Analyzing Flood Fatalities in Vietnam Using Statistical Learning Approach and National Disaster Database

Chinh Luu and Jason von Meding

Abstract

Floods and storms have had a severe impact on the people of Vietnam over many years, particularly regarding an unacceptably high death toll. However, it still lacks studies on flood-related fatalities in Vietnam. This research aims to explore flood fatalities on a national scale and analyze damage-influencing attributes related to flood fatalities using the national disaster database of Vietnam and statistical learning approach. Records covering 27 years from 1989 to 2015 indicate at least 14,927 flood mortalities in Vietnam. The analysis results of statistical learning methods show that housing impact factor has the most considerable influence on flood fatalities. The results can provide implications for housing policies for the poor in flood-prone areas. The objective of reduction in mortality in disasters is under Goal 11 of Sustainable Development Goals.

Keywords

Flood fatalities · Flood damage · Multiple linear regression · Random forest

C. Luu (✉) · J. von Meding
University of Newcastle, Newcastle, Australia
e-mail: ThiDieuChinh.Luu@uon.edu.au; Jason.VonMeding@newcastle.edu.au

1 Introduction

A flood is a very complex phenomenon, that is, the links of natural environment, people, and social system (Slobodan 2012). Global flood exposure and flood frequency are projected to increase especially in many low-latitude regions in Asia and Africa (Hirabayashi et al. 2013). The global disaster database, EM-DAT (Fig. 1), shows that the frequency of flood and storm events is correlated with flood fatalities.

Disaster data collection and analysis are increasingly prominent (UNISDR 2015). The analysis could provide necessary information for policy-setting and decision-making process in disaster risk reduction (IRDR 2014). Various disaster databases are available on national and global scales, which are summarized and listed in by Grasso and Dilley (2013) and Simpson et al. (2014). The exploration of these databases could contribute to better understanding of disaster risk, which is the first priority in the Sendai Framework (UNISDR 2015).

The Centre for Research on the Epidemiology of Disasters supported Vietnam and other Asian countries to develop their national disaster databases (Below et al. 2010). The Vietnamese national disaster database is Damage Assessment and Needs Analysis or DANA. The Central Committee for Flood and Storm Control of

A. Asgary (ed.), *Resettlement Challenges for Displaced Populations and Refugees*,
Sustainable Development Goals Series, https://doi.org/10.1007/978-3-319-92498-4_15

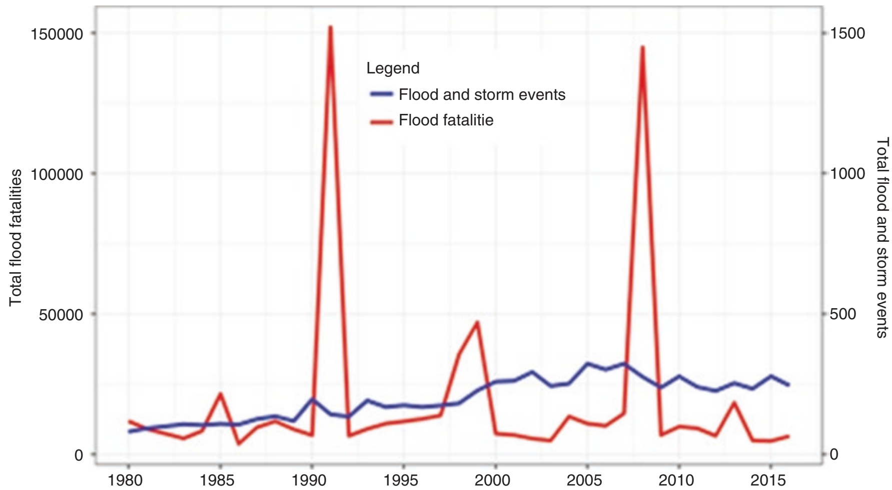

Fig. 1 Flood fatalities and flood events between 1980 and 2016 at global scale (compiled from EM-DAT database http://emdat.be)

Vietnam had been jointly developing the DANA database (MARD 2006; Hughey et al. 2011).

Analysis of flood damage data could provide crucial information to decision-makers in the field of flood risk management and adaptation planning (Merz et al. 2010). Although previous research investigated DANA, such as Hughey et al. (2011) and Nhu et al. (2011), there has not been any investigation related to flood fatalities.

There are many studies on flood-related deaths in developed countries (e.g., Coates 1999; Ashley and Ashley 2008; FitzGerald et al. 2010; Di Mauro and de Bruijn 2012; Sharif et al. 2015) while only a few studies in developing countries (e.g., Paul and Mahmood 2016). The presence of studies on flood-related mortalities in developing economies is minimal despite the fact that the mortality rate in these countries is significant. There is no systematical research on flood-related fatalities in Vietnam.

Studies on flood fatalities often have two main approaches: (1) developing predictive models (Di Mauro et al. 2012; Jonkman and Vrijling 2008; Zhai et al. 2006; Di Mauro and de Bruijn 2012; Jonkman et al. 2002) and (2) analyzing the causes of fatalities (Ashley and Ashley 2008; Paul and Mahmood 2016; Sharif et al. 2015; Coates 1999; Jonkman et al. 2009; Jonkman and Kelman 2005).

Statistical learning is the combination of classical statistics and computer science (James et al. 2013a). The applications of statistical learning are increasingly applied in genetics, medical science, business, and flood risk management field (Merz et al. 2013; Hasanzadeh Nafari et al. 2016). However, there is a lack of research on the application of statistical learning techniques to analyze the relationship between flood damage attributes on flood fatalities.

The present study aims to explore flood fatalities on a national scale and analyze damage-influencing attributes related to flood fatalities using the national disaster database of Vietnam and statistical learning approach.

2 Disaster Database

The Central Committee for Flood and Storm Control of Vietnam has developed the national disaster database since 1989 through DANA database. Flood damage data is collected via one template including 12 categories with many flood impact indicators, for example, fatalities, agriculture impacts, housing damages, infrastructure damages, and economic loss. The database provides hydrometeorological disaster damage information at the national level (Hughey et al.

2011). It only stores direct losses on monetary of reconstruction and recovery of damaged property and infrastructure, and does not report indirect losses, for instance, business disruption and production interruption (Wang et al. 2010).

3 Flood Fatalities in Vietnam

EM-DAT defined death or fatality including person confirmed dead and missing person presumed dead (Below et al. 2010). The loss of human life is considered one of the most critical indicators in assessing flood risk (Maaskant et al. 2009). The flood mortalities are very low in developed countries such as Australia (Mojtahedi and Oo 2016) and Scotland (Crichton 2004). Meanwhile, the flood fatalities in Vietnam are unacceptably high, with at least 14,972 flood mortalities between 1989 and 2015 (Fig. 2).

Spatial patterns of flood fatalities by provinces in Vietnam are generated using the compiled flood damage data from DANA database and ArcGIS 10.1 software in Fig. 2. Quang Nam, Ca Mau, and Quang Ngai provinces have the highest death toll of more than 800 people during the observation period. The second highest flood fatality level is from 401 to 800 people in nine provinces including Nghe An, Thanh Hoa, Da Nang, Thua Thien Hue, Binh Dinh, Khanh Hoa, Kien Giang, Dong Thap, and An Giang.

4 Statistical Learning Methods

Based on the DANA database, this study uses the two following statistical learning methods, multiple linear regression model and random forest, to measure the relative influence of flood damage attributes on fatalities. We present the detail underpinning methodology of these methods in this section.

4.1 Multiple Linear Regression Model

Multiple linear regression technique is employed to analyze the independent factors relating to the flood fatalities, which is set as a dependent variable or an outcome. After that, LMG method is used to obtain the relative importance of flood fatality determinants based on the regression model.

Multiple linear regression model aims to find an equation to describe the relationship between *X* (independent variables) and *Y* (dependent variable). A multiple linear regression model (population) with *p* independent variables is described as in Eq. (1) to examine the linear relationship between one dependent (*Y*) and two or more dependent variables (x_i).

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon \quad (1)$$

where

β_0: intercept
$\beta_1, \beta_2, \ldots, \beta_p$: regression coefficients
Y: dependent variable
x_{pi}: independent variables
ε: residuals

Weights of independent variables can be generated based on relative importance of variables (Hair Jr. et al. 2014). Bi (2012) reviewed new methods for generating relative importance from regression methods and recommended the LMG method for raw data. LMG indicator, which was proposed by Lindeman, Merend, and Gold in 1980 (Lindeman et al. 1980), is applied to assess the relative importance of variables of the multiple linear regression model in this study. The explanation of LMG is as follows (Lindeman et al. 1980; Bi 2012):

$$\mathrm{LMG}\left(x_k\right) = \frac{1}{p!} \sum_{r\mathrm{permutation}} \mathrm{seq}R^2\left(\left\{x_k\right\} | r\right) \quad (2)$$

where $r = 1, 2, \ldots, p!$ and $seqR^2(\{x_k\}|r)$ denotes sequential sum of squares for the regressor x_k in the ordering of regressors in *r*-th permutation.

4.2 Random Forest

Another method, random forest, for regression is applied to the same database. Random forest algorithm, for both regression and classification,

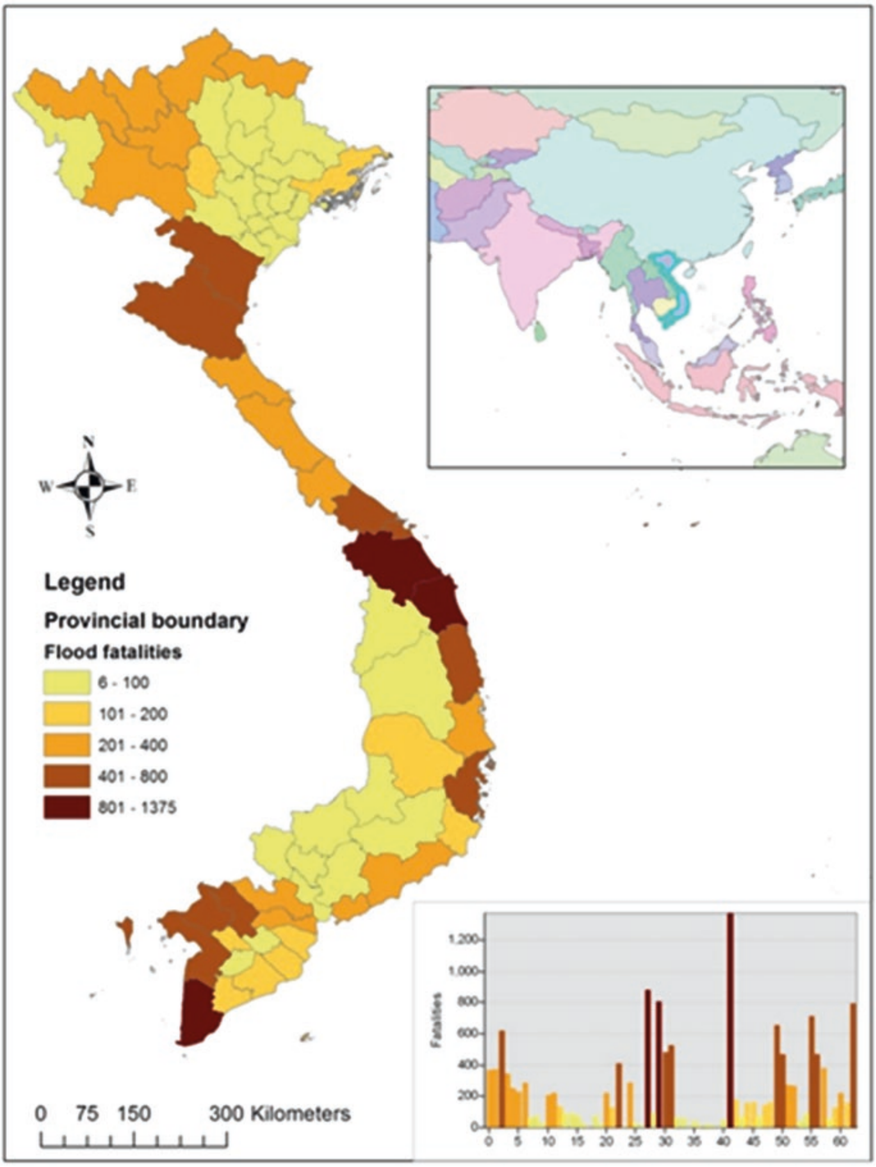

Fig. 2 Spatial patterns of flood fatalities by provinces in Vietnam from 1989 to 2015

is a panacea for all data science problems. This method constructs multitude of decision trees and selects the best as the final result which can be used to build predictive models. The random forest algorithm can be presented in the following steps (Liaw and Wiener 2002):

1. Draw n_{tree} bootstrap sample from the original dataset. A sample of these n_{tree} is taken at random with replacement.
2. For each of these samples, develop an unpruned regression tree: randomly sampling m_{try} of predictors and selecting the best split from these variables at each node.
3. Predict new dataset by aggregating the predictions of the trees (average for regression).

Random forest for regression is constructed by growing trees depended on a random vector such that the predicted tree takes numerical values as opposed to class labels. C. Strobl et al. (2008) suggested that the conditional importance for random forest should be applied for the highly correlated predictor variables when this tool reflects the actual impact of each variable. The conditional importance technique is chosen for the dataset of this study.

5 Application

5.1 Multiple Linear Regression Analysis

Multiple linear regression model is applied to flood damage data of DANA. The collected dataset includes 27 samples in 63 provinces (27 years from 1989 to 2015). One year is considered an observation, so there are 1701 observations. Flood fatality is set as a dependent variable. Flood damage attributes are set as independent variables from X1 to X10. Data transformation with logarithm function is applied to both outcome and independent variables for better fitting the normal distribution due to large variation in the flood damage data (Zhou et al. 2017).

Cross-validation methods are used to validate the multiple linear regression model. Cross-validation aims to evaluate whether a model has the good predictive ability for a new dataset. K-fold cross-validation and leave-one-out cross-validation (LOOCV) techniques can provide the best cross-validation estimate (James et al. 2013b). The k-fold cross-validation and LOOCV results are generated using "boot" package (Canty and Ripley 2016) in R statistical software (R Core Team 2016). The results in Fig. 3 show that mean square errors of both LOOCV and k-fold CV models with a degree of polynomial from 2 to 10 are low values and approximately the same. Therefore, the model is validated.

We use R statistical software to run the multiple linear regression model with the transformed data. The model has adjusted R-squared of 0.601 and residual standard error of 0.82. After that, we run "*relaimpo*" package (Grömping 2006) in R (R Core Team 2016) to generate the weights or relative importance of attributes. The weights of attributes or independent variables are generated based on LMG indicator as in Eq. (2). The result is shown in Table 1.

5.2 Conditional Importance for Random Forest

Random forest algorithm aimed to find the relative influence of independent factors (flood impacts from X1 to X10) to a dependent outcome (flood fatalities). The conditional importance for random forests was conducted using '*party*' package (Hothorn et al. 2006; Strobl et al. 2007, 2008) in R statistical software. The result is shown in Fig. 4. Housing impact factor (variable X1) has the highest influence on fatalities of this model.

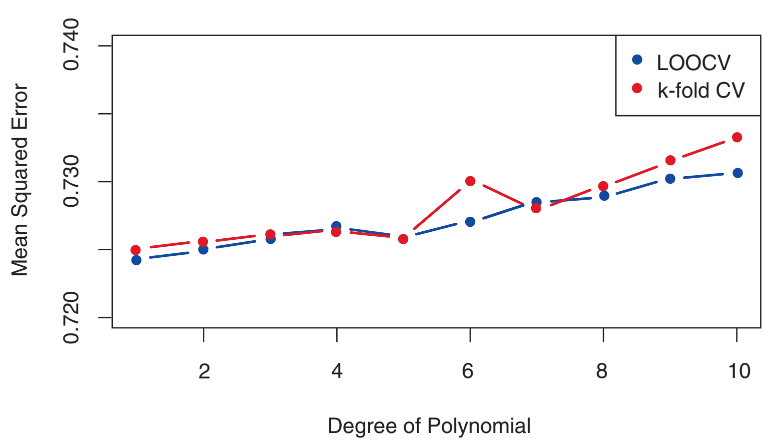

Fig. 3 Error curves of tenfold CV and LOOCV with different random splits and ten times repeated of the two cross-validation methods

Table 1 The relative importance of flood fatality determinants based on multiple linear regression model and LMG indicator after normalization

Determinants	Explanatory	Unit	Symbol	Relative importance	P value
Housing	Number of houses damaged		X1	0.226	<0.0001
Education	Number of classrooms damaged		X2	0.141	<0.0001
Healthcare	Number of clinics damaged		X3	0.085	0.0013
Agriculture	Area of paddy inundated and farm produce damaged	ha	X4	0.118	0.007
Irrigation	Volume of earth and rock eroded, washed away, and redeposited (of dikes, canals, and reservoirs)	m^3	X5	0.104	0.0122
Transportation	Volume of earth and rock eroded, washed away, and redeposited (of roads and highways)	m^3	X6	0.114	<0.0001
Fisheries	Area of fish and shrimp feeding damaged	ha	X7	0.099	<0.0001
Telecommunication	Number of telephone poles broken		X8	0.027	0.6374
Electricity	Number of high voltage electrical towers and electrical distribution poles collapsed		X9	0.055	0.7024
Materials	Volume of cement damaged, salt lost, clinker wetted, coal drifted	ton	X10	0.03	0.0059

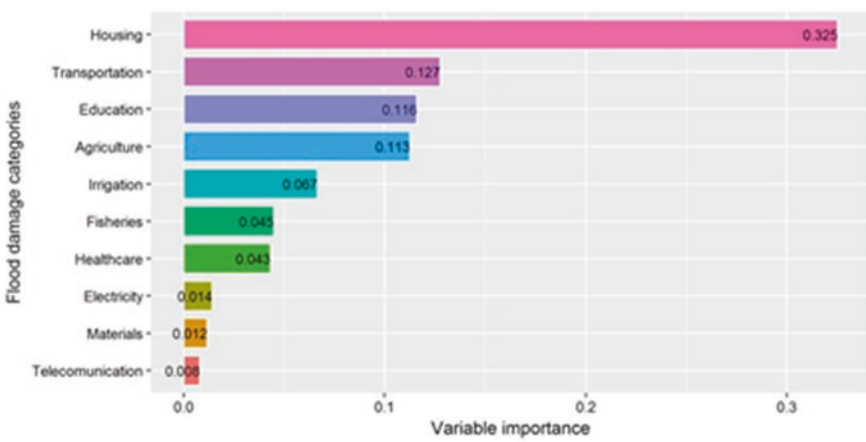

Fig. 4 Variable importance scores for the ten predictors in regression tree model by conditional variable importance for random forest

6 Discussion and Conclusion

The interrelation between sustainable development and disaster risk reduction is recognized in Sendai Framework (UNISDR 2015). Part of that, the objective of reduction in mortality in disasters is stated under Goal 11 of Sustainable Development Goals (UN 2015). A summary of flood fatalities in Vietnam and the investigation of damage-influencing attributes on flood fatalities in this study can support future efforts to mitigate fatality in flood disaster and have implications for flood risk management activities.

The recorded damage data suggests that flood hazards have had severe impacts on Vietnamese people, with over 14,927 dead and missing or at least 553 flood mortalities per year between 1989 and 2015 (Fig. 2). The susceptibility has been continuing with 264 flood fatalities in 2016 (UN Country Team in Vietnam 2016). The flood fatalities are unexpectedly high due to the passive response activities of local authorities and residents to flood and storm events. Therefore, flood risk management needs focusing on a proactive approach or mitigation and preparedness activities.

The analysis results from multiple linear regression analysis (Table 1) and conditional importance for random forest (Fig. 4) show that housing damage factor has the most significant influence on flood mortalities. The more houses damaged and flooded, the more fatalities are. The people who lived in a house that is destroyed or flooded are more likely to be killed. The houses of farmers (over 70% of Vietnamese are farmers) are mostly one-story and in poor conditions. They are not strong enough to withstand storms or floods, and they provide no room to escape in high flood depth cases.

Floods and storms affected low-income communities disproportionately, especially in riverine and coastal areas in Vietnam. The rural poor are particularly vulnerable to flood events. The affected households often received extremely limited financial support from the local government, which was stipulated in Decree 67/2007/ND-CP of the government. The poverty rate, as is common on a global scale, is invariably linked to disaster impact.

The results may provide information on community awareness and safety regulations. In particular, our results can be used to recommend government policies that focus on supporting the poor in upgrading their houses in flood-prone areas to mitigate flood fatalities.

This study used the data available in the national disaster database for the analysis. It is, therefore, limited to analyzing the relative influence of damage attributes on fatalities in Vietnam. We call for more detailed studies on flood fatalities such as predictive models and the causes of flood fatalities. However, it is required the disaster loss database documented details on the causes, age, and gender of flood fatalities.

In conclusion, the present study proposes an approach to investigate the damage-influencing attributes related to flood fatalities using statistical learning techniques and a national disaster database. The results show that housing damage factor has the most significant influence on flood mortalities in Vietnam. Our research provides a better understanding of flood fatalities in Vietnam by analyzing and reporting on flood mortalities using statistical learning approach and a national disaster database. The output can produce a reference for the decision-making process in flood risk management in Vietnam.

References

Ashley, S. T., & Ashley, W. S. (2008). Flood fatalities in the United States. *Journal of Applied Meteorology and Climatology, 47*(3), 805–818. https://doi.org/10.1175/2007jamc1611.1.

Below, R., Vos, F., & Guha-Sapir, D. (2010). *Moving towards harmonization of disaster data: A study of six Asian databases*. Brussels: Centre for Research on the Epidemiology of Disasters.

Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies, 27*(2), 87–101. https://doi.org/10.1111/j.1745-459X.2012.00370.x.

Canty, A., & Ripley, B. (2016). Boot: Bootstrap R (S-plus) functions. *R package version, 1*, 3–18.

Coates, L. (1999). Flood fatalities in Australia, 1788-1996. *Australian Geographer, 30*(3), 391–408. https://doi.org/10.1080/00049189993657.

Crichton, D. (2004). *Flood risks in the former Grampian Region since Devolution* (A Research Report for WWF by David Crichton, Vol. 16 November 2015): WWF Scotland.

Di Mauro, M., & de Bruijn, K. M. (2012). Application and validation of mortality functions to assess the consequences of flooding to people. *Journal of Flood Risk Management, 5*(2), 92–110. https://doi.org/10.1111/j.1753-318X.2011.01131.x.

Di Mauro, M., De Bruijn, K. M., & Meloni, M. (2012). Quantitative methods for estimating flood fatalities: Towards the introduction of loss-of-life estimation in the assessment of flood risk. *Natural Hazards, 63*(2), 1083–1113. https://doi.org/10.1007/s11069-012-0207-4.

FitzGerald, G., Du, W., Jamal, A., Clark, M., & Hou, X. Y. (2010). Flood fatalities in contemporary Australia (1997-2008). *Emergency Medicine Australasia, 22*(2), 180–186. https://doi.org/10.1111/j.1742-6723.2010.01284.x.

Grasso, V. F., & Dilley, M. (2013). *A comparative review of country-level and regional disaster loss and damage databases*. New York: United Nations Development Programme Bureau for Crisis Prevention and Recovery.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*(1), 1–27. https://doi.org/10.18637/jss.v017.i01.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Harlow: Pearson Education Limited.

Hasanzadeh Nafari, R., Ngo, T., & Mendis, P. (2016). An assessment of the effectiveness of tree-based models for multi-variate flood damage assessment in Australia. *Water, 8*(7), 282. https://doi.org/10.3390/w8070282.

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change, 3*(9), 816–821. https://doi.org/10.1038/nclimate1911.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics, 7*(3), 355–373.

Hughey, E., Bell, H., & Chatman, M. (2011). Who needs what? A case study of post-disaster damage and needs assessment (DANA) in Vietnam. *Risk, Hazards & Crisis in Public Policy, 2*(4), 1–24. https://doi.org/10.2202/1944-4079.1097.

IRDR (2014). *IRDR Peril Classification and Hazard Glossary* (DATA Project Report No. 1). Beijing: Integrated Research on Disaster Risk (IRDR).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). Introduction. In *An introduction to statistical learning: With applications in R* (pp. 1–14). New York: Springer New York.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). Resampling methods. In *An introduction to statistical learning: With applications in R* (pp. 175–201). New York: Springer.

Jonkman, S. N., & Kelman, I. (2005). An analysis of the causes and circumstances of flood disaster deaths. *Disasters, 29*(1), 75–97. https://doi.org/10.1111/j.0361-3666.2005.00275.x.

Jonkman, S. N., Maaskant, B., Boyd, E., & Levitan, M. L. (2009). Loss of life caused by the flooding of New Orleans after hurricane Katrina: Analysis of the relationship between flood characteristics and mortality. *Risk Analysis, 29*(5), 676–698. https://doi.org/10.1111/j.1539-6924.2008.01190.x.

Jonkman, S. N., van Gelder, P. H. A. J. M., & Vrijling, J. K. (2002). Loss of life models for sea and river floods. *Flood defence, 1*, 196–206.

Jonkman, S. N., & Vrijling, J. K. (2008). Loss of life due to floods. *Journal of Flood Risk Management, 1*(1), 43–56. https://doi.org/10.1111/j.1753-318X.2008.00006.x.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18–22.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview: Scott, Foresman and Company.

Maaskant, B., Jonkman, S. N., & Bouwer, L. M. (2009). Future risk of flooding: An analysis of changes in potential loss of life in South Holland (the Netherlands). *Environmental Science & Policy, 12*(2), 157–169. https://doi.org/10.1016/j.envsci.2008.11.004.

MARD (2006). Guideline on natural disaster damage and needs assessment. http://www.ngocentre.org.vn/webfm_send/1533. Accessed 14 Oct 2016.

Merz, B., Kreibich, H., & Lall, U. (2013). Multi-variate flood damage assessment: A tree-based data-mining approach. *Natural Hazards and Earth System Science, 13*(1), 53–64. https://doi.org/10.5194/nhess-13-53-2013.

Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article "assessment of economic flood damage". *Natural Hazards and Earth System Science, 10*(8), 1697–1724. https://doi.org/10.5194/nhess-10-1697-2010.

Mojtahedi, S. M. H., & Oo, B. L. (2016). Coastal buildings and infrastructure flood risk analysis using multi-attribute decision-making. *Journal of Flood Risk Management, 9*(1), 87–96. https://doi.org/10.1111/jfr3.12120.

Nhu, O. L., Thuy, N. T. T., Wilderspin, I., & Coulier, M. (2011). *A preliminary analysis of flood and storm disaster data in Vietnam* (Global Assessment Report on Disaster Risk Reduction, Vol. 30 Sept 2016): United Nations Development Programme Vietnam.

Paul, B. K., & Mahmood, S. (2016). Selected physical parameters as determinants of flood fatalities in Bangladesh, 1972–2013. *Natural Hazards*. https://doi.org/10.1007/s11069-016-2384-z.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sharif, H. O., Jackson, T. L., Hossain, M. M., & Zane, D. (2015). Analysis of flood fatalities in Texas. *Natural Hazards Review, 16*(1), 04014016. https://doi.org/10.1061/(asce)nh.1527-6996.0000145.

Simpson, A., Murnane, R., Saito, K., Phillips, E., Reid, R., & Himmelfarb, A. (2014). *Understanding risk in an evolving world: Emerging best practices in natural disaster risk assessment*. Washington DC: Global Facility for Disaster Reduction and Recovery, the World Bank.

Slobodan, P. S. C. (2012). *Floods in a changing climate: Risk management (part of international hydrology series)*. Cambridge, UK: Cambridge University Press.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*, 307. https://doi.org/10.1186/1471-2105-9-307.

UN. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development (a/RES/70/1)*. New York: United Nations.

UN Country Team in Vietnam. (2016). Vietnam: Situation reports. https://reliefweb.int/organization/unct-vietnam. Accessed 16 Oct 2017.

UNISDR. (2015). *Sendai framework for disaster risk reduction 2015–2030*. Geneva: United Nations Office for Disaster Risk Reduction (UNISDR).

Wang, X., Mahul, O., & Stutley, C. (2010). *Weathering the storm: Options for disaster risk financing in Vietnam*. Washington, DC: Global Facility for Disaster Reduction and Recovery, the World Bank.

Zhai, G., Fukuzono, T., & Ikeda, S. (2006). An empirical model of fatalities and injuries due to floods in Japan. *JAWRA Journal of the American Water Resources Association, 42*(4), 863–875. https://doi.org/10.1111/j.1752-1688.2006.tb04500.x.

Zhou, Q., Leng, G., & Feng, L. (2017). Predictability of state-level flood damage in the conterminous United States: The role of hazard, exposure and vulnerability. *Scientific Reports, 7*(1), 1–11. https://doi.org/10.1038/s41598-017-05773-4.