Ocean Engineering 166 (2018) 135–144

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Combining zero-inflated negative binomial regression with MLRT techniques: An approach to evaluating shipping accident casualties

Jinxian Weng[a], Dong Yang[b,*], Ting Qian[c], Zhi Huang[c]

[a] College of Transport and Communications, Shanghai Maritime University, Shanghai, 201306, China
[b] Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hong Kong, China
[c] Navigation Institute, Jimei University, Xiamen, China

ARTICLE INFO

Keywords:
Maritime safety
Negative binomial regression
Maximum likelihood regression tree
Shipping accident

ABSTRACT

This study aims to develop a maximum likelihood regression tree-based (MLRT) ZINB (zero-inflated negative binomial) model to predict shipping accident mortality, and also to examine the factors which affect the loss of human life in shipping accidents. Based upon 23,029 sets of shipping accidents observations collected from 2001 and 2011 in global water areas, a tree comprising 7 terminal nodes is built, each of which is assigned by a separate ZINB model. Model results indicate that the large number of shipping accident casualties are closely related to collision, fire/explosion, sinking, contact, grounding, operating time, capsizing, docking condition, hull/machinery damage, and miscellaneous causes. In addition, it is found that there is a larger casualty count for the accidents occurring under adverse weather conditions or far away from coastal/port areas. In addition, sinking is recognized as the accident type which causes the largest number of casualties. This study can help the decision makers to propose effective strategies to reduce shipping accident casualties.

1. Introduction

The international shipping activities account for approximately 90% of the world trade, thus the safety of ship is critical to the global economy. From 2007 to 2016, the number of annual shipping losses decreased from 171 to 85 (Lloyd's List Intelligence Casualty Statistics). There were 2611 reported shipping causalities in 2016, 4% decreased compared to 2015. Although both the accidents and causalities both witnessed a decline in number, a growing complexity and interconnectivity of ship rick can be ascertained (Safety and Shipping Review 201). For example, the over-supply of ships in shipping market and world economic integration could accelerate the pace of development of larger ships. The big-sized ships carrying more passengers and/or crew members may lead to catastrophic consequences in terms of both property damage and human life loss, in 2016, the top ten largest vessel lost are mainly caused by large ship (Safety and Shipping Review, 2017).

To operate ship under the complicated and high-risk environments, it is of great importance for decision-makers to propose effective navigational safety strategies to decrease the loss of human life once an accident does occur. Hence, it is necessary for them to fully understand the causal factors that affect the fatality in shipping accidents within the limited resources and budgets. For example, the different types of ship accidents, such as collision, fire/explosion, sinking, contact, grounding, capsizing, hull/machinery damage, navigation status and adverse weather conditions could affect the occurrence likelihood of shipping accidents in many different ways. If the correlations between casualties and these shipping accidents can be recognized, the resources and budgets can be maximally utilized.

So far, many studies have been conducted on the analysis of the relationship between the contributory factors and the risk of shipping accidents (e.g., Sahin and Senol, 2015; Senol and Sahin, 2016). However, the data sources of most studies only cover specific water areas, which means the analysis results may not be applied to other water areas. This study extends the previous literature on exploring the contributory factors influencing human life loss caused by shipping accidents occurring to the global marine areas. The Zero-Inflated Negative Binomial (ZINB) regression technique based on classification regression tree is employed to realize the research objective.

2. Literature review

One stream of the previous literature on historic shipping accidents analysis can be divided into specific ship types. A large number of researches have paid their attentions on fishing vessel accidents. (e.g., Jin et al., 2001; Jin and Thunberg, 2005; Perez-Labajos et al., 2006;

* Corresponding author.
E-mail addresses: dong.yang@polyu.edu.cn, wjx206@gmail.com (D. Yang).

https://doi.org/10.1016/j.oceaneng.2018.08.011
Received 11 April 2018; Received in revised form 25 June 2018; Accepted 8 August 2018
Available online 13 August 2018
0029-8018/

Roberts et al., 2010). Other ship types such as tankers (Eliopoulou and Papanikolaou, 2007), passenger ships (Talley et al., 2006), RoPax vessels (Endrina et al., 2018) and cellular type containerships (e.g., Eliopoulou et al., 2013) have also been addressed. These studies failed to provide information on the effects of ship type on the shipping accident consequences. There are also studies which investigated the relationship between the contributory factors and the risk of shipping accidents considering all ship types. For instance, some researchers (e.g., Ozsoysal and Ozsoysal, 2006; Birpinar et al., 2009; Uluscu et al., 2009; Aydogdu et al., 2012) analyzed shipping accidents and proposed many navigation safety enhancement strategies considering multiple ship types, but the studies only cover the Istanbul waters. Weng et al. (2012) examined the effects of time and traffic directions on shipping accident frequency in the Singapore Strait. He also took all the ship types into account. However, these studies were limited to Singapore Strait water areas. The results and findings from these studies may not be applicable for other water areas. In addition, the formal safety assessment techniques have also been applied to evaluate shipping accident consequence. For example, Chai et al. (2017) built a quantitative risk assessment model to estimate ship accident risk combining accident frequency and consequence for different types of ships (e.g., container ships, Ro-Ro/passenger ships, and cargo ships) in Singapore port fairways.

To date, various statistical methods have been applied in order to evaluate the shipping accident consequences. Jin et al. (2001) estimated total losses and crew injuries in commercial fishing vessel accidents using Pobit and negative binomial regression methods. Talley et al. (2006) determined the total loss, injuries and deaths/missing people in passenger vessel accidents with Tobit, negative binomial and Poisson regression techniques. Yip (2008) applied the negative binomial regression technique to describe the injuries and casualties caused by ship accidents in Hong Kong Waters. Jin (2014) developed an ordered Probit model to estimate the ship damage and crew injury severity in fishing vessel accidents. Afenyo et al. (2018) applied Bayesian network approach to identify the most significant causative factors for various consequences. It is noticed, the fatalities or injuries may not occur in a shipping accidents, in other words, the data such as the loss of human life often have a large number of zero outcomes in maritime safety analysis, thus the fore-mentioned methods are not applicable to the shipping accident casualty analysis. Alternatively, the zero-inflated distribution is a commonly used method for the problem of excess zeros despites it is rarely used in shipping accident consequence analysis.

The least square tree methods have been broadly employed by researchers to analyze accident injury severity for other transportation modes. Kuhnert et al. (2000) applied multivariate adaptive regression splines (MARS), classification and regression trees (CART), and logistic regression to analyze motor vehicle injury data. Yan et al. (2010) presented the hierarchical tree-based regression (HTBR) approach to investigate train–vehicle crashes at highway–rail grade crossings. Weng et al. (2013) adopted a tree-based logistic regression method to assess work zone casualty risk. The results of the above studies indicated that the proposed approach outperformed the decision tree approach and the logistic regression approach. Nevertheless, one of the drawbacks of the least square tree method is that it cannot be applied to analyze a dependent variable with a large variance. As an extension of the least square regression tree, Torgo (2000) built a least absolute deviation regression tree. It was found that the built tree can alleviate the effect of large variance to some extent. However, both least absolute deviation regression trees and least square regression trees have poor stability. Hence, some researchers (e.g., Su, 2002; Mohamed et al., 2013) have proposed a maximum likelihood regression tree, which has a rigorous mathematical justification and better tree stability than the least square regression tree.

In summary, the existing literature clearly indicates that a zero-inflated negative binomial (ZINB) regression model considering maximum likelihood regression tree-based methods are applicable for shipping accident analysis. In order to produce better model performance, each terminal node in the regression tree is assigned a ZINB regression model.

3. Objectives & contributions

The objective of this study is to propose a maximum likelihood tree-based zero-inflated negative binomial regression model to predict the shipping accident mortality. Using the proposed model, we can accurately examine the various effects of influencing factors under different situations. The contributions of this study are two-fold. First, the proposed model could provide higher prediction accuracy than the conventional statistical regression analysis method and zero-inflated negative binomial regression models for shipping accidents proposed in our earlier studies (e.g., Weng and Yang, 2015; Weng et al., 2016). In addition, one single zero-inflated regression model could not account for the fact that one influencing factor may exhibit different exposure effects on the human life loss caused by shipping accidents under different situations in reality. In this study, we will build a zero-inflated negative binomial regression model for each leaf node of the maximum likelihood regression tree. Hence, the second contribution is that the proposed model containing several zero-inflated negative binomial regression models could fully capture the heterogeneous effects of some influencing factors in shipping accidents.

4. Methodology

4.1. Maximum likelihood regression tree-based model

A decision tree can present the decision-making process intuitively and accurately. Tree structure is a finite set of one or more nodes. A complete tree should contain the following three kinds of nodes: (i) a root node; (ii) internal nodes; (iii) leaf nodes. Like the root of a big tree, the root node contains all the statistics and has no previous node. Internal nodes, including parent nodes and child nodes, can be divided continuously. Leaf nodes indicate a certain attribute of the research subject and cannot be further divided. The number of leaf nodes and tree layers represents the size of the tree thus also can reflect the complexity of the tree. The larger the tree size, the lower the accuracy of the underlying decision making scheme is and the poorer the stability is.

In principle, the decision tree has two categories: classification tree and regression tree. Since the loss of human life in shipping accidents is a continuous target variable, we adopt the regression tree in this study. In the regression tree, the growth of the tree is based on the result of testing the partition data on the parent node. In order to find the node to which the data point belongs, one can trace a path down the tree according to the features of the data from the root node. In the meantime, each terminal node will be assigned a Zero-inflated negative binomial regression model to describe the casualties of shipping accidents.

In general, if there are a lot of zero values in the data and the probability of occurrence of its zero value observation exceeds the probability that its distribution can bear, the data is regarded as zero-inflated (ZI) data. For instance, in the study of forest fires, the total number of forest fires is zero-inflated due to the significant increase of non-fire risks. In this study, the number of casualties in shipping accidents is also zero-inflated data.

A zero-inflated distribution is actually a mixture of two distributions including a distribution on zero ("δ state") and a distribution on the non-negative integers ("σ state"). In general, a data record is in the δ state with probability φ and in the σ state with probability $1-\varphi$. If the data record is in the δ state, it takes only the value of zero. If the record is in the σ state, it follows a distribution on nonnegative integers (including the value of zero). Lambert (1992) proposed a Zero-Inflated Poisson (ZIP) regression model with covariates, and since then the ZIP