Geomorphology 298 (2017) 118–137

Contents lists available at ScienceDirect

Geomorphology

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

Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion

Omid Rahmati [a], Nasser Tahmasebipour [a,*], Ali Haghizadeh [a], Hamid Reza Pourghasemi [b], Bakhtiar Feizizadeh [c]

[a] Department of Watershed Management, Faculty of Natural Resources and Agriculture, Lorestan University, Lorestan, Iran
[b] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran
[c] Department of Remote Sensing and GIS, University of Tabriz, Tabriz 51368, Iran

ARTICLE INFO

Article history:
Received 7 September 2016
Received in revised form 6 August 2017
Accepted 8 September 2017
Available online 27 September 2017

Keywords:
Gully erosion
Spatial prediction
Machine learning
Robustness

ABSTRACT

Gully erosion constitutes a serious problem for land degradation in a wide range of environments. The main objective of this research was to compare the performance of seven state-of-the-art machine learning models (SVM with four kernel types, BP-ANN, RF, and BRT) to model the occurrence of gully erosion in the Kashkan-Poldokhtar Watershed, Iran. In the first step, a gully inventory map consisting of 65 gully polygons was prepared through field surveys. Three different sample data sets (S1, S2, and S3), including both positive and negative cells (70% for training and 30% for validation), were randomly prepared to evaluate the robustness of the models. To model the gully erosion susceptibility, 12 geo-environmental factors were selected as predictors. Finally, the goodness-of-fit and prediction skill of the models were evaluated by different criteria, including efficiency percent, kappa coefficient, and the area under the ROC curves (*AUC*). In terms of accuracy, the RF, RBF-SVM, BRT, and P-SVM models performed excellently both in the degree of fitting and in predictive performance (*AUC* values well above 0.9), which resulted in accurate predictions. Therefore, these models can be used in other gully erosion studies, as they are capable of rapidly producing accurate and robust gully erosion susceptibility maps (GESMs) for decision-making and soil and water management practices. Furthermore, it was found that performance of RF and RBF-SVM for modelling gully erosion occurrence is quite stable when the learning and validation samples are changed.

1. Introduction

Gullies are an important sediment sources and often cause environmental problems within their reach (on-site effects) and downstream (off-site effects) (Boardman et al., 2003; Smolska, 2007; Wu et al., 2008). A gully is usually defined as an erosion channel with a cross-sectional area of > 1 ft^2 (Poesen et al., 1996) that is too big to be obliterated by conventional tillage (FAO, 1965; USDA-SCS, 1966). Generally, gully erosion results in different consequences: (i) significant land degradation and loss of productive capacity, (ii) high sediment yields and sediment discharge, which can transport both nutrients and pollutants, and (iii) sedimentation of reservoirs (reducing the water capacity of the reservoirs) and damage to the infrastructure and transport routes (Bufalo and Nahon, 1992; Poesen et al., 1998; Valentin et al., 2005; Lesschen et al., 2007a, 2007b; Ekholm and Lehtoranta, 2012; Fox et al., 2016). A contribution of 10–94% of the total sediment yield by gully erosion landforms in the watershed scale has been reported by Poesen et al. (2003). From an ecological viewpoint, gully erosion can cause related ecological problems such as eutrophication and acceleration of desertification processes (Carpenter, 2005; Valentin et al., 2005; Rekolainen et al., 2006; Ekholm and Lehtoranta, 2012). There is unanimity in the literature that the detection of potential gully prone areas in watersheds is notably important work for mitigation and prevention (Popp et al., 2000; Sidorchuk et al., 2003; Kumar and Nair, 2006; Li et al., 2016). Effective analysis of gully erosion susceptibility could provide planners with the foreknowledge of susceptible zones and thereby help with watershed management, water and soil conservation measurements, and infrastructure planning (Ni et al., 2008; Bouaziz et al., 2011).

A gully erosion susceptibility assessment is the first step towards estimation of gully erosion hazard and risk (Conoscenti et al., 2014). In the literature, several GIS-based models have been applied for gully erosion susceptibility mapping such as the frequency ratio (Conforti et al., 2011; Lucà et al., 2011), weights of evidence (Dube et al., 2014; Rahmati et al., 2016), logistic regression (Martínez-Casasnovas et al., 2004; Akgün and Türk, 2011; Lucà et al., 2011; Conoscenti et al., 2014), linear regression (Chaplot et al., 2005a), conditional analysis (Conoscenti et al., 2008; Magliulo, 2010, 2012; Conoscenti et al., 2013), the analytical hierarchy process (Zakerinejad and Maerker, 2014), classification and regression trees (Bou Kheir et al., 2007; Geissen et al., 2007; Gómez Gutiérrez et al., 2009b; Märker et al., 2011), multivariate adaptive regression splines (Gómez Gutiérrez et al., 2009a, 2009b; Gómez-Gutiérrez et al., 2015), and maximum entropy (Zakerinejad and Maerker, 2014). Additionally, some studies have addressed gully erosion susceptibility analysis

* Corresponding author.
E-mail address: Tahmasebi.n@lu.ac.ir (N. Tahmasebipour).

https://doi.org/10.1016/j.geomorph.2017.09.006
0169-555X/

using index-based methods such as the stream power index (Zakerinejad and Maerker, 2015), the gully density index (Hughes et al., 2001), the normalized topographic method (Castillo et al., 2014), and spatial information technology (Martínez-Casasnovas, 2003a, 2003b). However, traditional statistical methods have some drawbacks for determining the relationship between geo-environmental factors and gully erosion occurrence because of the definition of prior statistical assumptions (e.g., assumptions on data distribution) for the analysis (Tehrany et al., 2013; Polykretis et al., 2015). For example, the assumption of the variables' independence (i.e., predictors) is considered a limitation in susceptibility analysis, which is usually violated in practice (Ballabio and Sterlacchini, 2012).

According to the literature, the advent of machine learning techniques, such as random forest (RF), boosted regression trees (BRT), artificial neural network (ANN), and support vector machine (SVM), has contributed significantly to the field of susceptibility mapping of landslide (Lee et al., 2003a, 2004; Gomez and Kavzoglu, 2005; Nefeslioglu et al., 2008a, 2008b, 2008c; Yao et al., 2008; Yilmaz, 2009; Catani et al., 2013; Pradhan et al., 2010a, 2010b; Xu et al., 2012; Pradhan, 2013; San, 2014; Dou et al., 2015; Gorsevski et al., 2016; Hong et al., 2016), debris flow (Yuan et al., 2006; Chang, 2007; Chang and Chao, 2006), and ground subsidence (Oh and Lee, 2011; Lee et al., 2012). In recent years, Svoray et al. (2012) employed different machine learning models, such as decision tree (DT), SVM, and ANN, for predicting gully initiation at the catchment scale, and then, they compared their results with the analytic hierarchy process (AHP) and topographic threshold (TT) methods. The results of this study indicated that machine learning models provide a better predictive ability of gully initiation points than the application of both AHP and TT methods. Recently, Fernandes et al. (2017) evaluated the application of a SIMWE (SIMulated Water Erosion) model for gully erosion susceptibility analysis in the Douro region, Adorigo. Their results indicated that the SIMWE model had a moderate performance in identifying the susceptible areas to gully erosion. Angileri et al. (2016) applied Stochastic Gradient Treeboost (SGT), which is a multivariate statistical model, in central-northern Sicily (Italy) to analyse and predict the spatial occurrence of rill–interrill erosion and gully erosion. They stated that SGT allowed them to understand the relationships between erosion landforms and geo-environmental factors. Kuhnert et al. (2010a, 2010b) applied the RF model to predict gully density and the gully erosion rate through a suite of environmental predictors and to estimate the prediction uncertainty. However, inspection of the literature revealed that evaluating and predicting the efficiency of the SVM and the ANN models for gully erosion mapping is rarely studied (e.g., Svoray et al., 2012), and therefore, they should be further investigated in other regions. Due to the complex nature of gully erosion, such as the soil condition, lithology, topography, hydrology, and human activities, producing a reliable spatial prediction of gullies is still a challenging task. Despite the many efforts that have been made in gully erosion susceptibility and hazard modelling, there is still a dispute over which model or technique is the best for the identification of gully prone areas. Additionally, the best model for an area depends not only on the quality of the data used but also strongly on the employed modelling approaches (i.e., model structure) (Bui et al., 2016). Therefore, the evaluation of these techniques, including their comparison with field data, is highly necessary to obtain an adequate background to draw some reasonable conclusions. To address this, a broad range of machine learning models have been proposed from different points of view to understand their controlling factors and to identify gully susceptible zones. Therefore, the principal justification of this study is to compare the performance of seven advanced machine learning models for predicting the spatial occurrence of gully erosion and to gain insights into the limitations and strengths of these models.

In the current study, seven machine learning models (RF, BRT, BP-ANN, L-SVM, RBF-SVM, P-SVM, and S-SVM) were selected for the spatial prediction of gully erosion because of the following reasons: (1) they can model the non-linear relationship between the predictors (i.e., gully conditioning factors) and output terms (i.e., gully erosion occurrence); (2) they can work with different types of independent variables and can handle data from various measurement scales; (3) they do not define strict assumptions prior to the study; (4) a review of the research indicated that only a few studies have employed SVMs and ANN models for assessing gully erosion susceptibility (e.g., Svoray et al., 2012); and (5) according to research background, there is no comprehensive study to compare the capability of RF and BRT models for assessing gullying erosion susceptibility. To address the research gaps, the original study was conducted in the Kashkan-Poldokhtar Watershed, Iran, as an area highly prone to gully erosion. Therefore, the specific objectives of this study are to (1) spatially predict the gully erosion occurrence in the Kashkan-Poldokhtar Watershed, (2) develop RF and BRT models for predicting gully susceptibility and comparing their results with BP-ANN and SVMs models, and (3) evaluate the capability and robustness of applied machine learning techniques using different sample data sets and evaluation criteria.

2. Study area description

The Kashkan-Poldokhtar Watershed is located between the Lorestan and the Ilam Provinces and covers an area of 245 km^2 (Fig. 1). The southern sector has a mountainous landscape with an average elevation of 706 m a.s.l. and is characterized by steep slopes. The northern areas include low-mountains and high slopes, whereas the central and western parts are generally characterized as plain and flat areas. According to climatic classification in Iran (IDWRM, 2013), the study area has a semi-arid climate (100–400 mm) with hot and dry summers and precipitation concentrated in the mild winters. The mean annual rainfall is approximately 385 mm and is distributed in 68 rainy days, whereas the mean annual temperature is 21 °C and the mean monthly temperatures range is between 6.6 °C in February and 35 °C in August. The soil erosion processes of the study area are strongly controlled by the geological and geomorphological setting. There is a low variability of soil types ranging from Inceptisols to Entisols (USDA, 2006) with soil profiles often characterized by erosive truncation caused by water erosion.

There are two main reasons for the selection of the Kashkan-Poldokhtar Watershed. First, some of the basic physiographic data, such as digital elevation model (DEM), land use, lithology, and soil, have been collected by the Department of Natural Resources Management (DNRM) and are available for this study. Second, it is highly susceptible to gully erosion occurrences and land degradation, which is partly related to its geological origin and somewhat related to human activities such as overgrazing, deforestation, and improper tillage practices. In the Kashkan-Poldokhtar Watershed, the mean width and depth of the gullies are relatively large with a minimum depth and width of approximately 3 and 5 m, respectively. Two field photographs of the recent gullies identified in the study area are shown in Fig. 2.

3. Methodology

The methodology used in the current study contains various main steps. Fig. 3 illustrates the methodological flowchart and an overview of the approach that was developed for probabilistic gully erosion analysis using ANN and SVM (with four kernel types) and the RF and BRT models.

3.1. Dataset used

3.1.1. Splitting the dataset and sampling

A gully inventory map provides some important baseline information on the distribution pattern of the events in the affected area and furthers the understanding of the gullying process and the assessment of relationships between conditioning factors and gully erosion (Wu and Cheng, 2005; Duman et al., 2005; Galli et al., 2008). Thus, the preparation of an inventory of gully landforms is a key step for gully