Journal: NHESS
Title: Analysing flood fatalities in Vietnam using national disaster database and tree-based methods
Author(s): Chinh Luu et al.
MS No.: nhess-2017-155
MS Type: Research article
Special Issue: Damage of natural hazards: assessment and mitigation

The objectives are (1) providing a comprehensive overview of flood fatalities in Vietnam, and (2) examining the damage-influencing variables (flood impacts) on flood fatalities. Accordingly, tree-based methods and DANA database were used.

Despite the objectives of the study which is interesting, the paper suffers from major shortcomings which prevent its publication. In what follows, major criticisms are first discussed, and then specific comments are supplied.

Major criticisms

I. With respect to the two objectives, neither a comprehensive overview nor examining of the damage-influencing variables (flood impacts) is analysed. To this aim, I expected the authors to analyse the significance of hazard and vulnerability influencing parameters on flood fatalities. Even, the importance of this matter has been stated by the authors in line 7-9 of page 1. However, the authors simply performed an exposure analysis which does not really need a tree-based model. In other words, as described in Table 2, the authors have only considered some damaged, exposed sectors with some quantities (not some ranges of variation), and damage-influencing variables are neglected entirely.

   On the other hand, It is obvious that a majority number of the people are usually trapped in houses or caught in roads (in the time of escaping). Then, a tree based model represents these two influenced sectors since the quantity of them are relatively high.

   All in all, exposure assessment would not be a damage influencing analysis when the authors have neglected the vulnerability (physical, social or systemic) of affected objects and the hazard intensity parameters.

II. To my understanding there are some methodological inconsistencies in the paper:

   1. For implementing the above comment, an event-based analysis is needed. The authors have not considered a scientific approach for distributing the cumulated number of fatalities (between 1989 and 2015) in each year or each region. In Page 1 line 15 and page 5 line 13, the authors have simply divided 14,927 fatalities to 27 years (between 1989 and 2018) and reached to 553 numbers of casualties which is not scientifically sound. Also, this number is not compatible with Figure 5 information. For that, they needed to

calculate the Average Annual Damage (AAD) based on the probability (return period) of each event and the extent of losses of that.

2. It is not clear that how the authors have calculated the information of Figure 7 (annual average of losses), without assessing the AAD explained earlier. If it again comes from a simple division, it is not scientifically correct. Also, this figure suffers from several problems (e.g. incomplete caption; wrong axis label "SSC instead of SCC"; unnatural distribution of standard errors; incorrect length of bar charts for SE and NE which are equal to 2 and 5 respectively), and its information is not compatible with Fig.5.
3. Authors should describe their methodological steps chronologically to avoid confusion. There are many examples of the information which are represented in an inappropriate and unrelated section or repeated several times.
4. Discussion and Conclusion parts, as the most important sections of each study, should be rewritten entity. In the presented format, the discussion part is a repetition of previous materials, and the conclusion part does not represent any outcome, finding, or contribution.
5. The main application of out-of-bag (OOB) data, exploring the feature importance, is not used in this study. Then, what was the advantage of using this technique besides cross-validation approach?
6. It is obvious, and it has been shown before that Bagging and Random Forests represent more stable and accurate outcome. Consequently, I expected that the authors use MSE and cross-validation test for choosing a tree with the most appropriate number of nodes. However, it is not clear that why the authors have selected a tree with seven leaves (shown in Figure 9), while its error in Figure 10 is more than other sizes of trees.
7. In Page 10 line 26, the authors have mentioned that "The results showed that the tree-based models were validated and applicable". In this study validation of the model is not discussed and the study includes only some error estimations.
8. Results of Figure 2 are not compatible with the findings of the grown trees. In Figure 2, most of the sectors have a high (more than 0.5) correlation coefficient with the number of fatalities. This matter reflects the issue of the inaccurate inputs explained in section I.
9. Based on the explanations of Table 2, the relationship of some categories like "irrigation, telecommunication, electricity and material categories" with the number of fatalities is not clear. Are they related?!

III. Transformation of the empirical data needs more clarification.

Specific comments

- Page 1, Line 10: "The number of fatalities is the most important indicator in flood risk assessment." Do you have any reference for justifying this statement?
- It is not clear enough that how we can use Figure 9 for predicting the number of fatalities. In this Figure, I do not understand that what "the small number on the top

of each rectangle" is for, and what does the number at the upper line of each rectangle mean? All of them need some explanations. Also, why the summation of the percentages of the end nodes (leaves) is 102%?!!!

- In Figure 9, there are considerable differences in the number of data related to each end node (625 in the first node and 44 in the fifth node). It shows that the tree is not grown soundly and the pruning technique is not hired perfectly.
- Page 6, Line 22: "The cross-validation procedure was undertaken to ensure that the parameter estimation and model generation of regression trees, bagging, random forests and boosting are **entirely independent of the test data**." Cross-validation test does not generate an independent dataset. The authors need to use data collected from another event if they are interested in testing the transferability and applicability of the model.
- Page 8 Line 5: "*Cross-validation method was operated to select the most accurate tree-based technique and to* ***check the data validation***". Cross-validation is a technique for validation of model compared to the real damage data. It is not an approach for checking the validation of the data.
- There is some information that does not have any contribution to the objectives of this study and the authors need to delete them such as Page 11 L 8-12; total injured people in Figure 5; Figure 14; Figure 15; and Table 1.
- Figures need some improvement (e.g. caption of Figure 2 needs more explanations about variables, presentation of Figure 3, SSC should be replaced by SCC in axis label of Figure 6, and NE should be replaced by NW in Figure 6 caption)
- Page 6 Line 7: "*Tree-based methods are supervised learning algorithms. The methodology of these methods is based on classification and regression tree (CART) of Breiman et al. (1984).*" Classification and regression tree (CART) is not the only algorithm of tree-based methods. Trees can be grown based on different algorithms such as RETIS (Karalic & Cestnik, 1991), M5 (Quinlan, 1992), REPTree, Random Tree, or CART (Breiman et. al., 1984).