Journal of Hydrology 527 (2015) 1130–1141

Contents lists available at ScienceDirect

Journal of Hydrology

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

Flood hazard risk assessment model based on random forest

Zhaoli Wang [a], Chengguang Lai [a,b,c,*], Xiaohong Chen [b,c], Bing Yang [b,c], Shiwei Zhao [a], Xiaoyan Bai [d]

[a] School of Civil and Transportation Engineering, South China University of Technology, Guangzhou 510641, China
[b] Department of Water Resource and Environment, Geography and Planning School of Sun Yat-Sen University, Guangzhou 510275, China
[c] Key Laboratory of Water Cycle and Water Security in Southern China of Guangdong Higher Education Institutes, Guangzhou 510275, China
[d] Department of Environmental Engineering, School of Environmental Science and Engineering, Guangdong University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 15 May 2015
Accepted 4 June 2015
Available online 12 June 2015
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Flood disaster
Risk assessment
Random forest
Classification trees
Dongjiang River Basin

SUMMARY

Floods, natural disasters that occur worldwide, have become more and more frequent in recent decades. Flooding is often unavoidable and unexpected; however, it can be controlled through appropriate measures to minimize losses and damage. Flood hazard risk assessment, a holistic approach that involves numerous evaluation indices in river catchments, is an increasingly effective and sustainable practice, but the complicated, non-linear relationship between evaluation indices and risk levels pose a significant challenge to accurate assessment. An intelligent learning machine called random forest (RF) can run efficiently on large databases and provide estimates regarding the importance of specific variables in the classification. This lends RF a considerable advantage in solving the non-linear problems inherent to risk assessment, as well as estimating the importance degree of each index. As such, in this study, an assessment model based on RF was adopted to evaluate regional flood hazard risk. The proposed flood hazard risk assessment method was implemented in Dongjiang River Basin, China. Eleven risk indices were selected and five thousand samples were created for training and testing. The support vector machine (SVM) was used for risk assessment as a comparison, as well as an analysis of index importance degree. Results show that (1) both the training and testing error rate of RF can be reduced by increasing the sample size and the number of classification trees, (2) the higher and highest-risk zones occupy approximately 19.09% of the total, and are primarily located in Baoan, Longgang, Huiyang, Huidong, the western area of Boluo, and the southern part of Dongguang, and (3) the indices maximum three-day precipitation (M3PD), runoff depth (RD), typhoon frequency (TF), digital elevation model (DEM), and topographic wetness index (TWI) are the top five most important out of the eleven risk indices, occupying 71.11% of the total risk; but normalized difference vegetation index (NDVI), stream power index (SPI), soil texture (ST), distance to the river (DR), slope (SL), and land use pattern (LUP) indices are less consequential, at only 28.89%. This study shows the potential to provide a novel and highly successful approach to flood hazard risk assessment. Evaluation results provide a reference for flood risk management, prevention, and reduction of natural disasters in the study basin.

1. Introduction

Floods are generally considered to be the most common natural disaster worldwide (Stefanidis and Stathis, 2013). Over the past several decades, flooding has caused significant economic damage and loss of life in every corner of the globe (Gaume et al., 2009), and despite substantial measures that have been enacted to prevent floods, resultant loss of human life and property persist at high levels (Alexander, 1993; Cui et al., 2002). According to shocking statistic data from 1900 to 2013, floods have caused approximately 7 million deaths and led to more than US $600 billion in losses (Disaster Profiles, 2013). Worse still, flooding events are expected to increase in frequency and intensity in coming years due to rising sea levels and more frequent extreme precipitation events (Ramin and McMichael, 2009; IPCC, 2007; Stijn et al., 2013; Jonathan et al., 2013). Within this context, defining optimum strategies for appropriate flood management is essential (Ballesteros-Cánovas et al., 2013).

The importance of flood hazard risk assessment in ensuring the healthy and sustainable development of human society cannot be overemphasized. Flood hazard risk, naturally, is usually measured by the probability that a flood will occur. A flood event generally results from a specific situation – high intensity rainfall plus

* Corresponding author at: Department of Water Resource and Environment, Geography and Planning School of Sun Yat-Sen University, Guangzhou 510275, China. Tel.: +86 013763316315.

E-mail address: laichengguang@foxmail.com (C. Lai).

http://dx.doi.org/10.1016/j.jhydrol.2015.06.008
0022-1694/

adverse geographical environment – but adequate information regarding these adverse factors, as well as the relationship between them, is yet lacking. Flood hazard risk assessment, a qualitative or semi-quantitative method which considers the combined influence of disaster-inducing factors and hazard-inducing environments, is considered an effective means of solving this problem (Stephane et al., 2013), and therefore has been widely applied to flood insurance, floodplain management, disaster warning systems, and evacuation planning, proving to be an important, scientific reference for flood disaster risk management and relevant decision-making (Zou et al., 2013).

The purpose of flood hazard risk assessment is obtaining accurate risk levels. The main difficulty associated with this process is the multi-variable and non-linear relationship between indices and risk levels. In effort to remedy this, numerous systematic methods such as the analytic hierarchy process (AHP) (Fernández and Lutz, 2010; Stefanidis and Stathis, 2013; Yang et al., 2013), set pair analysis (SPA) (Zou et al., 2013; Guo et al., 2014), and fuzzy comprehensive evaluation (FCE) (Jiang et al., 2009; Li, 2013; Lai et al., 2015) have been applied to this field, with mixed results. Although these methods are acceptable for analyzing flood risk, there remain uncertainties and vulnerabilities due to the complexities and inconveniences found in their practical application. For example, index weights must be calculated by manual intervention before the AHP method can be used, resulting in high subjectivity. Additionally, there is more qualitative than quantitative data in AHP evaluation; consequently, results are less than satisfactory (Stefanidis and Stathis, 2013). Evaluation results of SPA and FCE are significantly influenced by index weight, and the necessary calculations are extremely complex (Feng and Luo, 2009; Zou et al., 2013). Alongside the development of artificial intelligence technology, a trend of applying intelligence algorithms to flood hazard risk assessment has emerged; for instance, use of support vector machines (SVMs) (Yeh et al., 2010; Deng and Zhou, 2013), decision trees (DTs) (Tingsanchali and Karim, 2010; Merz et al., 2013), and artificial neural networks (ANNs) (Ni and Xue, 2003; Li et al., 2013). These methods, which improve computing significantly, can better solve non-linear problems, but still exhibit a number of weaknesses. For example, SVMs are complex mathematical functions which are rather incomprehensible for human users (Martens et al., 2007). Considerable pre-treatment is required for using a DT (Kubal et al., 2009), and it readily falls into local optimization (Liu et al., 2008). The ANN method shows over-learning and slow convergence speed problems (Li and Yeh, 2002). Even worse, these intelligence algorithms are unable to estimate each index's contribution to the total risk. Though notable achievements have been made to rectify these weaknesses, efficient and effective methods are still urgently required.

Random forest (RF), a machine-learning algorithm proposed by Breiman, is a combination classification method based on statistical learning theory (Breiman, 2001). In a random forest, multiple samples are drawn using the resampling bootstrap method, and classification trees are built corresponding to each bootstrap sample. Eventually, all forecast classification trees are combined and final classification results are obtained by voting. The RF algorithm is a natural and non-linear modelling tool that provides estimates regarding the hierarchy of variables in the classification, and thus is able to estimate each index's contribution to the total risk. The RF algorithm has been applied to fields such as earthquake-induced damage classification (Solomon and Liu, 2010), prediction of rock burst classification (Dong et al., 2013), genomic data analysis (Chen and Ishwaran, 2012), tree species classification (Immitzer et al., 2012), gene selection (Deng and Runge, 2013), and computer-aided diagnosis (Mihailescu et al., 2013). A great deal of theoretical and empirical studies have detailed the many advantages of RF, including high forecast accuracy, acceptable tolerance to outliers and noise, and easy avoidance of over-fitting problems. Based on this body of knowledge, RF should be, in theory, highly applicable to flood hazard risk assessment and able to rectify multi-variable and non-linear issues; however, few applications in this field have been previously reported.

The primary objectives of this study are: (1) developing a systematic procedure for assessment using RF, (2) proving that RF is a feasible and reasonable method of flood hazard risk assessment, and (3) successfully analyzing the flood hazard risk distribution of the study basin. This study provides a novel approach to flood hazard risk assessment, showing significant scientific and practical merits in terms of flood insurance, flood risk management, prevention, and reduction of natural disasters in the study basin and beyond.

2. Study area and data

2.1. Study area

The Dongjiang River, a major tributary of the Pearl River, China, is 562 km in length with a drainage area of 27,363 km^2, accounting for approximately 5.96% of the Pearl River Basin (Fig. 1). The Dongjiang River Basin has a subtropical climate, with a mean annual temperature of approximately 21 °C. Front and typhoon-type rainfalls are predominant in the basin, and annual rainfall ranges from 1500 mm to 2400 mm (Liu et al., 2010). Large seasonal variations in rainfall and runoff exist within the basin, with 80% of the annual rainfall and runoff occurring in the rainy season and only 20% occurring in the dry season.

The Dongjiang River Basin, an economically advanced and densely populated area, is comprised mostly of six cities: Ganzhou, Heyuan, Huizhou, Dongguan, Guangzhou, and Shenzhen. The river is the primary water source for these cities as well as Hong Kong. In fact, the proportion of Dongjiang water in Hong Kong's annual water supply has steadily increased, from only 8.3% in 1960 to approximately 70% or even slightly over 80% in recent years (Jiang et al., 2007). The subtropical climate of South China causes substantial rainfall in the basin every year, making it a flood-prone area. For example, in the Xintian and Heyuan precipitation stations in the basin, maximum 24-h precipitation was measured up to 448 mm and 327.2 mm, respectively, in June 1959. This rainfall formed a super flood that killed 78 people, injured 443, submerged 159,000 hm^2 of farmland, and destroyed 11,900 water conservancy projects. Worse still, two downstream cities – Guangzhou and Shenzhen, are ranked as the first and fifth highest flood risk in 2050, among the top 136 major coastal cities at risk (Stephane et al., 2013). Undoubtedly, the negative influence of both natural and social factors contributes to the dire need for flood risk management in the area. To attempt to prevent floods and reduce potential loss as much as possible, a systematic evaluation of flood hazard risk in the basin is vital.

2.2. Data

The multi-variable and non-linear relationship between indices and risk levels is the primary challenge inherent to flood hazard risk assessment. The first task is selection of suitable risk indices. Risk index variables differ between study areas according to the specific characteristics of each location, however (Tehrany et al., 2013). An index which shows a high degree of impact on flood hazard risk in a specific area may not rank similarly in other areas (Kia et al., 2012). After considering the actual conditions of floods and relevant characteristics in the study area and reviewing recommendations provided by previous research (Jiang et al., 2009; Zou et al., 2013; Yang et al., 2013), three indices of disaster-inducing factors and eight indices of hazard-inducing