Science of the Total Environment 609 (2017) 764–775

Contents lists available at ScienceDirect

Science of the Total Environment

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

Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling

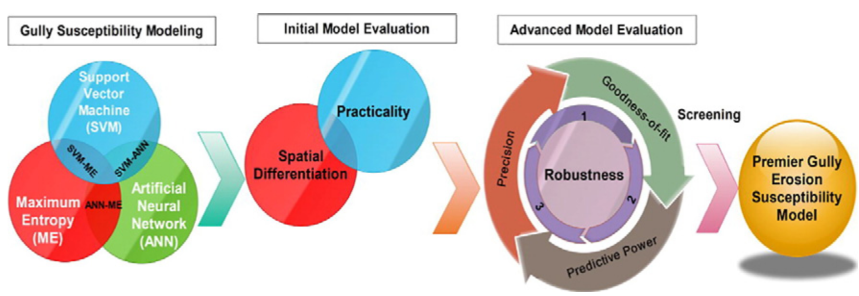
Hamid Reza Pourghasemi [a], Saleh Yousefi [b], Aiding Kornejady [c], Artemi Cerdà [d,*]

[a] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran
[b] Department of Watershed Management Engineering, Faculty of Natural Resources, Tarbiat Modares University, Iran
[c] Department of Watershed Management Engineering, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran
[d] Soil Erosion and Degradation Research Group, Departament de Geografia, Universitat de València, Blasco Ibàñez 28, 46010 Valencia Spain

HIGHLIGHTS

- Gully erosion susceptibility mapping models were evaluated.
- The ME model showed 45% of the study area as highly susceptible to gullying.
- ANN-SVM model shown 34% of the study area as highly susceptible.
- The role of ensemble modeling in relevant to building accurate and generalized models.
- Results prepare an outline for further biophysical designs on gullies scatter.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 June 2017
Received in revised form 22 July 2017
Accepted 22 July 2017
Available online 29 July 2017

Editor: D. Barcelo

Keywords:
MaxEnt
SVM
ANN
Goodness-of-fit
Prediction power
Robustness
Vulnerability

ABSTRACT

Gully erosion is identified as an important sediment source in a range of environments and plays a conclusive role in redistribution of eroded soils on a slope. Hence, addressing spatial occurrence pattern of this phenomenon is very important. Different ensemble models and their single counterparts, mostly data mining methods, have been used for gully erosion susceptibility mapping; however, their calibration and validation procedures need to be thoroughly addressed. The current study presents a series of individual and ensemble data mining methods including artificial neural network (ANN), support vector machine (SVM), maximum entropy (ME), ANN-SVM, ANN-ME, and SVM-ME to map gully erosion susceptibility in Aghemam watershed, Iran. To this aim, a gully inventory map along with sixteen gully conditioning factors was used. A 70:30% randomly partitioned sets were used to assess goodness-of-fit and prediction power of the models. The robustness, as the stability of models' performance in response to changes in the dataset, was assessed through three training/test replicates. As a result, conducted preliminary statistical tests showed that ANN has the highest concordance and spatial differentiation with a chi-square value of 36,656 at 95% confidence level, while the ME appeared to have the lowest concordance (1772). The ME model showed an impractical result where 45% of the study area was introduced as highly susceptible to gullying, in contrast, ANN-SVM indicated a practical result with focusing only on 34% of the study area. Through all three replicates, the ANN-SVM ensemble showed the highest goodness-of-fit and predictive power with a respective values of 0.897 (area under the success rate curve) and 0.879 (area under the prediction rate curve), on average, and correspondingly the highest robustness. This attests the important role of ensemble modeling in congruently building accurate and generalized models which emphasizes the necessity to examine different models integrations. The result of this study can prepare an outline for further biophysical designs on gullies scattered in the study area.

* Corresponding author.
E-mail addresses: hr.pourghasemi@shirazu.ac.ir (H.R. Pourghasemi), artemio.cerda@uv.es (A. Cerdà).

http://dx.doi.org/10.1016/j.scitotenv.2017.07.198

1. Introduction

Water erosion is an integral part of geological and geomorphological cycle of the earth system. It also causes massive damages to agricultural lands and sometimes has irrecoverable and destructive impacts on dams, reservoirs, and water quality in semi-humid and arid areas (Mekonnen et al., 2017; Comino et al., 2016; Kheir et al., 2007; Buttafuoco et al., 2012). Amplified soil erosion rates have recently been related to the so called "environmental land use conflicts" (Pacheco et al., 2014; Valle Junior et al., 2014), which bear on uses of the land that deviate from its capability (natural use), and that these higher erosion rates tend to reduce soil fertility by important reductions in organic matter content (Valera et al., 2016). Gully erosion is a morphologically emerged process (Maslov, 2005) formed by water erosion with a substantial flow rate in a determined area. Generally, it causes deep cuts with tens of meters in depth and width which is imperceptibly initiated on a hillside and scours soil (Billi and Dramis, 2003). Gullies dramatically decrease soil productivity by incising agricultural lands and consequently cause restrictions in land use, roads, fences, and structures (Takken et al., 2008; Akgün and Türk, 2011; Mekonnen et al., 2017; Zakerinejad and Märker, 2015). As the most prominent feature, gullies remove upland soils along drainage lines by surface runoff and make it hard to conduct tillage operations (USDA-SCS, 1966). They are one of the most dominant causes of geo-environmental degradation in the west (Rahmati et al., 2016a) and the north part of the Iran due to present land uses and geoclimatic agents.

From data availability view point, data mining and statistical methods have been indisputably coped with data scarcity issue, especially those geophysical and geochemical data that are being used by the gully physical models such as CREAMS (Chemicals, Runoff, and Erosion from Agricultural Management Systems), EGEM (Ephemeral Gully Erosion Mode), and WEPP (Water Erosion Prediction Project) (Knisel, 1980; Flanagan and Nearing, 1995; Woodward, 1999). As noted by Conoscenti et al. (2013), these physical methods need to be tested before being used. Moreover, they do not assess gully erosion susceptibility, while susceptibility maps are the most important level of conceiving the exposition of an area to gullying. Different data mining, bivariate, and multivariate statistical methods have been used in many environmental fields. Some of these have been used for assessing gully erosion susceptibility including classification and regression trees (CART) (Gómez-Gutiérrez et al., 2009a, b; Märker et al., 2011), logistic regression (LR) (Chaplot et al., 2005a, b; Lucà et al., 2011; Conoscenti et al., 2014; Kornejady et al., 2015), information value (Conforti et al., 2011); weights of evidence (WofE) (Dube et al., 2014); frequency ratio (FR) (Rahmati et al., 2016a); multivariate adaptive regression splines (MARS) (Gómez-Gutiérrez et al., 2015), and random forest (RF) (Kuhnert et al., 2010). Thus, a wide range of data mining methods has still remained unused. For instance, maximum entropy model has been widely employed in different fields such as environmental and ecological science (Phillips et al., 2004; Phillips and Dudík, 2008; Fourcade et al., 2014; Ariyanto, 2015; Cao et al., 2016) and landslide susceptibility mapping (Kim et al., 2015; Davis and Blesius, 2015; Dickson and Perry, 2016; Kornejady et al., 2017). The support vector machine (SVM) and artificial neural network (ANN) also were applied in landslide field (Tsangaratos and Benardos, 2014; Ren et al., 2015; Tien Bui et al., 2016; Chen et al., 2017). On the other hand, ensemble modeling is getting more popular nowadays due to their accurate results. It can integrate models to achieve high performance results in terms of goodness-of-fit and predictive power in an efficient amount of time (Moonjun, 2007; Nefeslioglu et al., 2010; Pradhan, 2013; Umar et al., 2014). This being a tangible gap in gully erosion assessments, the objectives of this work are the following: 1) Use of three individual models (ANN, SVM, ME) and their ensembles (ANN-SVM, ANN-ME, and SVM-ME) in Aghemam Watershed; 2) conducting initial performance assessment statistics; 3) Calibrating the models by assessing the goodness-of-fit; and 4) validating the results by assessing the predictive power, precision, and robustness.

2. Material and methods

2.1. Study area

The Aghemam Watershed has an area of 2595 ha and is situated in the east of Golestan Province in northern Iran (Fig. 1), with an altitude range between 357 and 822 m asl. Average slope of study watershed is about 13% and the maximum length is 9.4 km. The main channel length in the study area is about 10.4 km with an average slope of 4%. Silty-loamy (about 87.5%) and Silty-loamy–loamy (about 12.5%) soils are the dominant soil textures. The main land uses in the study area comprise accordingly rangelands (66.5%), farmlands (33.16%), and bare lands (0.26%). The prevailed land covers in the study area include *Artemisia* + *Poa* species (34.86%), croplands (33.16%), *Poa* + *Medicago* species (25.69%), *Paliurus* + *Artemisia* species (6.01%), and bare lands (0.26%). Geologically, the study area covered by Marl and Shaleston (245 ha), Loos (2230 ha), and quaternary alluvial fans (120 ha).The variation of annual precipitation in the study area is about 75 mm with an average of 491 mm (Mohammad-Ebrahimi et al., 2015). The average annual temperature of the study area stands at 28 °C.

2.2. Methods

The methodological process of the current study is presented in Fig. 2. As shown, the flowchart comprises three main steps including: 1) data preparation; 2) modeling process (objective 1); 3) performance analysis consists of initial performance assessment (objective 2) and advanced performance assessment (objective 3 and 4).

2.2.1. Inventory map of gullies

The gully erosion inventory map was prepared using field surveys with a DGPS (Differential Global Positioning System) device. Series of linear and digitated gullies with U-shaped and V-shaped cross-sections with tens of meters in width and depth are evident in the study area, mostly located nearby roads and foot of the hills (Figs. 1 and 3). In total, 25 gullies are scattered in the study area with an area of about 105.88 ha including six digitated (64.45 ha) and 19 linear gullies (41.43 ha), where the locations of all 25 gullies were recorded, mapped as polygons, and then all the cells intersected by gullies (2647 pixels with a 20 m resolution as positive samples) were used for modeling. 70% of gullies were randomly selected for training (18 gullies; equivalently 1985 pixels) and the 30% rest were set aside to validate the built models (7 gullies; equivalently 662 pixels) (Youssef et al., 2015; Hussin et al., 2016). The same configuration was applied for negative points (non-gully areas) where the same number and percentage of negatives was used in calibration and validation procedures (Lombardo et al., 2014; Cama et al., 2016; Kornejady et al., 2017). As proposed by Conoscenti et al. (2014), the positive and negative training/test sets were altered three times in order to assess the robustness of the models (Fig. 4). For better graphical check, only positive training and test sets are presented.

2.2.2. Preparation of the conditioning factors

The main geo-environmental factors affect gully erosion are rainfall features such as intensity, period, and spatial distribution (Kheir et al., 2007; Magliulo, 2012; Capra et al., 2012), topography derived factors such as contributing drainage area, distance from ridges, slope steepness, slope aspect, and slope curvatures (Montgomery and Dietrich, 1992; Samani et al., 2009; Capra et al., 2012; Conoscenti et al., 2013), lithology and soil related properties and features such as erosivity, soil water content, soil texture, and sub-surface flow (Parras-Alcántara et al., 2016; Marzolff et al., 2011; El Maaoui et al., 2012) and land use/ land cover (LU/LC) (Poesen et al., 2003; Takken et al., 2008;