential for future regulatory activities. Starting now in constructing a solid database for the future is fundamental (Jongtae Jeong, 2010;). IMO has established a global system of casualty investigation with the harmonized procedures. Casualty investigation mechanism is essential for adoption the new code of casualty investigation by MSC 84. (Metin Celik, 2010). The causes of an accident have usually been described with the closest-to-accident factors. Researchers, however, have recently tended to analyze an accident more thoroughly – not only the accident itself but also the activities and factors prior to and subsequent to the accident. Some accident patterns were found to be preventable not by correcting driving behaviors but by adjusting behaviors prior to driving. Modeling accidents or combinations of factors and consequences has become an alternative way to understand the process of accidents. The derived chains imply causality between factors and consequences, and are usually called causal chains (Elvik, 2003) or scenarios (Fleury and Brenac, 2001).

The existing approaches to accidents analysis can be either analyzed by extracting the circumstances of certain accident consequences or by comparing the differences of circumstances between consequences. These approaches may represent a step forward to managing safety but may not be enough to address the vague and uncertainty of safety effectively. An accident may not occur if one or more undesirable activities in this process were removed. The causality between factors and consequences is interpreted by the derived outcomes, which are usually represented as combinations of factors, trees or rules. So there is a need to adopt a relatively new method called rough set theory as a complement to the complicated accident analysis.

RST is suitable for processing qualitative information that is difficult to analyze by standard statistical techniques which has been successfully applied to different fields. This paper presents the RST on maritime transportation safety and reliability. Firstly, this is due to the fact that maritime accident is relatively a rare phenomenon; therefore, it is difficult to collect statistically significant number of samples to derive conclusive analysis and prediction. Secondly, the classical statistical methods often include a priori assumptions on the sample distribution (Iftikhar U. Sikder * & Toshinori Munakata, 2009). A two-stage approach was then proposed for the purposes. Rough set theory and statistical tests were adopted at the first stage to derive rules for accidents. A multinomial logistic regression model was applied at the second stage to evaluate the effects of factors in accident outcomes for different types of accidents.

3.1 Data Preprocessing

2003-2009 JiangSu Province's maritime accident data is chosen to demonstrate the feasibility and usefulness of rough set theory and the proposed framework in accident chain analyses. The total number of maritime accidents, excluding invalid cases, was 216. The collected attributes and their corresponding categories are summarized in Table 2. Accident type is chosen as the decision attribute while the other attributes are considered as condition attributes. Rose2 is used in this paper. The data set includes premonitory factors for earthquakes consisting of seismic activity on 155 records of weekly measures of the concentration measured at three different locations (attributes A1–A3) and four measures of vessel factors (attributes A4–A8) and four measures of environmental factors. Route Segment type is the decision attribute.

The preprocessing algorithm is applied to these attributes. Firstly, discretization the death & injured data and economical loss data, second supplement the missing data. In this stage two different preprocessing tools are used to obtain the dependent attributes. The accident statistics with 15 attributes is used in the analysis. As a result of the analysis, {duty time, educational background, type of vessel, Port of registry, Speed limit, Type of through, Visibility } attribute set is effective on the fatality resulted accident, attributes that affects the fatality are shown in Table 1.

Table 1. The Attribute and Category Table of the Importance of Factors Contribute to Accidents

| Dimension | Parameter name | Parameter definition |
|--------------------------|------------------------------------|---|
| Seaman characteristics | duty time (A1) | Daytime/nighttime |
| | Seaman's License certificates (A2) | Valid, invalid, unknown |
| | educational background (A3) | Professional/other |
| vessel characteristics | Port of registry (A4) | CHINA,FOREIGN Registries, other |
| | Type of vessel (A5) | Cargo ship, HSC-passenger, Barge, Container, Fishing vessel etc |
| | Type of through (A6) | Non stopped vessel, stopped vessel |
| | Speed limit (A7) | Low, common, above |
| Navigational Environment | Type of accident (A8) | Collision & Contact, Foundering/sinking, Fire/explosion |
| | Direction of Underway (A9) | In/out |
| | Climate (A10) | Sunny or cloudy, rainy, other |
| | Visibility (A11) | Good, poor |
| Route Segment | Type of waterway (D) | Main Channels |
| | | Fairways |
| | | Precaution area |
| | | Anchorage & berthage |
| Fatal injured | Death(DJ) | Above 3, 1-3, injured but no death, no injured |
| Economic loss | Direct & Indirect(D) | Very high, high, common, low |

3.2 Reduction of Attributes

The basic construct in rough set theory is called a reduct. The extraction of reducts from data involves construction of minimal subset of attributes ensuring the same quality of sorting as that of all attributes. The term reduct was initially defined for sets rather than objects with input and output features or for decision tables with decision attributes and

outcomes. Reducts of the objects in a decision table have to be computed with consideration given to the value of the output feature. The original definition of reduct considers features only. In this paper, each reduct is viewed from four perspectives – feature, feature value, object, and rule perspective.

Reduct sets were determined by choosing Genetic Algorithm (GA) based reducts and selector method. An information table is sent to the integrated system for the GA-based reducts and selector. It is employed to reduce the input attribute set and conduct the optimization operation of GA. Input to a reducer algorithm is used as a decision table, and then a set of reducts is returned. The returned reduct set may possibly have a set of rules attached to it as a child. A reduct is a collection of attribute indices into the table which the reduct belongs to. Reducts of both these types can be computed modulo which the decision attributes or not.

3.3 Rule Identification Algorithm

Rule extraction is a relatively straightforward procedure. Reducts are used to generate decision rules from the decision table. The objective is to generate basic minimal covering rules or minimal number of possibly shortest rules covering all the cases. The classical LEM2 algorithm was used to derive minimal set of rules covering all the objects from learning set.

3.4 The Rule-validation Procedure

The predictive performance of the rules derived is tested for new instances using 10 cross-validation method. The following steps are applied to examine the objects in the testing data set to estimate the validity of the rules derived from the above algorithm (Z. Zou, *et al.*, 2011):

Step 1: Compare each decision rule derived from the rule composing algorithm with each new object from the testing data set. Calculate the number of objects that match with the rule.

Step 2: Repeat the comparisons of the decision rules with the objects from the testing data set until no decision rule is left.

Step 3: Calculate the accuracy of each rule by using the total matched objects divided by the summation of the total correctly matched objects and the total incorrectly matched objects. If the accuracy of the rule is greater than the predefined threshold value of confidence, then go to step 4; otherwise, remove the rule. Note that an incorrectly matched object means that the object contains the identical known value of conditional attributes with the rule, yet the outcomes are different from the rule.

Step 4: Stop and output the results of validated rules.

This method of cross-validation is particularly appropriate for the given dataset for several reasons. First, it allows using maximum possible number of training instances for learning tree. Secondly, since the number instances of decision variable is very small, 10-fold stratified cross-validation or percentage split method would not have produced sufficient number of training set for the learning tree. Thirdly, the procedure is essentially deterministic.

4. Empirical Study for Exploring Accident Occurrence and Losses

4.1 Identification Route Segment Prone to Accident

We evaluate the importance condition attributes for definition of objects' classification by values of chosen accident location decision attributes. Giving the representation of the dependencies between the minimum subset of condition attributes and decision ones in a form of a set of decision rules, the algorithm generates 16 minimum covering rules. The

level of discrimination indicates the ratio of the number of covered positive objects to the number of all objects covered by the rule.

Rules are generated from the accident database by rough set theory, and the significant rules for each accident type are shown. The accuracy of approximation for accidents is higher when more condition attributes are included, the hit rate (the percentage of correct prediction) for the precaution area is up to 70% when all condition attributes are considered, the hit rate for other route segment all range from 20% to 40% (Table 2), this suggests that that the occurrence of a precaution accident may follow similar paths and is more predictable because the precaution area has the most heavy traffic volume. But for other accident types, the rules generated from their training cases may not be representative since their occurrences are mostly random. Different route segment types have their corresponding useful condition attributes. For example, the condition attributes of vessels characteristics are useful for the precaution area and the other accident route segment, and those of environmental factors are useful for main channels accidents (Table 3). All these results are helpful for devising adequate countermeasures.

Table 2. Rough Set Results

| Accident Route Segment | Generated rules | Accuracy (%) | Quality of classification (%) | Hit rate (%) | Overall Hit rate (%) |
|------------------------|-----------------|--------------|-------------------------------|--------------|----------------------|
| Main Channels | 16 | 87.95 | 90.66 | 23.33 | 53.00 |
| precaution area | | 100.00 | | 66.67 | |
| Anchorage & berthage | | 74.12 | | 40.00 | |
| Fairways | | 66.84 | | 25.00 | |

Table 3. Description of Significant Rules

| Accident Route Segment | Rule description |
|--------------------------------|--|
| Main Channels ^a (8) | (A4 = 2) & (A10 = 2) & (A11 = 2) |
| | (A4 = 0) & (A6 = 2) & (A8 = 2) & (A11 = 0) |
| | (A3 = 0) & (A6 = 1) |
| | (A4 = 1) & (A6 = 1) & (A7 = 1) |
| | (A2 = 1) & (A9 = 0) |
| | (A2 = 1) & (A9 = 1) & (A11 = 2) |
| precaution area (15) | (A2 = 1) & (A6 = 2) & (A7 = 0) |
| | (A4 = 0) & (A8 = 0) |
| | (A3 = 2) & (A4 = 0) & (A6 = 0) & (A9 = 2) |
| | (A1 = 0) & (A3 = 2) & (A7 = 0) & (A11 = 0) |
| Anchorage & berthage (8) | (A1 = 0) & (A8 = 2) & (A11 = 1) |
| | (A6 = 2) & (A9 = 0) |
| | (A2 = 2) & (A5 = 1) |
| Fairways (7) | (A1 = 1) & (A6 = 0) & (A8 = 1) |
| | (A2 = 1) & (A4 = 1) |
| | (A1 = 1) & (A4 = 2) & (A8 = 2) |

^a The value represents the rule strength.

4.2 Identification of Injury Severity & Death and Economic Losses

The consequence of an accident is injury severity, death and economic losses. We evaluate the condition attributes for definition of objects' classification by values of chosen injury severity & death and economic losses decision attributes as Table 4. The classification results show that most of the injury severity & death accidents are assigned to the cargo ship and bulk ship least into the HSC-passenger types. This

suggests that, while most accidents are associated with some critical condition attributes which lead to the similar classification pattern, very high and high severity&death rate is related to very distinctive characteristics. This also implies that some similarities may exist in the occurrence of the injured and death types since they are all related to nighttime navigation and Seaman characteristics.

Table 4. Rough Set Results of Injury & Death

| injury & death | Generated rules | Accuracy (%) | Quality of classification (%) | Hit rate (%) | Overall Hit rate (%) |
|----------------------|-----------------|--------------|-------------------------------|--------------|----------------------|
| Above 3 | 18 | 47.25 | 80.54 | 36.67 | 50.50 |
| 1-3 | | 86.47 | | 58.33 | |
| Injured but no death | | 42.39 | | 10.00 | |
| No injured | | 51.54 | | 14.52 | |

Table 5. Description of Significant Rules of Injury & Death

| Accident Route Segment | Rule description |
|--------------------------|---------------------------------|
| Above 3 ^a (7) | (A10 = 2) & (A5 = 0) |
| | (A1 = 0) & (A3 = 0) & (A6 = 1) |
| | (A3 = 1) & (A9 = 1) & (A10 = 1) |
| 1-3 (19) | (A3 = 2) & (A6 = 0) & (A5 = 1) |
| | (A2 = 1) & (A8 = 0) |
| | (A1 = 0) & (A5 = 1) |
| | (A4 = 1) & (A8 = 2) |
| | (A3 = 0) & (A7 = 1) & (A5 = 2) |
| | (A6 = 1) & (A5 = 3) |
| Injured & no death (9) | (A2 = 2) & (A3 = 0) & (A5 = 3) |
| | (A5 = 1) & (A8 = 2) & (A11 = 2) |
| | (A2 = 1) & (A9 = 0) |
| | (A3 = 1) & (A9 = 2) & (A5 = 2) |
| | (A8 = 1) & (A10 = 1) |
| No injured (12) | (A6 = 0) & (A5 = 3) |
| | (A1 = 1) & (A5 = 0) & (A8 = 0) |
| | (A3 = 1) & (A8 = 2) & (A10 = 1) |
| | (A2 = 1) & (A5 = 2) & (A7 = 1) |

^a The value represents the rule strength.

Economic losses are mainly the loss of damage of ships which is counted by money. These similarities are the reasons for the low hit rates for the No injured and Injured but no death types, since they can be easily assigned to the collision accidents due to the fact that the when a collision happens, more or less it will produce the economic loss. As a consequence, more rules associated with the occurrence of the common economic losses are generated and dominate the classification pattern. On the other hand, the low or non-economic losses are more closely related to daytime, foreign registered.

Table 6. Rough Set Results of Economic Losses

| injury & death | Generated rules | Accuracy (%) | Quality of classification (%) | Hit rate (%) | Overall Hit rate (%) |
|----------------|-----------------|--------------|-------------------------------|--------------|----------------------|
| Very high | 9 | 32.36 | 76.05 | 35.00 | 58.50 |
| high | | 65.64 | | 55.00 | |
| common | | 72.74 | | 59.00 | |
| low | | 56.39 | | 45.00 | |

Table 7. Description of Significant Rules of Economic Losses

| Accident Route Segment | Rule description |
|--|--|
| Very high ^a (7) | (A1 = 0) & (DJ = 0) |
| | (A7 = 2) & (DJ = 0) |
| High (18) | (A4 = 0) & (DJ = 1) |
| | (A1 = 1) & (DJ = 1) |
| | (A3 = 2) & (A7 = 0) & (A8 = 2) & (A11 = 0) |
| Common (10) | (DJ = 2) |
| | (A6 = 0) & (A8 = 0) |
| low (13) | (D = 3) & (DJ = 3) |
| | (A3 = 2) & (DJ = 3) |
| ^a The value represents the rule strength. | |

5. Discussion and Conclusion

5.1 Discussion

While both methods provide algorithm for evaluating conditioning attributes, their inherent significance is entirely different. While the concept of reduct in rough set is based on elimination of superfluous or redundant attributes in a decision table. The focus is to identify minimal set of attributes that preserve the indiscernibility relation.

Compared with the analysis of negative binominal regression (NB), while maximal frequency of occurrence in rough set reducts is collision, environmental factors appear also significant suggesting a strong relationship of radon emanation and associated environmental factors, the result of NB accordant to the rough set analysis, which port of registry attribute has no significant effect on the happen of accidents, collision of cargo and bulk ships dominates the main accident pattern, more injured and death happen when the ship sink, daytime and nighttime are equivalent to accident happen but nighttime is more often to cost economic losses.

The result of negative binominal regression is as:

$$y1 = @exp(-2.456020768 \times a1 - 2.301870088 \times a2 + 1.675619819 \times a4 + 1.174598196 \times a6 + 2.20949467 \times a7 + 2.899588413 \times a9 + 0.05406722134 \times a10)$$

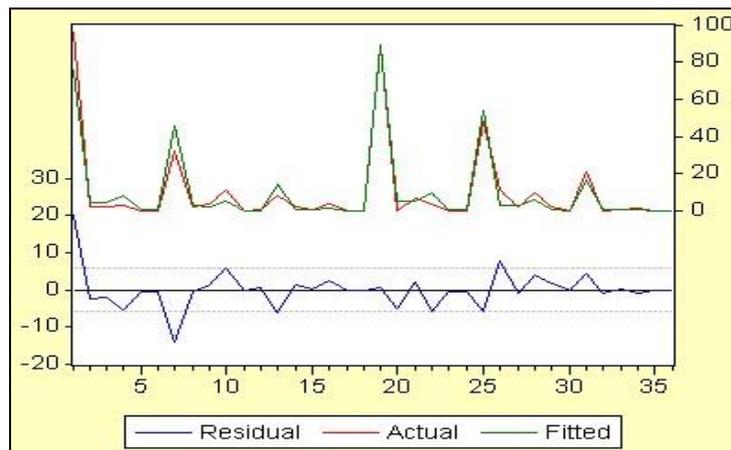


Figure 2. The Regression Result of Injured and Death

5.2 Conclusion

Rough set provide transparent method for inductive learning from data. This is especially important when the seismic activity need to be understood in term of its causal structure involving multiple factors and their interactions. Thus, the rules generated from these machine learning techniques could provide further insight into the complex dynamics of accidents.

Taking advantages of rough set, this research implemented the idea that the occurrence of an accident is a series of errors or mishandling. The illustrated case shows that it is feasible to apply rough set theory to analyze the links among affecting factors and accident types. The proposed factor structure can be easily transformed and extended based on an analyst's knowledge and on-hand accident databases, large number of condition attributes were included without any prior judgments except when being grouped with respect to the temporal and logical sequence of the occurrence of an accident. A condition attribute was dropped only when the removal did not have any impact on defining accident types. This procedure differs from conventional statistical approaches where non-significant attributes are usually immediately dropped and are sometimes claimed to have no impact on the occurrence of an accident.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (51149001).

References

- [1] L. de Zhong, S. X. Duan and C. Y. Sheng, *et al.*, "Characteristics of freeway traffic crashes in China", *Road Traffic & Safety*, vol. 4, no. 7, (2007), pp. 16-21.
- [2] M. Celik, S. M. Lavasani and J. Wang, "A risk-based modelling approach to enhance shipping accident investigation", *Safety Science*, vol. 1, no. 48, (2010), pp. 18-27.
- [3] F. Goerlandt and P. Kujala, "Traffic simulation based ship collision probability modeling", *Reliability Engineering & System Safety*, vol. 1, no. 96, (2008), pp. 91-107.
- [4] Z. Xiao, L. Chen and B. Zhong, "A model based on rough set theory combined with algebraic structure and its application: Bridges maintenance management evaluation", *Expert Systems with Applications*, vol. 7, no. 37, (2010), pp. 5295-5299.
- [5] I. U. Sikder and T. Munakata, "Application of rough set and decision tree for characterization of premonitory factors of low seismic activity", *Expert Systems with Applications*, vol. 1, no. 36, (2009), pp. 102-110.
- [6] K. Thangavel and A. Pethalakshmi, Dimensionality reduction based on rough set theory: A review.
- [7] G. Xie, J. Zhang, K. K. Lai and L. Yu, "Variable precision rough set for group decision-making: An application", *International Journal of Approximate Reasoning*, vol. 2, no. 49, (2008), pp. 331-343.
- [8] J. R. W. Merrick and J. R. Harrald, "Making decisions about safety in US ports and waterways Interfaces", vol. 3, no. 37, (2007), pp. 240-252.

- [9] B. Fabiano, F. Curro, A. P. Reverberi and R. Pastorino, "Port safety and the container revolution: A statistical study on human factor and occupational accidents over the long period", *Safety Science*, vol. 8, no. 48, (2010), pp. 980-990.
- [10] M. A. Yazici and E. N. Otay, "A Navigation Safety Support Model for the Strait of Istanbul", *Journal of Navigation*, vol. 4, no. 62, (2009), pp. 609-630.
- [11] O. S. Uluscu, B. Ozbas, T. Altioik, *et al.*, "Risk Analysis of the Vessel Traffic in the Strait of Istanbul", *RISK ANALYSIS*, vol. 10, no. 29, (2009), pp. 1454-1472.
- [12] P. Antao and C. G. Soares, "Causal factors in accidents of high-speed craft and conventional ocean-going vessels", *Reliability Engineering & System Safety*, vol. 9, no. 93, (2008), pp. 1292-1304.
- [13] J. T. Wong and Y. S. Chung, "Comparison of Methodology Approach to Identify Causal Factors of Accident Severity", *Transportation Research Record*, no. 2083, (2008), pp. 190-198.
- [14] O. F. Knudsen and B. Hassler, "(2011)IMO legislation and its implementation: Accident risk", vessel deficiencies and national administrative practices, *Marine Policy*, vol. 2, no. 35, (2006), pp. 201-207.
- [15] J.-F. Balmat, F. Lafont, R. Maifret and N. Pessel, "MARitime RiSk Assessment (MARISA), a fuzzy approach to define an individual ship risk factor", *Ocean Engineering*, vol. 15-16, no. 36, (2009), pp. 1278-1286.
- [16] E. Vanem and R. Skjong, "Designing for safety in passenger ships utilizing advanced evacuation analyses-A risk based approach", *Safety Science*, vol. 2, no. 44, (2006), pp. 111-135.
- [17] T. L. Yip, "Port traffic risks - A study of accidents in Hong Kong waters", *Transportation Research Part E: Logistics and Transportation Review*, vol. 5, no. 44, (2008), pp. 921-931.
- [18] P. Trucco, E. Cagno, F. Ruggeri and O. Grande, "A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation", *Reliability Engineering & System Safety*, vol. 6, no. 93, (2008), pp. 845-856.

Authors



Zhang Hao, He is a PHD from Shanghai Maritime University major in maritime safety, transportation information engineering and control.