

# Combining Zero-inflated Negative Binomial Regression with MLRT Techniques: an Approach to Evaluating Shipping Accident Casualties

Jinxian Weng<sup>a</sup>, Dong Yang<sup>b,\*</sup>, Ting Qian<sup>c</sup>, Zhi Huang<sup>c</sup>

<sup>a</sup>*College of Transport and Communications, Shanghai Maritime University, Shanghai, China 201306*

<sup>b</sup>*Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hong Kong, China*

<sup>c</sup>*Navigation Institute, Jimei University, Xiamen, China*

## 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.

Keywords: Maritime safety; Negative Binomial regression; Maximum likelihood regression tree; Shipping accident

## 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

---

\*Dr. Dong Yang, Email: [dong.yang@polyu.edu.hk](mailto:dong.yang@polyu.edu.hk)

Casualty Statistics). There were 2,611 reported shipping casualties in 2016, 4% decreased compared to 2015. Although both the accidents and casualties both witnessed a decline in number, a growing complexity and interconnectivity of ship risk 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; 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 despite 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 (“ $\delta$  state”) and a distribution on the non-negative integers (“ $\sigma$  state”). In general, a data record is in the  $\delta$  state with probability  $\phi$  and in the  $\sigma$  state with probability  $1-\phi$ . If the data record is in the  $\delta$  state, it takes only the value of zero. If the record is in the  $\sigma$  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 regression model is gaining more and more attention. In order to solve the problem of over dispersion in zero-inflated data, Martin (2001) studied the Zero-Inflated negative binomial (ZINB) regression model and demonstrated that it is more appropriate than the ZIP model for processing ZI data with large deviations. Specifically, the probability function of a ZINB regression model can be expressed by

$$f\langle y_i | X_i, \beta, \tau, \alpha \rangle = \begin{cases} \phi_i + (1 - \phi_i) \left( \frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} & , y_i = 0 \\ (1 - \phi_i) \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1) \Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left( \frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{y_i} & , y_i > 0 \end{cases} \quad (1)$$

$$\ln(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} = \mathbf{X}_i \boldsymbol{\beta} \quad (2)$$

$$\Phi_i = \Phi_i \langle y_i = 0 | \mathbf{X}_i, \boldsymbol{\beta}, \tau \rangle = \varphi(\tau \mathbf{X}_i \boldsymbol{\beta}) \quad (3)$$

where  $\mathbf{X}_i$  is the vector of the explanatory variables for the negative binomial regression model,  $\boldsymbol{\beta}$  is the vector of coefficients,  $\phi_i$  is the probability of  $\delta$  state,  $\alpha$  is the dispersion parameter,  $\tau$  is the adjustment parameter,  $\varphi$  denotes the cumulative distribution function of a standard normal random variable.

The estimates of  $\boldsymbol{\beta}$ ,  $\tau$  and  $\varepsilon$  can be determined by maximizing the log-likelihood function  $\ln L\langle \boldsymbol{\beta}, \tau, \alpha | y, \mathbf{X} \rangle = \sum_{i=1}^n \ln f\langle y_i | X_i, \beta, \tau, \alpha \rangle$ . The procedures for estimating the coefficients of the generalized linear models and generalized additive models are not applicable for the ZINB regression models due to the fact that it is unknown which state the zero-valued observations belong to. We can use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to determine the maximum likelihood estimates. In the EM algorithm, one can introduce a new random variable  $K$ , which is only partially observable. If  $y$  is strictly greater than zero, the value of  $Z$  is known (i.e., zero) while it is unknown otherwise. More specifically, the EM algorithm at the  $j^{th}$  iteration includes the following three steps:

(i) E-step: Compute

$$K_i^j = E\langle K_i | y_i, \beta^{j-1}, \tau^{j-1}, \alpha^{j-1} \rangle = \begin{cases} \frac{\Phi_i^{j-1}}{\Phi_i^{j-1} + (1 - \Phi_i^{j-1}) \left( \frac{\alpha^{j-1}}{\alpha^{j-1} + \lambda^{j-1}} \right)^{\alpha^{j-1}}} & , y_i = 0 \\ 0 & , y_i > 0 \end{cases}$$

(ii) M-step for  $\beta$  and  $\alpha$  : Determine the estimates  $\beta^j$  and  $\alpha^j$  by fitting the negative binomial regression model using the weights  $(1 - K_i^j)$  and the response variable  $y_i$ .

(iii) M-step for  $\tau$  : Determine the estimate  $\tau^j$  by fitting the Probit regression model using the  $\beta^j$  and response variable  $K_i^j$ .

#### 4.2. Construction of maximum likelihood tree structure

In general, the classification process of the tree analysis method contains two steps: the first is tree growing and followed by tree pruning. All the data samples should be divided into training samples and validation samples. The two parts of data correspond to the two stages of the tree structure respectively.

##### 4.2.1. Tree growing

A top-down recursive method is applied to construct the tree structure. When the training data enters the root node of a maximum likelihood tree, a maximum likelihood splitting (MLS) algorithm is used to search for all possible splits among all variables. The MLS algorithm is a greedy search algorithm that aims to maximize the log-likelihood of the tree-based model by splitting the data in a parent node into several subgroups. More specifically, the following steps can further explain how to determine which variable can be used to optimally separate the parent node.

(a). The maximum log-likelihood for the node  $k$ , denoted by  $LL(k)$ , is calculated by

$$LL(k) = \sum_{i=1}^{n_k} \ln f\langle y_i | X_i, \beta, \tau, \alpha \rangle$$

where  $n_k$  is the number of observations in node  $k$ , and  $y_i$  is the shipping accident



casualties of the  $i^{\text{th}}$  observation.

(b). Calculate the maximum log-likelihood increment on the parent node  $k$  caused

$$\text{by a split } \omega. \quad \Delta LL(x, \omega, k) = LL(k_R) + LL(k_L) - LL(k)$$

$k_R$  and  $k_L$  are the right and left child nodes of the parent node  $k$  respectively. The best split  $\omega^*$  that causes the maximum increment in log-likelihood can be selected by searching all possible splits of the explanatory variable  $x$ .

$$\Delta LL(x, \omega^*, k) = \max_{x \in W} \Delta LL(x, \omega, k)$$

$W$  represents the set of all possible splits for the variable  $x$ .

(c). The best splits for all explanatory variables are determined by repeating Step (b). Among these best splits, the split  $\omega^*$  of the variable  $x^*$  that causes the global maximum increment in log-likelihood can be determined

$$\Delta LL(x^*, \omega^*, k) = \max_{x \in X} \Delta LL(x, \omega^*, k)$$

where  $X$  represents the explanatory variable set.

(d). If  $\Delta LL(x^*, \omega^*, k) \leq 0$ , the parent node  $k$  will be considered as the terminal node. Otherwise, we choose the variable  $x^*$  and split  $\omega^*$  to split the parent node  $k$ .

(e). When  $\Delta LL(x^*, \omega^*, k) \leq 0$ , or the tree depth and the number of leaf nodes reaches the threshold value, the process stops. Otherwise, return to step (a).

#### 4.2.2. Tree pruning

Constructing a basic tree structure ensures that the tree fits perfectly with the data. In fact, this may cause the model to over-fit the data. Overfitting problem is a crucial issue in the tree analysis method, which will reduce the applicability of the model and also make the model out of the actual situation. Therefore, the tree structure construction process must prune the grown tree. According to Akaike (1974), the

Akaike Information Criterion (AIC) statistics can be used to represent the cost complexity of a maximum likelihood regression tree  $Q$ . Namely,

$$AIC(Q) = -2LL(Q) + 2 \cdot (|\tilde{Q}| + 1)$$

where  $AIC(Q)$  is the  $AIC$  statistic of the tree  $Q$ ,  $LL(Q)$  is the maximum log-likelihood of the tree  $Q$ , and  $|\tilde{Q}| + 1$  is the total number of parameters of the ZINB models in the tree  $Q$ . The small tree with the best performance can be selected by minimizing the  $AIC$  statistic.

In general, pruning process is to verify the internal node from the bottom up. If the reduction of internal nodes and their associated leaf nodes reduces the  $AIC$  value of the entire model, the internal node will be subtracted.

## 5. Data description

In this study, we collected shipping accident data from the shipping accident database managed by the Lloyd's List Intelligence Company. This database has records of shipping accidents occurred in 33 major worldwide water areas. In the shipping accident database, each set of the recorded data includes the following information: (i) vessel; (ii) casualty date and time; (iii) casualty area; (iv) precis; (v) commercial operator; (vi) accident types; (vii) the loss of human life, including passengers and crew members who was dead or missing in the accident. In this study, the casualty date and time was divided into daytime and night-time periods. The daytime period is defined as the period from the local time of sunrise to the time of sunset. We classified the accident location based on the distance to the harbor or port. Using the Centroid Clustering (CC) algorithm, the optimum number of categories can be determined. In the CC algorithm, the objective function is to minimize the sum of the squared distances from the category means. The algorithm terminates when the number of iterations arrives at the preset maximum value. The optimum number of categories was found to be two: Category 1 (0, 20 km) and Category 2 (20 km,  $\infty$ ). Hence, the accident location was divided into these two categories: Category 1 (0, 20

km) indicating “near the coastal area/harbor/port” and Category 2 (20 km,  $\infty$ ) indicating “far away from the coastal area/harbor/port”. Similarly, the ship type was classified into two groups: (i) cruise ships (e.g., cruise ships, ferries, passenger ships and Ro-Ro ships); (ii) other ship types (e.g., container carriers, cargo ships).

The accident types we investigated include collision, fire/explosion, sinking, contact, grounding, hull damage (e.g., cracks, structural failure) and machinery damage (e.g., lost rudder, fouled propeller), capsizing, and accidents due to miscellaneous causes. What calls for special attention is that a collision refers to the situation where a ship struck or was struck by another ship on the water surface. Contact is defined as a situation where the ship struck any fixed or floating objects other than those included under collision or grounding. Grounding refers to a situation where a ship is in contact with the sea bottom or a bottom obstacle. Capsizing is defined as a situation where a ship is turned on its side or she is upside down. Note that capsizing does not mean that the ship will sink. In addition, capsizing is not the only prerequisite of sinking. The occurrence of contact, collision and grounding might also lead to the sinking of ship.

A total of 23,029 shipping accidents occurring in the global marine areas were recorded from 2001 to 2011. Of these accidents, there are at least one death and/or missing people in 2 ship accidents. Table 1 presents the variables and their descriptive statistics. The mean statistics reveal that collision, fire/explosion, sinking, contact, grounding, capsizing, hull and machinery damage and accidents caused by miscellaneous reasons account for 19.1%, 8.38%, 7.72%, 8.43%, 18.8%, 1.44%, 37.6% and 12.8% of total shipping accidents in the global marine areas respectively. The majority of shipping accidents occurred in good weather conditions (94.0%), and 40.4% of accidents occurred at night, 9.0% of shipping accidents occurred when the ship was docked. The loss of human life resulting from shipping accidents ranges from 0 to 1800, with an average value of 0.43 fatalities or missing persons per accident.

Figure 1 shows the comparison results of the frequencies of variables in both two cases. According to the loss of human life, we divide the data into two groups, one is

the frequencies of explanatory variables when no fatalities or missing persons occur, the other is the frequencies of explanatory variables when at least one person was killed or missing. According to the figure, the differences of the frequencies when “operating time = 1” is relatively little in both two conditions, which indicates that the operating time has an insignificant effect on the loss of human life. In other words, whether the accident occurs in night-time period or not, the mortality will not change a lot. Moreover, when at least one person is killed or missing, the frequency of “sinking=1” and “capsizing = 1” is much bigger than that of the other condition, which means if a sinking or a capsizing occurs, the probability of death will increase. The frequency of “hull and machinery damage = 1” is also much bigger than that of the other condition when death or missing of human occurs, which surprisingly indicates if a hull or machinery damage occurs, the probability of death will decrease.

In addition, the probability of death will rise in the shipping accident, if fire/explosion, sinking, capsizing accidents occurring far away from the coastal area/harbor/ports in adverse weather conditions. Unexpectedly, Figure 1 also shows that the occurrence of collision, contact, and hull or machinery damage accidents are unlikely to cause fatal accident.

One possible reason for these results may be that such kind of univariate statistical technique only allows the analysis of a single factor at a time. However, the consequence of a shipping accident is influenced by multiple factors simultaneously. Therefore, biased or incorrect results may be produced by isolating a single factor for analysis while treating other factors as fixed. It is more applicable to employ multivariate analysis, for example, ZINB regression technique.

## **6. Results analysis**

The process of the maximum likelihood tree-based model fusing the 19441 training data records and 3588 validation data records is illustrated in Fig. 2. It can be observed that the negative log-likelihood for the training data decreases with the increasing number of leaf nodes. However, when the number of leaf nodes is smaller

than 7, the decrease of the number of leaf nodes does not contribute to the decrease in the AIC value for the validation data. Therefore, the tree has 10 leaf nodes.

The structure of the maximum likelihood tree is presented in Fig. 3. With the training data, the initial split at the root node is based on the sinking that causes the largest increment in the log-likelihood. The fact that the largest log-likelihood increment is caused by sinking can be put down to the fact that shipping accidents involving sinking are very fatal. The tree directs the sinking accident attribute “Sinking” to the left forming leaf node 1, and the attribute “No Sinking” to the right forming Internal Node 1. The maximum likelihood splitting algorithm then continues splitting Internal Node 1 based on the distance into two groups (Internal Nodes 2 and 3). Internal Node 3 is further split based on the factors of weather, miscellaneous non-classified causes, fire or explosion, and operating time.

It can also be found from Fig. 3, sinking is strongly connected with a large number of deaths and missing people in a shipping accident. For non-sinking incidents, accident location has the biggest influence on the shipping accident casualties. According to the structure of the maximum likelihood tree, the loss of human life in shipping accidents is significantly influenced by the following explanatory variables in sequence: sinking, accident location, weather, miscellaneous causes, fire/explosion, and operating time.

Fig. 3 shows that the observed mean of the shipping accident casualties caused by sinking accident is 5.2 people, which is 85.2 times of casualties caused by non-sinking accidents. The reason why the observed mean value of shipping accident casualties caused by sinking accident is significantly higher than other causes can probably be explained by the fact that many factors could eventually lead to sinking accidents. Obviously, there will be more casualties when a shipping accident is caused by more than one factor. Similar to sinking, accident location also has a significant influence on the shipping accident casualties. By comparing Leaf node 2 and Internal Node 2, we can see that the mean value of casualties caused by accidents far away from the coastal area/harbor/port is 79.6% higher than that for accidents near the coastal

area/harbor/port.

The Poisson regression model, negative binomial regression model, ZIP regression model and ZINB regression model are also tested to demonstrate that the proposed model provides the best fit with these data. Using the measures such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Deviance Information Criterion (DIC), one can compare the model performance and select the optimal model. The AIC model is defined as  $AIC = -2 \ln L + 2\mu$ , where  $\ln L$  represents the fitted log-likelihood and  $\mu$  is the number of parameters for the model. The BIC is expressed as  $BIC = -2 \ln L + \mu \ln(N)$ , where  $N$  is the sample size. As a hierarchical modeling generalization of AIC and BIC, the DIC model is defined as  $DIC = -2 \ln L + \frac{Var(-2 \ln(p(y|\pi)))}{2}$ , where  $p(y|\pi)$  is the likelihood function and  $\pi$  are the parameters for the model. In general, the lower the values of the AIC, BIC and DIC statistics are, the better fit the model can perform. Fig. 4 shows the measures of fit comparison results of different regression models. It is obviously that our model provides the best fit to the shipping accident data owing to the smallest AIC, BIC and DIC statistics. Furthermore, Table 2 presents the mean absolute percentage errors (MAPE) of different regression models. It can be observed that the MAPEs from our model are the smallest for the training and validation data, respectively. This implies that our model could provide higher accuracy in predicting the number of fatalities than other models.

Table 3 shows the ZINB regression model results in each leaf node. From the table, it can be seen that “ship type”, “collision”, “weather” and “accident location” have positive coefficients in the ZINB model for Leaf node 1. This suggests that, for shipping accident casualties involving sinking accidents, the loss of human life will increase if the shipping accident occurs on a cruise under adverse weather conditions or far from the coastal area/harbor/ports. The results in Table 3 also indicate that shipping accident casualties are significantly influenced by the following explanatory

variables: sinking, accident location, weather, miscellaneous causes, fire/explosion, and operating time.

It can be seen from Table 3 that many of the factors affecting shipping accident casualties do not take effect independently. Instead, they cause the loss of human life jointly. Table 3 shows that the effect of the same factor on different types of accidents may be different or sometimes even be the opposite. For example, the role of collision variables in leaf node 1 and leaf node 2 is different because the coefficients of the ZINB model expressions for the two leaf nodes are different. Therefore, a collision accident exerts different influences in the event of sinking other than in the event when the accident is away from the coast. Specifically, when a sinking accident occurs, the accident casualties resulting from a collision accident will be less than that when there is no sinking accident but the accident location is far away from the coast.

As can be seen from Figure 3 and Table 3, the same kind of factors, such as collision accidents, may affect the casualties caused by many types of shipping accidents at mean time. The traditional statistical regression analysis method and the ZINB model can only analyze the impact of the independent variables on the target variables. Our model divides the target variable into several subspaces with category attributes and analyzes the respective effect caused by each factor in the data subspace. In other words, our model can explore the impact of the same factors on the loss of human life caused by different types of shipping accidents. For example, it can be found from Leaf node 1 that the sinking accident (i.e., Leaf node 1) occurred far away from the coastal area/harbour/ports could result in 24.9% higher mortalities, which is quite similar to the finding in the South China Sea (Weng et al. 2016). However, the accident location exhibits no effect for non-sinking accidents (i.e., Leaf node 2). This implies that maritime authorities should take corresponding navigational safety strategies for different kinds of shipping accidents.

As mentioned by Weng and Yang (2015), Australasia has a much lower proportion of collisions than the East Mediterranean & Black Sea, whereas the former is associated with bigger proportion of fire/explosion accidents. Similarly, the effect of

each influencing factor not only depends on the accident circumstances but also varies substantially with different seas/water areas. For instance, the collision was found to be the accident type that could result in the maximum increase of human life loss in the US Coast areas (Talley et al., 2006) while the accident type associated with the largest number of mortalities was found to be fire/explosion in Hong Kong waters (Yip, 2008) and sinking in the South China Sea (Weng et al., 2016).

## **7. Conclusions**

This study employed the ZINB regression technique based on classification regression tree to evaluate the relationship between shipping accident casualties and the corresponding contributory factors. The study was conducted with 23,029 sets of shipping accidents observations from 2001 and 2011 in the global marine areas. These data are used to train and validate the developed maximum likelihood regression tree-based (MLRT) model. A tree comprising 7 terminal nodes was constructed to predict the loss of human life in shipping accidents. For each terminal node, A ZINB model is built.

The results of our developed model show that the high human life loss is associated with sinking accidents, fire/explosion accidents and miscellaneous shipping accidents occurring under adverse weather far away from the coastal area/harbor/ports during the night-time periods. As was expected, probability of human life loss is likely to increase for the accidents occurring under adverse weather conditions and/or far away from the coastal area/harbor/port. This study implies that it is critical to prevent ships from sinking when an accident occurs so that the resulting loss can be reduced significantly. The policy-makers could employ our model to assess various safety measures and strategies. Insurance companies can also make use of our model to evaluate the possible loss of human life and damage to ships. It is noteworthy that the ZINB regression technique is only applicable when there are excess zeros and the distribution of the observed counts is over-dispersed.

For comparison, four widely applied techniques, including Poisson regression,



Negative binomial regression, ZIP regression and ZINB regression were conducted using the same training data. The results of the comparison show that our developed model provides the best fit to the shipping accident data owing to the smallest AIC, BIC and DIC statistics. This confirms the advantage of our developed model in predicting the fatality in shipping accidents.

Our model is able to explain the heterogeneity effect and to provide a better goodness-of-fit in predicting the loss of human life in shipping accidents. In addition, on account of the fact that it is able to avoid the over-fitting problem owing to the use of the AIC-based pruning criterion, our developed model distinctly outperforms the traditional ZINB model. This also demonstrates that the developed model is a good alternative for predicting human life loss in shipping accident.

One limitation of this study is that the injuries in shipping accidents were not considered in this study. The injuries resulting from shipping accidents will be taken into account in future studies. Because of data limits, some factors affecting shipping accident consequence such as the education and training of the crew, ship hull material and structure and the ship speed have not been considered in this study. However, it should be pointed out that some of the influencing factors may be highly related to these factors. For example, the consequence for ship collisions may be affected by the ship speed and collision angle. In this situation, the effects of some influencing factors considered in this study may be biased. Therefore, our future study will have to take into account these missing factors after collecting more data.

## **References**

- Afenyo, M., Khan, F., Veitch, B., Yang, M., 2018. Arctic shipping accident scenario analysis using Bayesian Network approach. *Ocean Engineering*, 133, 224-230.
- Agarwal, D.K., Gelfand, A.E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9:341-355.
- Birpinar, M.E., Talu, G.F., Gonencgil, B., 2009. Environmental effects of maritime

- traffic on the Istanbul Strait. *Environmental Monitoring and Assessment*, 152, 13–23.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1), 1–38.
- Eliopoulou, E., Hamann, R., Papanikolaou, A., Golyshev, P., 2013. Casualty analysis of cellular container ships. *Proceedings of the IDFS 2013, Shanghai*, 25–27.
- Endrina, N., Rasero, J.C., Konovessis, D., 2018. Risk analysis for RoPax vessels: a case of study for the strait of Gibraltar. *Ocean Engineering*, 151, 141–151.
- Garay, A.M., Hashimoto, E.M., 2011. On Estimation and Influence Diagnostics for Zero-Inflated Negative Binomial Regression Models. *Computational Statistics and Data Analysis*, 55, 1304–1318.
- Hall, D.B., 2000. “Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study” . *Biometrics*, 56, 1030–1039.
- Jin, D., 2014. The determinants of fishing vessel accident severity. *Accident Analysis and Prevention*, 66, 1–7.
- Kuhnert, P.M., Do, K.-A., Mc Clure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. *Comput. Stat. Data Anal*, 34 (3), 371–386.
- Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- Minami, M., Lennert-Cody, C. E., Gao, W., Roman-Verdesoto, M., 2007. Modeling shark by catch: the zero-inflated negative binomial regression model with smoothing. *Fisheries Research*, 84, 210–221.
- Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Saf. Sci*, 54, 27–37.
- Ridout, M., Hinde, J., Demetrio, C. G. B., 2001. A Score Test for Testing a Zero-Inflated Poisson Regression Model against Zero-Inflated Negative Binomial

- Alternatives. *Biometrics*, 57: 219-223.
- Sahin, B., Senol, Y.E., 2015. A novel process model for marine accident analysis by using generic fuzzy-AHP algorithm. *Journal of Navigation*, 68(1), 162-183.
- Senol, Y.E., Sahin, B., 2016. A novel real-time continuous fuzzy fault tree analysis (RC-FFTA) model for dynamic environment. *Ocean Engineering*, 127, 70-81.
- Su, X., 2002. Maximum likelihood regression trees. In: *ASA Proceedings of the Joint Statistical Meetings*. American Statistical Association, 3379–3383.
- Su, X., Wang M., Fan, J., 2004. Maximum Likelihood Regression Trees. *Journal of Computational & Graphical Statistics*, 13(3), 586-598.
- Torgo, L., 2000. Inductive learning of tree-based regression models. *AI Commun*, 13 (2), 137–138.
- Talley, W. K., Jin, D., Kite-Powell, H. L., 2006. Determinants of the severity of passenger vessel accidents. *Maritime Policy and Management*, 33, 173–186.
- Weng, J., Meng, Q., Qu, X., 2012. Vessel collision frequency estimation in the Singapore Strait. *Journal of Navigation*, 65, 207–221.
- Weng, J., Meng, Q., Wang, D., 2013. Tree-based logistic regression approach for work zone casualty risk assessment. *Risk Analysis*, 33 (3), 1539–6924.
- Weng, J., Yang, D., 2015. Investigation of shipping accident injury severity and mortality. *Accident Analysis and Prevention*, 76, 92-101
- Weng, J., Ge. Y., Han, H., 2016. Evaluation of Shipping Accident Casualties using Zero-inflated Negative Binomial Regression Technique. *Journal of Navigation*, 69.433-448.
- Yip., 2008. Port traffic risks – A study of accidents in Hong Kong waters. *Transportation Research Part E*, 44, 921–931.
- Yan, X., Richards, S., Su, X., 2010. Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings. *Accident Analysis and Prevention*, 42 (1), 64–74.

**Table 1**  
Variable descriptions

<b>Variables</b>	<b>Attributes</b>	<b>Mean</b>	<b>Stdev</b>	<b>Minimum</b>	<b>Maximum</b>
<b>Dependent variable</b>					
Loss of human life	Number of persons who were killed or missing	0.430	14.381	0	1800
<b>Independent variable</b>					
Ship type	1 if the ship is a cruise, 0 otherwise	0.001	0.029	0	1
Accident type					
Collision	1 if a collision is occurred, 0 otherwise	0.191	0.393	0	1
Fire/explosion	1 if a fire/explosion is occurred, 0 otherwise	0.084	0.277	0	1
Sinking	1 if a sinking occurs, 0 otherwise	0.077	0.267	0	1
Contact	1 if a contact occurs, 0 otherwise	0.084	0.278	0	1
Grounding	1 if a grounding occurs, 0 otherwise	0.188	0.391	0	1
Cpasizing	1 if a cpasizing occurs, 0 otherwise	0.014	0.119	0	1
Hull/Machinery damage	1 if a hull damage or machinery damage occurs, 0 otherwise	0.376	0.484	0	1
Miscellaneous	1 if the accident is caused by miscellaneous non-classified causes, 0 otherwise	0.128	0.334	0	1
<b>Operating conditions</b>					
Weather	1 if the accident is occurred under adverse weather conditions, 0 otherwise	0.043	0.203	0	1
Accident location	1 if the accident location is far from the coastal area/harbour/ports, 0 others	0.062	0.241	0	1
Operating time	1 for the night-time period, 0 otherwise	0.404	0.491	0	1
Docking condition	1 if the ship is docked or moored, 0 otherwise	0.090	0.286	0	1

**Table 2**

Prediction accuracy for different regression models

$MAPE^*$	Poisson regression model	Negative binomial regression model	ZIP regression model	ZINB regression model	Our model
Training data	15.32%	13.20%	13.45%	12.33%	10.85%
Validation data	10.30%	9.39%	9.43%	8.86%	7.65%

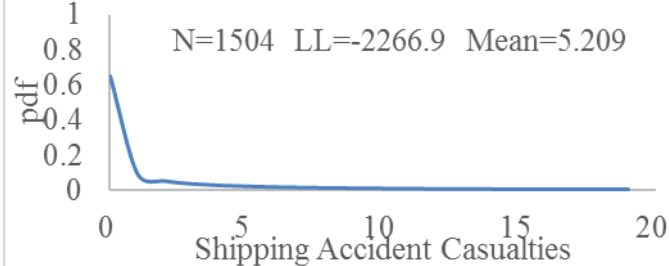
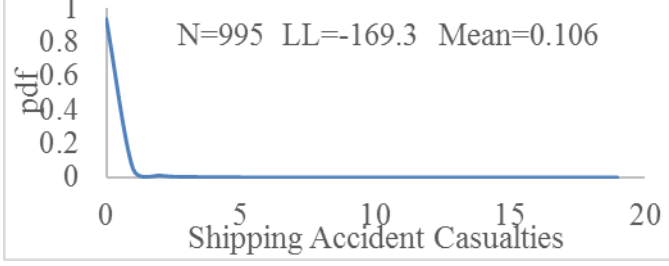
\* $MAPE$  represents the mean absolute percentage error that has been widely used to evaluate the model accuracy. It can be expressed by

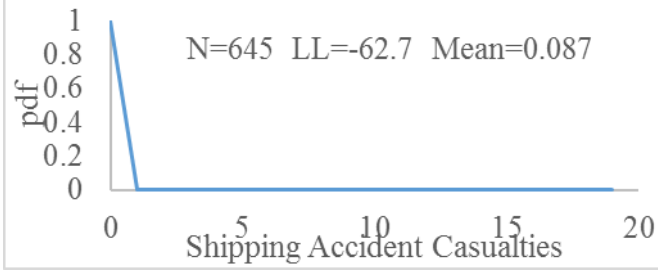
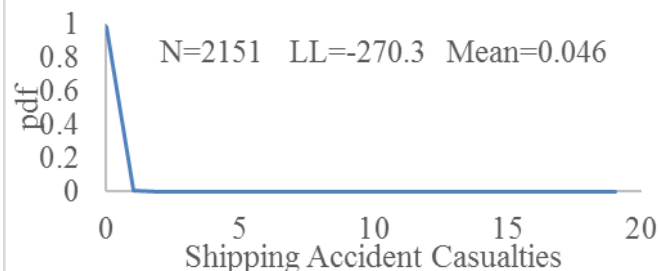
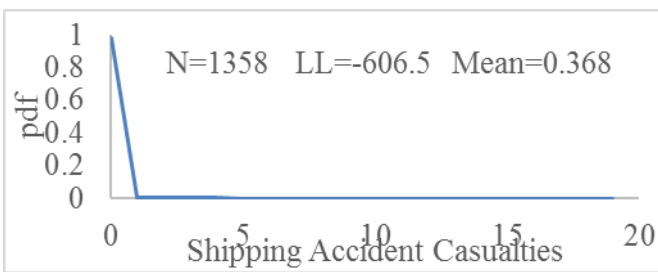
$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i^o - y_i^p}{y_i^o} \right|, \text{ where } n \text{ is the number of observations, } y_i^o \text{ is the observed value for the } i^{\text{th}} \text{ observation and } y_i^p \text{ is the predicted value}$$

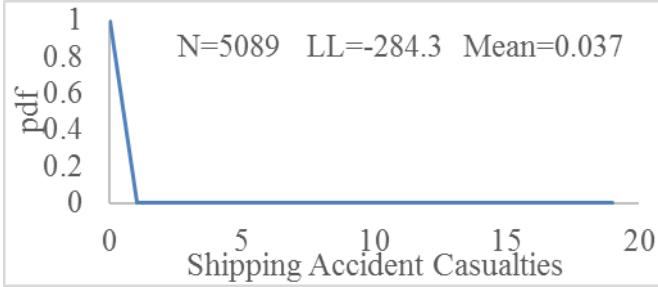
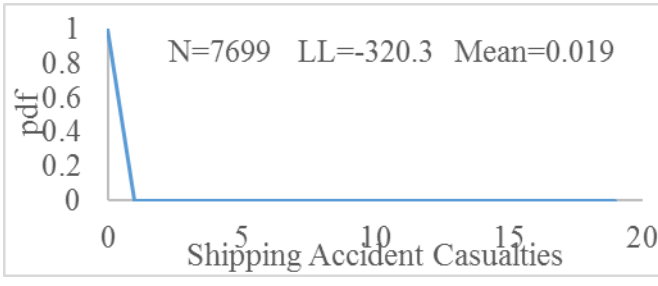
for the  $i^{\text{th}}$  observation.

**Table 3**

ZINB model results for each leaf node

Leaf node i	ZINB models			
	$\lambda$	$\tau$	$\alpha$	Probability density function
Leaf node 1	$=\exp(1.0177+4.6888*\text{ship type}+0.2038*\text{collision}-0.2003*\text{fire/explosion}-0.1958*\text{contact}-0.6359*\text{grouding}-0.7468*\text{miscellaneous}+0.7031*\text{weather}+0.2225*\text{accident location}-1.1796*\text{docking condition}-0.2307*\text{hull/machinery damage})$	-1.0311	5.1613	
Leaf node 2	$=\exp(-2.4213+4.203*\text{collision}+1.2461*\text{fire/explosion}+5.2728*\text{grouding}-3.9153*\text{weather}+0.6443*\text{operating time}-29.8322*\text{docking condition}-4.3114*\text{hull/machinery damage})$	1.1661	7.5356	

Leaf node 3	$= \exp(-0.0296 + 0.013 * \text{grouding} + 0.0002 * \text{miscellaneous} + 0.0058 * \text{operating time} + 1.4632 * \text{capsizing} + 0.0034 * \text{docking condition} + 0.0014 * \text{hull/machinery damage})$	-151.98	8.7146	
Leaf node 4	$= \exp(-2.8647 - 29.2394 * \text{ship type} - 29.0668 * \text{collision} - 28.742 * \text{contact} - 29.1481 * \text{grouding} + 0.8331 * \text{operating time} + 2.7472 * \text{capsizing} + 0.7834 * \text{docking condition} + 0.6183 * \text{hull/machinery damage})$	-0.0843	26.137	
Leaf node 5	$= \exp(-0.1168 + 4.1858 * \text{ship type} + 0.3776 * \text{collision} + 0.1210 * \text{contact} + 0.0255 * \text{grouding} + 0.0773 * \text{operating time} + 0.0208 * \text{docking condition} + 0.1198 * \text{hull/machinery damage})$	-12.363	5.7315	

Leaf node 6	$= \exp(-4.0414 - 26.4778 * \text{ship type} + 0.752 * \text{collision} + 1.4094 * \text{contact} + 1.9326 * \text{grounding} - 27.7421 * \text{capsizing} - 2.4699 * \text{docking condition} + 0.2206 * \text{hull/machinery damage})$	0.4451	290.14	
Leaf node 7	$= \exp(-10.779 + 7.6772 * \text{collision} + 3.9873 * \text{contact} + 6.7325 * \text{grounding} + 1.5858 * \text{capsizing} + 0.3385 * \text{docking condition} + 4.7936 * \text{hull/machinery damage})$	1.685	225.05	



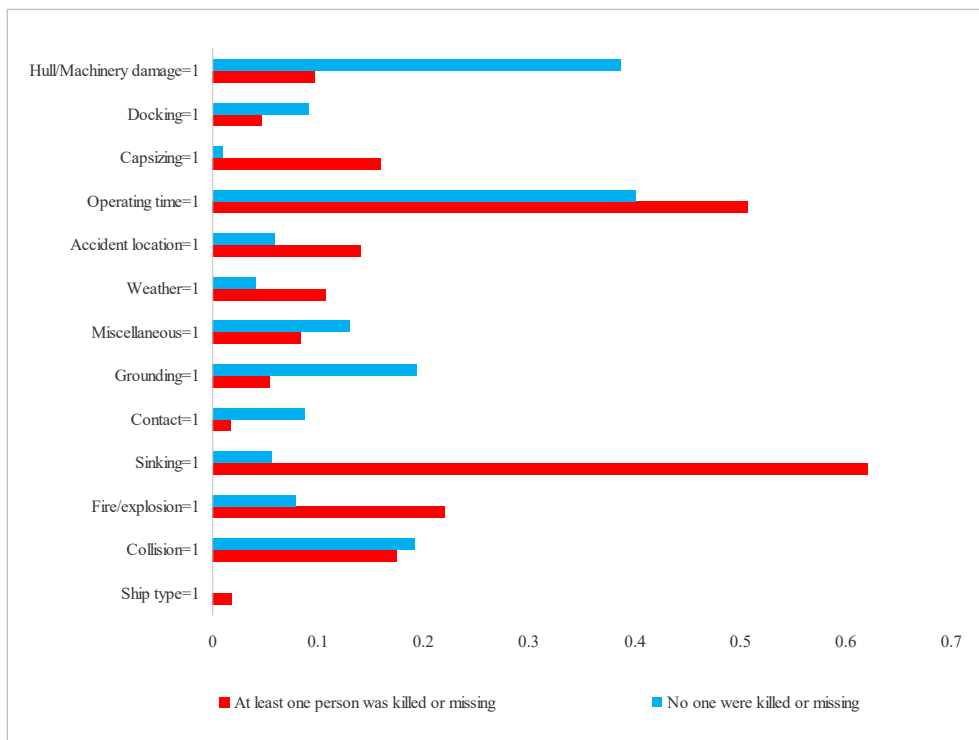


Fig.1. Comparison results of the mean of variables in two cases

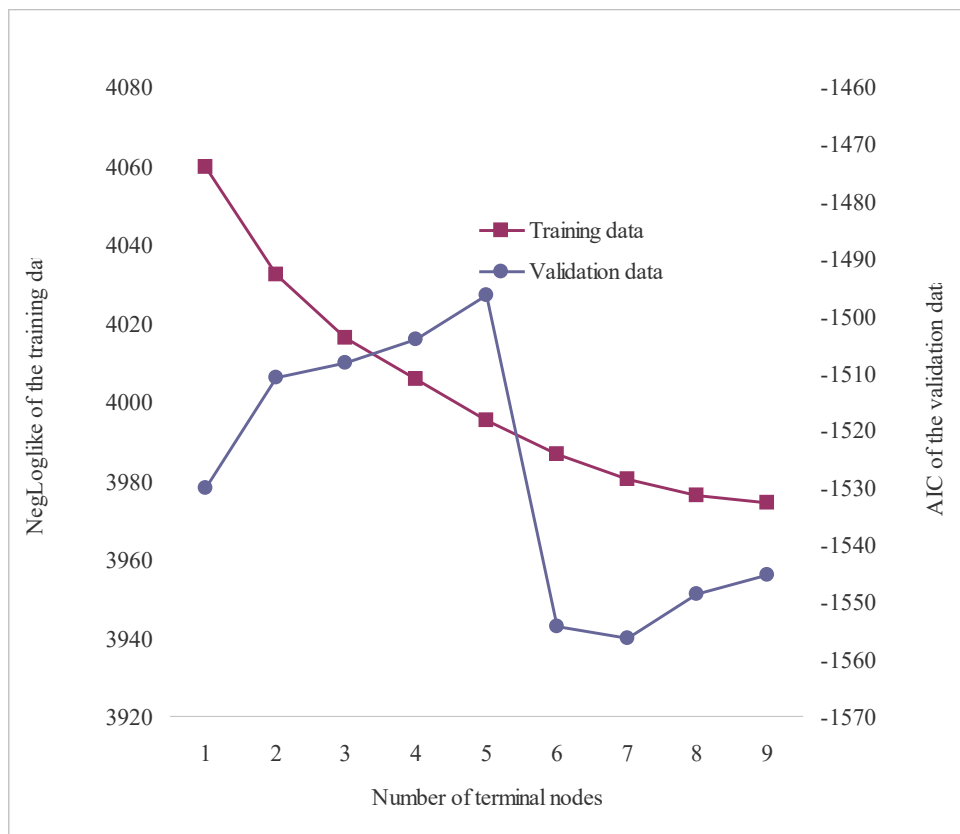


Fig. 2. Learning process of the maximum likelihood tree-based model

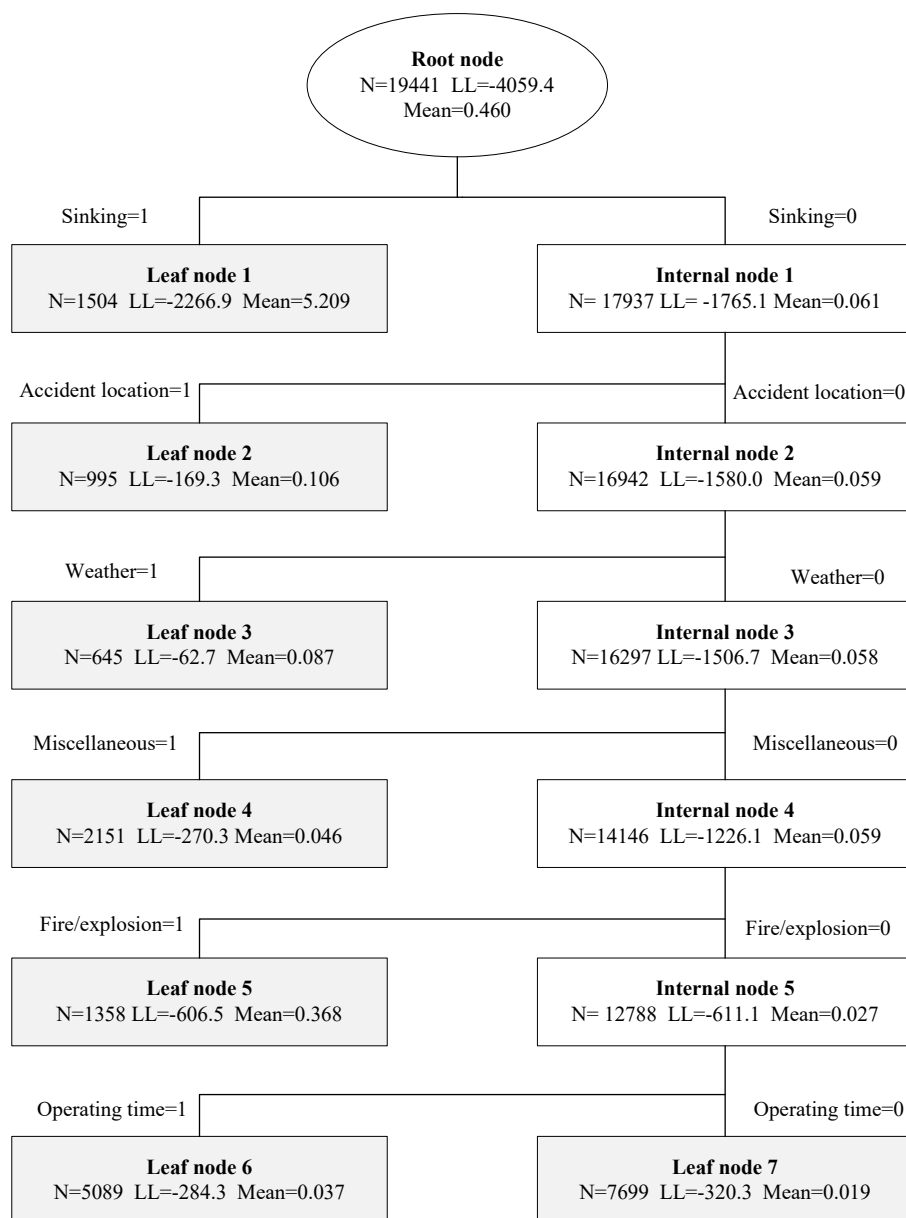


Fig.3 Maximum likelihood regression tree for Shipping Accident Casualties

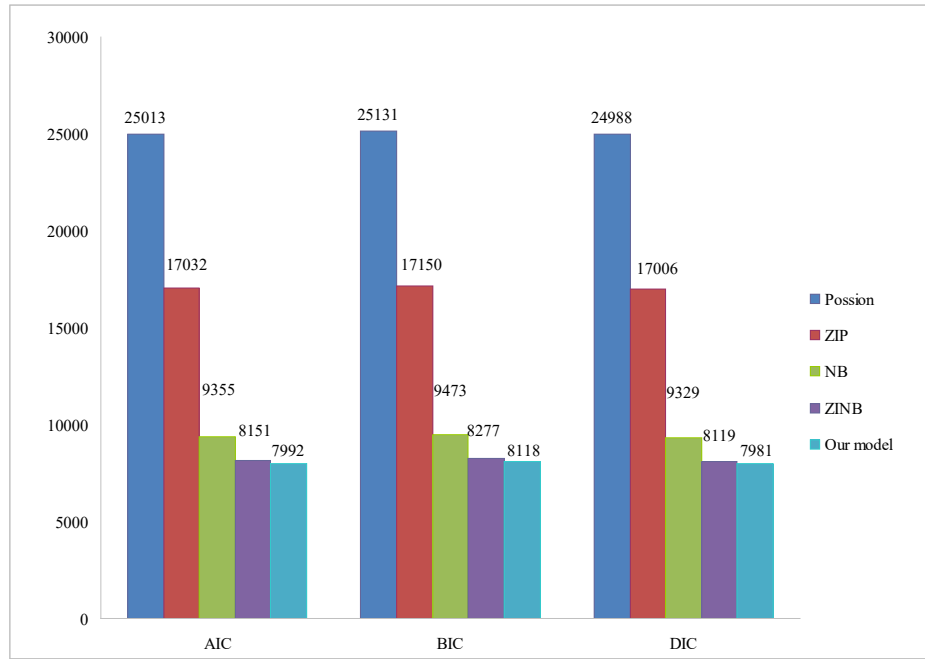


Fig. 4. Comparison results of different regression models